

INTRODUCTION TO LOG-LINEAR AND LATENT CLASS MODELS

Richard Breen
CIQLE/ Department of Sociology
Yale University
richard.breen@yale.edu

Saturday November 21st 2009

INTRODUCTION:

Program:

Morning: Log-linear models

Afternoon: Latent class models

LEM program:

You can download the program from here:

<http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html>

One limitation of *LEM* is that it has no facilities for data manipulation – it wants the data exactly as they need to be for the model you want to fit. So it is convenient to call the program from within R using the ‘system’ command as follows:

```
system("C:/lem/lemwin/lem95 example_2.inp example_2.out")
```

PART 1: Analysis of Cross-Classified Categorical Data using Log-linear models

Notation:

Simplest case is a 2-way table:

Alcohol Use	Cigarette Use	
	Yes	No
Yes	1449	500
No	46	281

Source: Agresti Table 8.3.

	j=1	j=2	j=3
i=1	f_{11}	f_{12}
i=2	f_{21}	f_{22}	...
i=3	f_{33}

Row variable - R or I - indexed by $i = 1, \dots, I$

Column variable - C or J - indexed by $j = 1, \dots, J$

f_{ij} is the observed frequency in the ij^{th} cell

$$N = \sum_i \sum_j f_{ij}$$

We use F_{ij} to mean the frequency in the ij^{th} cell expected under some model, say, M.

We usually work with logarithms of frequency counts, $\log(f_{ij})$ and $\log(F_{ij})$ where \log is the natural log.

This extends straight forwardly to a 3-way or higher order crosstabulation. Then we have, say, f_{ijkl} etc. with variables I, J, K and L (or R, C, L and T, e.g.)

Marginal frequencies are usually written $f_{i.}$ or $f_{.i+}$ to mean the frequencies of the row variable.

$$\text{So } f_{i.} \equiv \sum_j f_{ij}$$

In a higher dimension table we might have: $f_{+++} = \sum_i \sum_j \sum_l f_{ijkl}$

Equally we could have

$$f_{ij++} = \sum_k \sum_l f_{ijkl}$$

and this is the two-way joint marginal distribution of I and J.

Example: Much mobility research focuses on O x E x D table where

O_i , $i=1, \dots, I$ is class origins

E_j , $j=1, \dots, J$ is educational level

D_k , $k=1, \dots, K$ is class destination

So $f_{i+k} = \sum_j f_{ijk}$ is the simple O x D table.

Chi - Square Test

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

F_{ij} is computed as $\frac{\sum_i f_{ij} \times \sum_j f_{ij}}{N^2}$ - i.e. the product of the marginal totals divided by N^2 .

This follows because if R and C are independent then $p_{ij} = \frac{f_{ij}}{N} = p_{i+} \times p_{+j}$.

Test statistic is distributed as χ^2 variate with df equal to $(I-1)(J-1) =$
 $(\# \text{ rows}-1) \times (\# \text{ columns}-1)$

Another way to think of this test is to begin with the model of independence that generates F_{ij} .

Total df before any parameters are fitted = # cells = $I \times J$.

The difference between $I \times J$ and $(I - 1)(J - 1)$ is $J+I-1$ - so in forming F_{ij} we have estimated $J+I-1$ parameters.

Example: Imagine a 4x3 table: $I \times J = 12$

$$(I-1)(J-1) = 6$$

$$\text{therefore } J+I-1 = 4 + 3 - 1 = 6$$

We could think of these as 6 parameters which we somehow needed to arrive at F_{ij} as

- (1) an effect that fixes the total size of the table, N
- (2) effects that fix the row totals
- (3) effects that fix the column totals.

So our model of independence fits $1+(I-1) + (J-1) = I+J-1$ effects as required.

Write a model for F_{ij} in which these three sets of effects are multiplicative

$$F_{ij} = \mu \tau_i^R \tau_j^C$$

$$\tau_i^R \propto p_{i+}$$

$$\tau_j^C \propto p_{+j}$$

μ scales $\tau_i^R \tau_j^C$ to the sample size.

We can write the model in additive form as:

$$\begin{aligned} \log(F_{ij}) &= \log \mu + \log \tau_i^R + \log \tau_j^C \\ &= \alpha + \beta_i^R + \beta_j^C \end{aligned}$$

Since we can only estimate (I-1) row effects and (J-1) column effects, we must constrain τ_i and τ_j in some way. There are two main ways:

- (1) dummy variable coding: set one of them to 1 (multiplicative) or to 0 (additive);
- (2) centered coding: force all of them to have the product 1 (equally, in the log form, to sum to zero).

Example:

Alcohol Use	Cigarette Use	
	Yes	No
Yes	1449	500
No	46	281

Fit the independence model: yields fitted values F_{ij}

$$\begin{array}{cc} 1280.21 & 668.79 \\ 214.79 & 112.21 \end{array} \quad \chi^2 = 451.40, \text{ df}=1.$$

Parameter estimates:

(a) Dummy variable coding:

	Multiplicative	Additive
μ or α	1.28E+0003	2.4849
τ^R or β^R (2)	0.1678	-1.7851
τ^C or β^C (2)	0.5224	-0.6493

Notice that the first row and column effects are set to 1 (multiplicative)/ 0 (additive).

Notice too that the multiplicative row parameter is equal to the ratio of the second row total to the first row total and likewise for the column parameter. *This is only true in models of independence. The proportionality relationship between the row and column marginal probabilities and the model row and column parameters only holds for independence and not otherwise.*

(b) Centered, or effect, coding:

μ or α	379.01	5.9376
τ^R or β^R (1)	2.4414	0.8926
τ^R or β^R (2)	0.4096	-0.8926
τ^C or β^C (1)	1.3836	0.3247
τ^C or β^C (2)	0.7228	-0.3247
	↓	↓
	Product = 1	Sum = 0

Model Fitting

Usual assumption is that the frequencies in each cell of the table follow a Poisson distribution.

