Appendix I: Loglinear Analysis

Loglinear analysis is used to analyze the relationship between three or more categorical variables. Unlike, for example, linear regression in loglinear analysis there is no dependent variable that is predicted. Instead, it is the cell frequencies that are predicted.

We saw in Chapter 8 that it can be very difficult to interpret contingency tables when analysing the association between three or more variables, for the number of possible models that could fit the data rises alarmingly, and the best fitting model cannot easily be determined by sifting through a series of separate Chi Square tests.

Loglinear analysis was developed to cope with this problem. It allows us to interpret in a systematic way the relationships between a number of categorical variables by fitting imaginary models of relationships between them in an attempt to represent the patterns in the data. The aim is to find the simplest model that fits the data satisfactorily such that the predicted cell frequencies are not significantly different from the observed cell frequencies.

MODELS AND EFFECTS

In any crosstabulation the cell frequencies are the result of a number of *effects.* In a two-way crosstabulation, for example, there are three effects that between them determine the individual cell values:

- The *grand mean*. This is the average cell frequency count.

- The *main effect* of each variable. The frequencies in each cell of a contingency table are constrained by the row and column marginal totals. If we look at Table I:1, for example, we can see that 158 people smoked (the second row marginal), and 139 people reported that their mothers had smoked (the second column marginal). These figures clearly set limits on the possible cell count in cell D (it could not possibly exceed 139, for example). Indeed, we saw in Chapter 8 that it

is possible to calculate from these marginal totals the number of cases we would expect to find in cell D, if there were no association between the two variables (i.e. the null hypothesis model). The effect that each of the variables has in determining the cell values is called the *main effect.*

- The *association effect* between the variables. This is the contribution that the association between the variables makes to each cell frequency. In the example above, the strength of association between being a smoker and having friends who smoke determines how far the actual (observed) frequency in cell E diverges from the expected frequency. We saw in Chapter 8 that the strength of this effect can be measured through odds ratios.

If we construct a model that includes all three of these effects, we end up with what is called a *saturated* model. Because all of the effects that determine the cell frequencies are included, our model will exactly reproduce the observed cell counts.

This, however, is completely pointless! We already know the frequencies in each cell. What we want to know is which effects are having the greatest impact on these frequencies, and which are having only a tiny influence. The task, in other words, is to find a *simplified model* – one that removes as many effects as possible while still coming close to predicting all the cell frequencies accurately (i.e. predicted cell frequencies should not be significantly different from observed cell frequencies, even though we have simplified the model).

We could for example, try removing the association effect between the two variables. We would then be left with a model that includes just the effects of the grand mean and the main effects. If we found that this model still fitted the observed data adequately, we might then decide to try to simplify the model even further by removing, say, both of the main effects so that the only effect we were left with was the grand mean. This radically simplified model would predict individual cell counts purely from the overall number of cases (i.e. it would predict that the observed frequencies in each of the cells would be equal to the average cell count) – an unlikely result in practice.

Loglinear modelling is simply a procedure that calculates the coefficients for all of the effects that explain the observed cell values, and then seeks the model with the fewest effects that fits the observed data.

Estimating the parameters for the effects

In loglinear analysis the first step is to calculate the contribution that each effect makes in accounting for the observed cell frequencies. The logic involved in this procedure is a bit like working out the relative quantities of each ingredient needed to make a cake.

Imagine that we wanted to produce a similar cake to the one that had been purchased from the local cake shop. We might know that the cake consists of five ingredients: water, flour, chocolate, eggs and sugar. However, before we can make the cake we need to know the exact quantities required of each of the ingredients.

Once we have calculated these quantities and made an accurate copy of the cake, we might decide to have a go at making a cake that tastes and looks the same but involves fewer ingredients. This is akin to the next step in loglinear modelling, but let us not run ahead of ourselves.

- The equivalent of the cake in loglinear analysis is the contingency table
- The equivalent of the ingredients are the effects
- The equivalent to the quantities of the ingredients are the coefficients for the effects.

Just as we know the ingredients in our cake, so we know what the effects are that determine the cell counts in a crosstabulation. If we want to remove one of these effects, we must first calculate the contribution that it makes to the cell counts so that this can be taken out of the model when predicting cell counts. In other words we must know the coefficient for each effect.

Calculating coefficients for effects in a bivariate table

With a bivariate crosstabulation, calculating the effects is straightforward. Taking the crosstabulation between respondents' smoking behaviour and mothers' smoking behaviour as our example (see Table I:1), the grand mean effect, the two main effects, and the association effect are calculated as follows:

Table I:1: The observed frequency counts for the crosstabulation of respondents' smoking behaviour by mothers' smoking history

Smoked in last week? * MUMSMOKE Crosstabulation

Count

		MUMSMOKE		
		No	Yes	Total
Smoked in last week?	no	A 109	B 63	172
	yes	C 82	D 76	158
Total		191	139	330

- The *grand mean effect* is the average cell count based on the geometric mean. To obtain this we take the fourth root of the product of the cell counts:

$$\sqrt[4]{(\text{cell A} \times \text{cell B} \times \text{cell C} \times \text{cell D})}$$

This calculation can be simplified by taking its natural logarithm.