The Poisson probability is

$$p(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x \geq 0, x \text{ integer.}$$

λ is the Poisson mean. For the Poisson, variance = λ also.

So, for a particular cell of a 3-way table we have $p(f_{ijk}) = \frac{F_{ijk} e^{-F_{ijk}}}{f_{ijk}!}$

The joint distribution over all $I \times J \times K$ cells of the table is the product of this for all cells:

$$p(f_{ijk}) = \prod_{ijk} \frac{F_{ijk} e^{-F_{ijk}}}{f_{ijk}!}$$

and so the log-likelihood is

$$\sum_{ijk} f_{ijk} \log(F_{ijk}) - \sum_{ijk} F_{ijk} - \sum_{ijk} \log(f_{ijk}!)$$

The last term is irrelevant and so the kernel of the log-likelihood consists of just the first two.

Replace $\ln(F_{ijk})$ with a model, e.g. $= \alpha + \beta_i^R + \beta_j^C$ whose parameters we want to estimate and then find the maximum of the kernel of the log-likelihood.

The likelihood equations are essentially the same for the Poisson and for two other sampling schemes: the multinomial and the product-multinomial.

Poisson – arises as the count of the # of events in a cell in a fixed time;

Multinomial - fixed N and each sample member is then classified according to his/her value on the marginal variables

Product multinomial – fixed numbers in each of the categories of one of the variables (e.g. row) and rest allocated as for multinomial

Capturing Association

The model:

$$\log(F_{ij}) = \alpha + \beta_i^R + \beta_j^C$$

postulates independence of R and C: $p_{ij} = p_{i+} \times p_{+j}$

But if R and C are not independent then this does not hold. In log-linear models we capture dependence between pairs of variables using odds ratios, θ .

In the observed data an odds ratio is

$$\frac{f_{ij} / f_{i'j'}}{f_{ij'} / f_{i'j}} = \theta_{ii'jj'}$$

In the special case where $j'=j+1$ and $i' = i+1$ we have an odds ratio formula for four adjacent cells, sometimes called adjacent or local or interstitial odds ratio. This is the only kind there is in a 2 x 2 table. But consider a 3x3 example:

	j=	1	2	3
i =	1	a	b	c
	2	d	e	f
	3	g	h	i

$\frac{a/b}{d/e}$ is an interstitial odds ratio whereas $\frac{d/f}{g/i}$ is not.

Notice that θ can also be written $\frac{f_{ij} \times f_{i'j'}}{f_{ij'} \times f_{i'j}} = \theta_{ii'jj'}$ or $\frac{ae}{bd}$.

In this case $\begin{array}{cc} 5 & 2 \\ 10 & 6 \end{array}$ $\theta = 30/20 = 1.5$, positive association

While here $\begin{array}{cc} 2 & 6 \\ 8 & 12 \end{array}$ $\theta = 24/48 = 0.5$, negative association

And here $\begin{matrix} 3 & 2 \\ 6 & 4 \end{matrix}$ $\theta = 12/12 = 1$, no association

+ve association: the odds of being found in column j rather than j' are larger if you come from row i rather than i' .

-ve association: the opposite

No association: the odds are the same i.e. $R \perp C$.

$\log(\theta)$ is widely used: so $\begin{matrix} +ve & > 0 \\ -ve & < 0 \\ indep & = 0 \end{matrix}$

In a table of dimensions $I \times J \ni IJ (I-1)(J-1) / 4$ odds ratios, θ .

e.g. in a 4x4 table there are $\frac{16 \times 9}{4} = 36$ odds ratios.

But they can all be written as a function of $(I-1)(J-1)$ θ s in any basis set.

$(I-1)(J-1)$ should be familiar.

In a log-linear model we add another set of terms to capture association:

$$\log(F_{ij}) = \alpha + \beta_i^R + \beta_j^C + \beta_{ij}^{RC}$$

The association parameters are the log of the odds ratios of a particular basis set – but which set depends on the parameterization of the main effects (Breen 2007 *SMR* discusses this in more detail and shows how to set association parameters to equal log-odds ratios of interest).

Under dummy variable coding they look like this (in a 3x3 case)

$$\begin{array}{ccc|ccc} 0 & 0 & 0 & a & b & c \\ 0 & \beta_{22} & \beta_{23} & d & e & f \\ 0 & \beta_{32} & \beta_{33} & g & h & i \end{array}$$

$$\beta_{22} = \ln [(a/b)/(d/e)]$$

$$\beta_{23} = \ln [(a/c)/(d/f)]$$

$$\beta_{32} = \ln [(a/b)/(g/h)]$$

$$\beta_{33} = \ln [(a/c)/(g/i)]$$

But because these are a basis set we can use them to compute any log θ that we want:

$$\ln ((b/c) / (e/f)) = \beta_{23} - \beta_{22}$$

$$\ln ((d/g) / (g/i)) = \beta_{33} - \beta_{23}$$

So any log odds ratio can be expressed as a function of the association parameters of a log-linear model.

Odds ratios are invariant to any scalar multiplication of the rows or columns of a table.

Alcohol Use	Cigarette Use	
	Yes	No
Yes	1449	500
No	46	281

Odds ratio is 17.703. Suppose we multiply the first column of the table by a and the second by b: then the odds ratio becomes $\frac{1449a/500b}{46a/281b} = 17.703$.

Log-linear models in matrix form

$$\log(\mathbf{y}) = \mathbf{X}\hat{\boldsymbol{\beta}}$$

Here \mathbf{y} is the cell frequencies from the table, laid out as a column vector.

$\boldsymbol{\beta}$ is the vector of coefficients whose length depends on the specification of the model we are fitting.

\mathbf{X} is the design matrix. It has the same number of rows as the vector \mathbf{y} but how many columns depends on the model.

\mathbf{X} is set up to fit all the main effects and interactions in the specified model. So a model like (AB) (BC) (AC) would require that we fit:

- An overall intercept or mean effect (1)
- Main effects for A (2), B(1) and C(2)
- Interaction terms for AB (2), BC (2) and AC (4)

What does \mathbf{X} look like? Using dummy variable coding with the last category being the reference:

$y(i,j,k)$	Mean	A1	A2	B	C1	C2	A1*B	A2*B	etc
1,1,1	1	1	0	1	1	0	1	0	
1,1,2	1	1	0	1	0	1	1	0	
1,1,3	1	1	0	1	0	0	1	0	
1,2,1	1	1	0	0	1	0	0	0	
1,2,2	1	1	0	0	0	1	0	0	
1,2,3	1	1	0	0	0	0	0	0	
2,1,1	1	0	1	1	1	0	0	1	
2,1,2	1	0	1	1	0	1	0	1	
2,1,3	1	0	1	1	0	0	0	1	
2,2,1	1	0	1	0	1	0	0	0	
2,2,2	1	0	1	0	0	1	0	0	
2,2,3	1	0	1	0	0	0	0	0	
3,1,1	1	0	0	1	1	0	0	0	
3,1,2	1	0	0	1	0	1	0	0	
3,1,3	1	0	0	1	0	0	0	0	
3,2,1	1	0	0	0	1	0	0	0	
3,2,2	1	0	0	0	0	1	0	0	

Patterns of Dependencies

Log-linear models are used to determine the pattern of associations among a set of variables. Imagine a 3-dimensional table, $R \times C \times L$. Then the simplest model we could have would be

$R \perp C \perp L$ - mutual independence

We write the corresponding model as (R) (C) (L) which is an abbreviated version of

$$\log(F_{ij}) = \alpha + \beta_i^R + \beta_j^C + \beta_k^L$$

The model (R) (C) (L), or R+C+L as it is also sometimes written, embodies a hypothesis about the relationship between R, C, and L and we test that hypothesis using the likelihood ratio χ^2 test, where we compare the fitted values implied by this model with the observed values. The statistic we use is the familiar minus twice the difference in the log-likelihoods of the models being compared. In this case the likelihood ratio test reduces to a simple formula:

$$G^2 \text{ or } L^2 \equiv 2 \sum_i \sum_j f_{ij} \log \frac{f_{ij}}{F_{ij}}$$

G^2 follows a chi-squared distribution with the usual degrees of freedom.

Conditional independence:

$$R \perp C \mid L \Rightarrow (RL)(CL).$$

Notice that this notation is hierarchical: thus $(RL)(CL) \Rightarrow$

$$\log(F_{ijk}) = \alpha + \beta_i^R + \beta_j^C + \beta_k^L + \beta_{ik}^{RL} + \beta_{jk}^{CL} \leftarrow \text{odds ratio } \theta_{CL} \text{ same at all levels of R}$$

↑ odds ratio θ_{RL} same at all levels of C

Could also have conditional independence conditioning on other variables.

Pairwise association: (RC)(RL)(CL).

How do we know whether conditional independence or pairwise association is a better model? Take the difference in their G^2 s and this has a χ^2 distribution with df equal to the difference in the number of parameters in the two models.

So if this were a $3 \times 3 \times 3$ table the number of extra parameters in the 'all pairwise association' model compared with conditional independence would correspond to the extra term $B_{ij}^{rc} \Rightarrow 4$ parameters.

This test is only valid if the models are nested (as is usual with likelihood ratio test).

Saturated model, (RCL) which adds, to the 'all pairwise association' model, the three way interaction term: β_{ijk}^{RCL} .

Notice that this model fits the data exactly and has $G^2=0$

It also has zero d.f. because, given a 3^3 table, it fits

1 overall effect
 6 \times main effects
 12 \times 2-way interactions
8 \times 3-way interactions
 27

What is a three way interaction term: β_{ijk}^{RCL} ?

An important point is that any margin that appears implicitly in a log-linear model is fitted exactly.

So, to go back to the conditional independence model, (RL)(CL)

This fits exactly the joint RL margin of the table and also the RC margin, as well as the univariate margins, R, C & L.

Examples (1)

Basic IEM instructions using Danish mobility example. {example_1.inp}

Model search

Stepwise search procedures are sometimes used to try to find a simple model that adequately represents the data. I don't recommend them unless you split your data, one part on which you carry out the search, the other part on which you validate the model that the search found.

But this is how it would work for a 4-way table.

There are 4 basic models:

- (0) Saturated model: (ABCD);
- (1) All-3-way interactions model: (ABC) (ABD) (ACD) (BCD);
- (2) All 2-way interactions model: (AB) (AC) (AD) (BC) (BD) (CD);
- (3) All 1-way effects: (A) (B) (C) (D).

Step I: find the best fitting of these four models by testing

$$(G_{m+1}^2 - G_m^2) \sim \chi^2, (df_{m+1} - df_m) \text{ starting at } m=0.$$

If this is not significant for $m=0$ proceed to $m=1$. Repeat and stop when this comparison is significant.

Step II: Suppose model (2) is best (in this sense that (1) is not a better fit and (3) is a worse fit). Then see if it can be improved by adding terms from model (1) - that is, 3-way interactions. Do this by testing

$(G_2^2 - G_{2a}^2) \sim \chi^2, (df_2 - df_{2a})$ where 2a means the augmented model 2. Notice that there are many possible models 2a.

Step III: having found the best model that lies between 2 and 1, see if there are any 2-way terms that are not part of a 3-way term in the model that can be removed, keeping their one-way components. Test

$$(G_{2d}^2 - G_{2b}^2) \sim \chi^2, (df_{2d} - df_{2b})$$

where 2b means the best model up to this point, and 2d is that model with some 2-way associations removed.

In this way you could end up with a simpler version of the all two-way interactions model, such as (AB)(BD)(CD) or a more complex model, like (ABC)(D).

Measures of goodness of fit

Bayesian information criterion, or *bic*.

$$bic = G^2 - df \times \log(N)$$

Bic is an approximation to minus twice the logarithm of the odds that, given the data, the model in question is true relative to the saturated model. So if the model in question is more likely to be true than is the saturated model, *bic* will return a negative value. The most preferred model is the one with the largest negative value of *bic*.

Bic works by penalizing the addition of an extra parameter by increasing the statistic by $\log(N)$ (because you lose one df). For $N > c. 50$ there can be cases in which the more complex model would be accepted using the likelihood ratio test but rejected (in favor of something simpler) by *bic*. So using *bic* alone makes us choose simpler models that capture the big picture - which or may not be what we want.

The index of dissimilarity:

$$\Delta = \frac{1}{2N} \sum_i \sum_j |f_{ij} - F_{ij}|$$

Can be interpreted as the proportion of cases that would have to change cells in order to give the observed and fitted frequencies the identical distribution.

Modeling Association

As well as modeling dependencies among variables log-linear models are also used to model the nature of the associations among variables.

$$\log(F_{ij}) = \alpha + \beta_i^R + \beta_j^C + \beta_{ij}^{RC}$$

This model would be saturated except that we now try to model the $R \times C$ association in a more parsimonious way by placing constraints on the β_{ij}^{RC} terms.

We will look at some common models.

① Quasi-independence (QI) or Quasi-perfect mobility (QPM):

$$\begin{aligned} \beta_{ij}^{RC} &= \beta_{ij}^{RC} \text{ if } i=j; \\ \beta_{ij}^{RC} &= 0 \text{ otherwise.} \end{aligned}$$

R and C are independent except that we fit a parameter to each of the cells on the main diagonal of the table. For a 4×4 table we could express this as

$$\begin{array}{cccc} 2 & 1 & 1 & 1 \\ 1 & 3 & 1 & 1 \\ 1 & 1 & 4 & 1 \\ 1 & 1 & 1 & 5 \end{array}$$

Where each number represents a distinct association parameter, and, in this case, 1 indicates $\beta_{ij}^{RC} = 0$.

② Constrained version of QI:

$$\begin{aligned} \beta_{ij}^{RC} &= \beta \text{ if } i=j; \\ \beta_{ij}^{RC} &= 0 \text{ otherwise.} \end{aligned}$$

$$\begin{array}{cccc} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{array}$$

The implied patterns of local log odds ratios:

	QI				CQI		
	1,2	2,3	3,4		1,2	2,3	3,4
1,2	$\beta_2 + \beta_3$	$-\beta_3$	1	1,2	$-\beta$	1	$-\beta$
2,3	$-\beta_3$	$\beta_3 + \beta_4$	$-\beta_4$	2,3	$-\beta$	1	$-\beta$
3,4	1	$-\beta_4$	$\beta_4 + \beta_5$	3,4	1	$-\beta$	1

③ Same ideas can be extended to off-diagonal cells too, as in the model of quasi-symmetry (QS): $\beta_{ij}^{RC} = \beta_{ij}^{RC}$

1	2	3	4
2	1	5	6
3	5	1	7
4	6	7	1

Quasi-symmetry, not symmetry, because the F_{ij} are not symmetric. For that we would need marginal homogeneity (MH): $\beta_i^R = \beta_j^C$ when $i=j$.

$$\text{MH} + \text{QS} = \text{Symmetry}$$

How many parameters are fitted by QS? Ans = $\frac{IJ-I}{2} = \frac{I(J-1)}{2}$ in this case $\frac{4(3)}{2} = 6$. One feature of QS is that although we do not need to specify separate parameters for each cell on the main diagonal, these cell frequencies are fitted exactly.

④ The parameters need not be symmetric - topological or levels models (Hauser *Social Forces* 1978).

Featherman and Hause (1978) model for association in 5x5 occupational mobility table:

2	4	5	5	5
3	4	5	5	5
5	5	5	5	5
5	5	5	4	4
5	5	5	4	1

This should be derived from theory about mobility patterns.

© Multiple parameters for cells

Mare and Schwartz (2005) crossings effects in educational assortative marriage:

		<i>Husband's years of education</i>				
<i>Wife's educ</i>	<u><10</u>	<u>10-11</u>	<u>12</u>	<u>13-15</u>	<u>≥16</u>	
<10	1	2	2+3	2+3+4	2+3+4+5	
10-11	2	1	3	3+4	3+4+5	
12	2+3	3	1	4	4+5	
13-15	2+3+4	3+4	4	1	5	
≥ 16	2+3+4+5	3+4+5	4+5	5	1	

Each number identifies a different β^{rc} $\beta_1^{rc} = 0$ and so there are 4 others that capture how unlikely it is to observe a marriage crossing a certain number of educational differences.

Examples (2)

Danish mobility example. {example_2.inp through example_4.inp}

Educational homogamy data { example_5.inp}

Association models

Goodman (1979) introduced association models where rows and columns are scored and the association or interaction parameters look like:

$$\beta_{ij}^{RC} = \beta x_i y_j \text{ where } x_i \text{ and } y_j \text{ are the row and column scores respectively.}$$

Use these when row and or column variables are ordinal rather than nominal.

When x_i and y_j are known, and they are equally spaced, this model is called uniform association (UA). Why? First we see that the scores for each cell ($= x_i \times y_j$) are proportional to:

	1	2	3	4
1	1	2	3	4
2	2	4	6	8
3	3	6	9	12
4	4	8	12	16

And so when we multiply by the association parameter we get

$$\therefore \beta_{ij}^{rc} = \begin{array}{cccc} \beta & 2\beta & 3\beta & 4\beta \\ 2\beta & 4\beta & 6\beta & 8\beta \\ 3\beta & 6\beta & 9\beta & 12\beta \end{array}$$

And the local log odds ratios are therefore all equal to β .

So local log odds ratios $\neq 0$ (unless $\beta=0$) but they are all same.

If X and Y are not evenly spaced, then $\log \theta$ is proportional to the distance between rows and columns.

RC models

Suppose X and Y are not known: then we have the following model which is log-multiplicative rather than log-linear:

$$\beta_{ij}^{RC} = \phi \mu_i \nu_j$$

Here, μ and ν are now estimated, rather than given, and conventionally the association parameter is called ϕ rather than β .

So all three quantities need to be estimated

$$\log(\theta_{ii',jj'}) = \phi[(\mu_i - \mu_{i'})(\nu_j - \nu_{j'})]$$

So log odds ratios depend on ϕ and on the distances between rows and between columns.

These models are very parsimonious. UA uses only one more df than independence. RC $1 + (I-1) + (J-2)$

RC can be made more parsimonious by constraining the row and column scores to be the same.

It is possible to fit models with more than one set of scores and thus multiple association parameters.

Examples (3)

Danish mobility example. {example_6.inp }

These models work better when rows and columns are clearly ordinal:

Socio-economic status and mental health data {example_7.inp }

Multidimensional Tables

Consider the saturated model for the three-way table:

$$\log(F_{ijk}) = \alpha + \beta_i^R + \beta_j^C + \beta_k^L + \beta_{ij}^{RC} + \beta_{ik}^{RL} + \beta_{jk}^{CL} + \beta_{ijk}^{RCL}$$

Often our concern is with a particular pairwise association, say RC, and how it varies over L . So we want a model for RC, and a model for how it varies over L , but we do not care much about RL or CL and so we fit them exactly. So now we focus on β_{ij}^{rc} and β_{ijk}^{rc} .

Examples: how does OD association vary over countries, birth cohort etc?
 how does gender gap in education vary by race?

Notice that if we use dummy variable coding, β_{ij}^{rc} tells us RC association at level 1 of L ($k=1$) and β_{ijk}^{rc} tells us the difference between the RC association at level k (>1) and at level 1.

Simplest approach is to take a model for RC association and let all its parameters vary across levels of L .

For example, consider a 4×4 table in which we model the RC association by QS.

QS uses 6 parameters in a 4×4 table \therefore 18 in the three tables. Could compare each of the 6 parameters across the three tables

$$\begin{array}{ccc} \beta_{11} & \beta_{12} & \beta_{13} \\ \beta_{21} & \beta_{22} & \beta_{23} \\ : & & : \\ \beta_{61} & & \beta_{63} \end{array}$$

The interaction effects, β_{ijk}^{rc} , may actually report

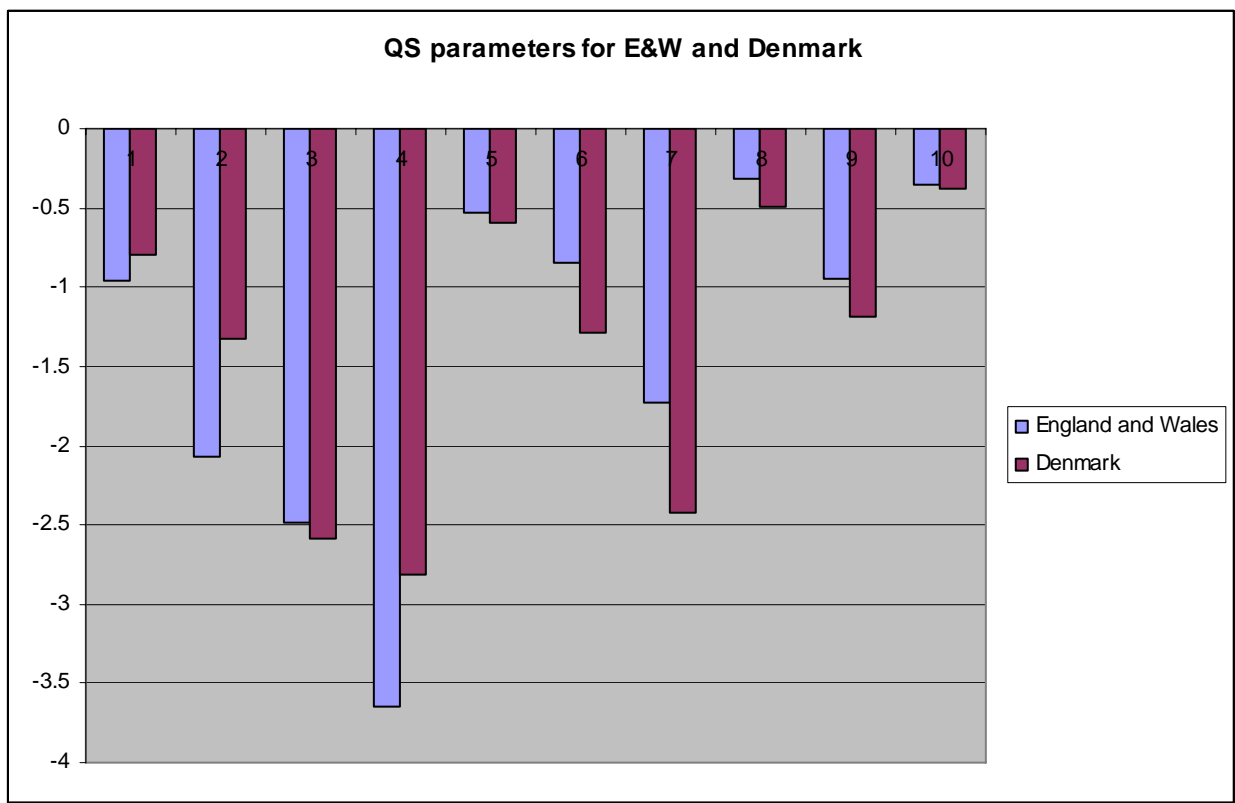
$$\begin{array}{ccc} \beta_{12} - \beta_{11} & \beta_{13} - \beta_{11} & \\ \beta_{22} - \beta_{21} & \beta_{23} - \beta_{21} & \text{etc} \end{array}$$

So tests of the sign of the interaction effects tests whether this difference is significant or not. (But IEM does not do it this way!)

Examples (4)

English and Danish mobility example. {example_8.inp }

Could allow only some of the parameters to differ over L and keep others constant.



Could apply the same approach to UA: then we would have K values of β_k .

Could apply it to RC models and then we could vary, across L, any or all of φ, μ or ν .
For example: $\varphi_k \mu_i \nu_j$.

Log-multiplicative layer effect model

Very widely used model (Xie 1992; Erikson and Goldthorpe 1992 “Unidiff”).

$$\ln \theta_{ijk} = \beta_k \ln \theta_{ij}$$

Conventionally, $\beta_1 = 1$: that is, this scale parameter is set to unity for one of the tables, and so in this case the fitted log odds ratios are simply equal to the baseline set.

$\beta_k > 1$: stronger RC association in the k^{th} table than in the reference table:

$\beta_k < 1$: weaker RC association in the k^{th} table than in the reference table.

Very flexible and parsimonious model for comparing across tables (esp. used in social mobility research) but rests on the assumption of common pattern of local odds ratios.

Could replace β_k , $k = 1, \dots, K$. with a linear trend instead:

$$\ln \theta_{ijk} = [1 + \beta \times (k - 1)] \ln \theta_{ij}$$

Intercept Slope

Here the tables are scaled or scored 0 to K-1 but we could replace this with some other table level covariate such as gdp/capita if each table represented a country.

Could model θ_{ij} itself in different ways. Often the set of θ_{ij} is based on the saturated model, but one could fit, say, QS to model θ_{ij} and then ‘unidiff’ the QS parameters over L.

Examples (5)

English and Danish mobility example. {example_9.inp }

Multivariate Unidiff

Breen and Luijkx (2004) introduced an extension of the log-multiplicative layer effect model by considering the case of a 4 way table in which pairwise associations between two of the variables are allowed to evolve log-multiplicatively over the other two.

Their example was a mobility table in which the OD association varies over period and countries.

Possible formulations are:

$$\ln \theta_{ijkl} = \beta_{kl} \ln \theta_{ij} \quad \text{country} \times \text{period -specific parameter}$$

$$\ln \theta_{ijkl} = \beta_k \beta_l \ln \theta_{ij} \quad \text{log-additive specification of cohort and period effects}$$

This is useful for testing trends: e.g. does the trend across periods persist when we allow for the trend across cohorts? (Breen and Jonsson 2007 *AJS*).

Goodman and Hout (1998 *SM*) extended the model in another way:

$$\ln \theta_{ijk} = \beta_k^1 \ln \theta_{ij}^1 + \beta_k^2 \ln \theta_{ij}^2$$

So there are two sets of local log odds ratios which evolve log-multiplicatively over k.

All these are easy to fit in IEM, although the latter model has not been widely used (Vallet 2004 is one example).

Log-linear and logit models

$\log(F_{ij}) = \alpha + \beta_i^R + \beta_j^C + \beta_{ij}^{RC}$ - this is a log linear model

Now consider

$$\begin{aligned} \log(F_{2j}) - \log(F_{1j}) &= \log\left[\frac{F_{2j}}{F_{1j}}\right] \\ &= \alpha + \beta_2^R + \beta_j^C + \beta_{2j}^{RC} - \alpha - \beta_1^R - \beta_j^C - \beta_{1j}^{RC} \\ &= [\beta_2^R - \beta_1^R] + [\beta_{2j}^{RC} - \beta_{1j}^{RC}] \end{aligned}$$

If we had used dummy variable coding this would reduce to:

$\log\left[\frac{F_{2j}}{F_{1j}}\right] = \beta_2^R + \beta_{2j}^{RC}$ and this is a logit model for the log odds of being in R=2 rather than R=1.

Had we used centered coding we would have

$$\log\left[\frac{F_{2j}}{F_{1j}}\right] = 2\beta_2^R + 2\beta_{2j}^{RC} - \text{and this is a logit model.}$$

The equivalence is general and holds for all cases including when the dependent variable has more than two categories - multinomial logit.

Suppose we have the log-linear model (AB) (AC) (BC) and we take A as the dependent variable: this corresponds to:

Logit [A] = linear/additive function of B and C

By contrast, the log-linear model (ABC) allows B and C to interact in their effect on A:

Logit [A] = multiplicative function of B and C

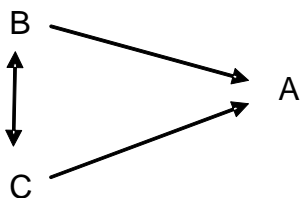
So log-linear and logit are formally equivalent - *but only if we include, in the log-linear model, the parameters needed to fit exactly the margins of the explanatory variables.*

In this case we must include (BC) in the model.

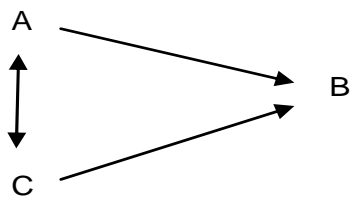
(ABC)(AD)(BCD) is equivalent to the model: $\text{logit}[A] = f(B*C + D)$, but (ABC)(AD)(BC)(BD)(CD) is not a logit model for A.

Log-linear systems of equations

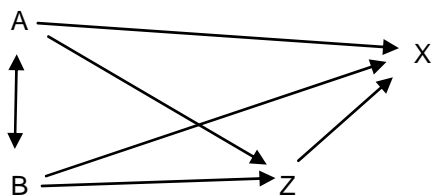
We could draw a figure to correspond to the logit model derived from (AB)(BC)(AC) as follows:



But the following logit model could also be derived from (AB)(BC)(AC):



Suppose we wanted to estimate more than one equation as in:



In the model for X we would need to fit (ABZ) exactly – yet the model for Z says that A and B have additive effects (ZA) (ZB), not interactive (ABZ).

We circumvent this as follows. We consider two tables: ABZ and ABZX. We use the first table to estimate the model for Z, which we can write $Z | AB$: in this case (ZA)(ZB)(AB). We use the second table to estimate $X | ABZ = (XA)(XB)(XZ)(ABZ)$.

In IEM this is easy:

mod $Z | AB \{ZA ZB\}$ * no need to include AB because in this notation it is automatically included

$X | ABZ \{XA XB XZ\}$ * ABZ is automatically fitted.

LEM gives parameter estimates for both models and fitted values for the ABZX table on which it bases goodness of fit statistics. How does it compute the fitted values?

We can write the joint distribution as follows:

$$p(ABZX) = p(X | ABZ)p(ZAB)$$

The first model is used to calculate $p(ZAB)$ and the second to calculate $p(X | ABZ)$ and their product yields the fitted values for the whole table. Notice that $X | ABZ$ tells us how X is distributed given ABZ (under the model) and $Z | AB$ tells us how ABZ is distributed, under the model.

A common application is in mobility research: $E | OC$ and $D | OEC$.

The two equations are assumed independent - there is no unobserved component that influences both X and Z.

This is usually called a log-linear path model.

Panel data with families observed as being in poverty or not at each wave. Call the variables A (poor or not at wave 1), B (poor or not at wave 2), etc though to F (wave 6). This gives the table/ joint probability distribution, for ABCDEF.

A reasonable model for this might be a first order Markov process:

```
mod A
  B | A
  C | B
  D | C
  E | D
  F | E
```

We could relax the Markov assumption:

```
mod AB
  C | AB
  D | BC
  E | CD
  F | DE
```

In /EM we can do many nice things: e.g. make the original model time homogenous:

```
mod A
  B | A
  C | B eq1 B | A
  D | C eq1 B | A
  E | D eq1 B | A
  F | E eq1 B | A
```

Can even set the transition probabilities to specific values, constrain certain of them to be equal, and so forth.

Examples (6)

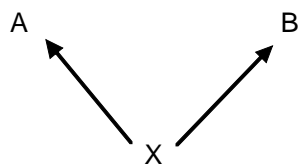
Markov process over 4 waves {example_10.inp }

Path model using Sewell and Shah data {example_11.inp }

PART 2: Latent Class Models

Imagine two categorical variables, A and B , that are associated, (AB) . There may exist a variable X , such that $A \perp B \mid X$ i.e. A and B are conditionally independent given X . This implies that the model $(AX)(BX)$ would fit the three-way ABX table as well as $(AX)(BX)(AB)$.

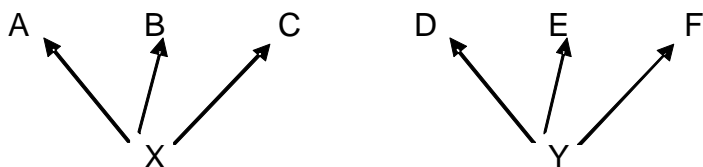
Diagrammatically:



The idea of latent class models is to construct a variable X such that A and B are conditionally independent. X is then said to be a latent variable. Since these variables are all categorical we think of X as categorical with T categories or classes. In the simplest case X is a latent variable with $T=2$ classes (since $T=1$ amounts to saying A and B are unconditionally independent).

This generalizes to cases with more than two observed or manifest variables. Given $(AB)(BC)(AC)$ we look for an X such that $(AX)(BX)(CX)$ provides as good a fit to the ABC table.

We might also have more than one latent variable.



In this case ABC is independent of DEF and both display what is called 'local independence' meaning that they are independent of each other given the latent variable.

Analogy with latent variable models as used in measurement models. In this case too a latent variable is often described as being a true underlying measure which gives rise to observed indicators.

For example, A and B are 5-category responses by husbands and wives to a question about the frequency of marital sex. Then we could think of X in (A|X)(B|X) as the true frequency distribution of marital sex and A and B as fallible indicators.

Another example: A, B and C are three measures of a person's health and so X in (A|X)(B|X)(C|X) could be some latent or true underlying health status of the person. A, B and C are unconditionally associated because, and only because, they each tap the same underlying dimension of health status.

Estimation of Latent Class Models

Observed data is the $A \times B$ table. We want to estimate $(XA)(XB)$ (so A and B must not be independent to begin with).

We treat this as a missing data problem: the true data is the $A \times B \times X$ table but the data on X are missing. Estimation is via the E-M algorithm which yields ML estimates.

Suppose A and B both have J categories (but same dimension is not necessary) and suppose X has two classes.

Step 1: Take the original data (a single 5×5 table) and randomly allocate the frequencies across two 5×5 tables, one for each value of X ;

Step 2: Fit the model $(XA)(XB)$ to these data;

Step 3: Derive the fitted values from step (2) for the AB table by collapsing the fitted values over X ;

Step 4: Weight the data in $ABX=1$ and $ABX=2$ tables by the ratio of their fitted values divided by the AB fitted values.

Keep repeating steps 2 through 4 until the fitted frequencies stop changing.

This is easy to program but often slow to converge. But there's no need because *IEM* and other packages have this built in.

We judge how well $(AX)(BX)$ fits the data by collapsing the fitted values over X and comparing the implied AB frequencies with the observed AB frequencies.

Parameters and degrees of freedom

The number of parameters fitted is exactly the same as if X had been observed. So in the case where $I = J = 5$ and $T = 2$ we have

$1 + 4 + 4 + 1$ (intercept and main effects of A, B and X)

$+4 + 4$ (the XA and XB terms)

$= 18$.

But we only have J^2 observations in the original data, so the model has $25-18=7$ df, whereas, had X been observed, we would have had 32df.

So this limits how large T (the number of latent classes) can be.

The parameters for X are β_i^X (main effects) and $\beta_{ii}^{XA}, \beta_{ij}^{XB}$ (interactions). Together they tell us (or we can derive from the fitted values) the marginal distribution of X (i.e. $p(X=1)$ and $p(X=2)$) and also the probability of an observation taking a specific value of A or of B or of their combination, given latent class membership: $p(A=i | X=t)$.

We can use Bayes' rule to calculate: $p(X=t | A=i)$.

Identification

No results about this except that identification requires that $df \geq 0$.

Need to check identification empirically via the eigenvalues of the information matrix from the model.

Even if the model passes this test this only proves local identification – ie. we might be at a local rather than a global maximum of the log-likelihood. Therefore we should always estimate the model several times using different starting values.

Notation

$$\log(F_{ijkt}) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C + \beta_t^X + \beta_{it}^{AX} + \beta_{jt}^{BX} + \beta_{kt}^{CX}$$

or

$$F_{ijkt} = \mu \tau_i^A \tau_j^B \tau_k^C \tau_t^X \tau_{it}^{AX} \tau_{jt}^{BX} \tau_{kt}^{CX}$$

Sometimes:

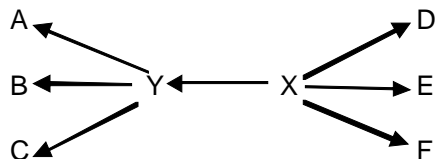
$$p_{ijkt} = \eta \tau_i^A \tau_j^B \tau_k^C \tau_t^X \tau_{it}^{AX} \tau_{jt}^{BX} \tau_{kt}^{CX} \text{ where } \eta = \mu / N$$

Examples (7) Basic latent class models in IEM

Single latent variable {example_12.inp } Sewell and Shah data

Latent class path models

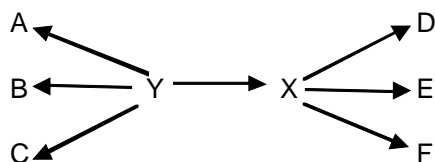
Same as path models we saw earlier but with latent variables.



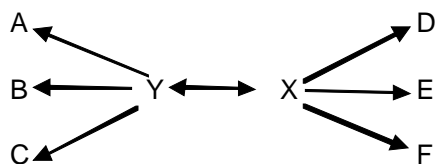
If A through F are manifest variables and X and Y are latent then this is a categorical counterpart to the well known MIMIC model.

We could write it as: $(AY)(BY)(CY)(DX)(EX)(FX)(XY)$

And its goodness of fit is tested against the observed ABCDEF table. Notice that this model is indistinguishable from



(which is a different SEM); or from



Which is not a SEM at all but a pure measurement model (we saw it a moment ago but with X and Y independent).

The general point is that any model (whether it contains latent constructs or not) is a hypothesis that we test against the data and against competing models.

Examples (7)

Latent Markov process over 5 waves {example_13.inp, example_14.inp }

Ashford and Snowden data {example_15.inp}

Latent Markov with multiple indicators {example_16.inp}

Latent class models as discrete approximations: mixture models

Lindsay (1983a, b) showed that any form of unobserved heterogeneity can be arbitrarily well approximated by a finite number of latent classes. So latent classes can be used in many places where other ways of modeling unobserved heterogeneity might be used.

For example, a random effects (RE) logit model is usually written:

$$p(Y = 1) = \int \frac{\exp(x'\beta)}{1 + \exp(x'\beta)} f(\beta_0) d\beta_0$$

Here β_0 is the first (intercept) element of the parameter vector β . So this is a random intercept model. If we integrate over the whole β vector we would have the general RE specification.

But this can be arbitrarily well approximated by:

$$p(Y = 1) = \sum_{t=1}^T \pi_t \left[\frac{\exp(x'\beta)}{1 + \exp(x'\beta)} \right]$$

Where π_t is the proportion in the t^{th} latent class and there are T classes.

Latent class models are widely used to model unmeasured heterogeneity especially when we do not know what parametric form it might take: e.g. Heckman and Singer (1982).

Examples (9)

Event history model with unmeasured heterogeneity {example_17.inp}

Recently we (Breen and Luijkx) applied this approach to cumulative probability models.

Let Y be a dependent ordinal variable with categories indexed $j=1,\dots,J$, and X a vector of explanatory variables. We write the probability that the value of Y for the i^{th} observation, y_i , is less than or equal to j , given X , as $\gamma_j(x_i)$. There exists a family of statistical models that sets $g(\gamma_j(x_i)) = t_j - \beta'x_i$ (McCullagh 1980), where g is a link function (such as the logit) that maps the $(0,1)$ interval into $(-\infty, \infty)$ and t_j are a set of thresholds.

Our basic model also explicitly models the scale of the distribution so our models start from

$$g(\gamma_j(x)) = (t_j - \beta'x_i) / \tau(x_i)$$

Our most general model is then

$$g(\gamma_j(x) | k) = \frac{t_j - \beta'x_i - \alpha_k}{\tau_k}$$

Where k denotes latent class membership.

One special case of this model is:

$$g(\gamma_j(x) | k) = \frac{t_j}{\tau_k} - \beta'x_i - \alpha_k$$

Which we apply on data about attitudes to pre-marital sex taken from the BSA in the early 1980s.

Examples (9)

Mixture model using BSA data {example_18.inp}