Logarithms have the advantage of transforming mathematical functions into simpler mathematical functions. For example, multiplication of two natural numbers is achieved by simple addition of their logs, and division of natural numbers is similarly accomplished in logarithms by simple subtraction. Logging thus transforms products into sums and divisions into subtractions.

If we do this, the equation becomes:

$$\text{Log (grand mean)} = 0.25(\log \text{cell A} \times \log \text{cell B} \times \log \text{cell C} \times \log \text{cell D}) =$$

Log (grand mean) = 0.25(4.69 + 4.41 + 4.14 + 4.33) = 4.39

- The *main effect* for *mumsmoke* on cell A is the product of cells A and B divided by the product of cells C and D. Thus:

$$\sqrt[4]{(\text{cell A} \times \text{cell B}) / (\text{cell C} \times \text{cell D})}$$

The log transformation of this equation simplifies the equation to:

Log (main effect of *mumsmoke* for cell A)
$= 0.25(\log \text{cell A} + \log \text{cell B} - \log \text{cell C} - \log \text{cell D}) = 0.156$

This is the main effect of *mumsmoke* for cell A, where the value of *mumsmoke* is 'no'. The main effect of this variable for cell B, where the value is yes, will be the same as for cell A, but with the opposite signs between the logs.

- The main effect for *var00003* (whether the respondent smoked in the last week) for cell A is calculated in much the same way:

$$0.25(\log \text{cell A} - \log \text{cell B} + \log \text{cell C} - \log \text{cell D}) = 0.024$$

- The *association effect* is the fourth root of the odds ratio of the two variables:

$$\sqrt[4]{(\text{cell A} \times \text{cell D}) / (\text{cell B} \times \text{cell C})}$$

The log transformation of this equation simplifies the equation to:

Log (association effect for cell A)

$= 0.25(\log \text{cell A} - \log \text{cell B} - \log \text{cell B} + \log \text{cell D}) = 0.118$

If we add all of these effects together we obtain the log cell frequency for cell A:

Log (cell A) = 4.39 (grand mean effect) + 0.156 (main effect for MUMSMOKE) + 0.0347 (main effect for var00003) + 0.118 (association effect) = 4.688

If we take the anti log of this in order to 'unlog' it we obtain:

Cell A = $e^{4.688} = 109$

What we have seen here is that through some relatively straightforward (although somewhat lengthy) calculations we have been able to calculate the contribution that each effect makes to the overall cell count. These contributions are referred to as the coefficients for each effect.

Calculating coefficients of effects with more than two variables

Unfortunately, it becomes far more difficult to calculate coefficients when we move to the analysis of the relationship between three or more variables because the number of effects increases as the number of variables is added.

- With two variables we only had to consider the grand mean effect, the main effects for two variables and the single association effect.

- With three variables, there are three main effects (one for each variable), three association effects (one for each pair of variables) and one interaction effect (the effect that the three-way relationship between the variables has on the observed cell count).

There are no simple formulae that can be used to calculate all of these effects. Instead, SPSS uses a method known as *iterative proportional scaling* which is an

algorithm similar in nature to that used for calculating the coefficients in logistic regression.

In simple terms, the procedure works by the 'try-it-out-and-see' method of making a series of progressively better guesses at the coefficients for each of the effects required to produce the cell frequencies for the saturated model (i.e. the observed frequencies). Eventually the algorithm will end when the observed cell frequencies can be estimated accurately from the coefficients for each of the effects.

Why 'loglinear'?

In loglinear analysis the cell frequencies that are predicted are logged. There are two good reasons for doing this:

- One is that it makes calculations easier. Logarithms turn multiplication into sums. In principle, it would be possible to calculate the coefficients for the effects that predict the cell counts directly (i.e. without logging), but as we saw when calculating the coefficients, an additive model makes things easier and a lot more manageable when trying to predict the cell frequencies.

- Secondly, logging the cell frequencies ensures that the model predicting the cell frequencies is an *additive* one. This parallels other statistical techniques such as multiple regression and ANOVA, for these linear techniques are, of course, additive.

Because it is an additive model, the equation for the loglinear model closely resembles the multiple regression equation. The equation predicting the log cell frequency for cell A ($Log(A)$) in the two way crosstabulation in Table I:1, for example, is:

$$Log(A) = \mu + \mu_1^A + \mu_1^B + \mu_{11}^{AB}$$

μ is the log of the grand mean. It is a constant that applies to all cell values.

μ_1^A is the log of the main effect for non-smoking respondents

μ_1^B is the log of the main effect for non-smoking mothers

μ_{11}^{AB} is the log of the effect of the association between non-smoking respondents and non-smoking mothers

On the right hand side of the equation are the logged effects that predict the logged frequency count. The μ (mean parameter) terms are analogous to the a and b terms in the linear regression equation (Y = a + bX). They represent all that we need to know to predict the cell frequency, for if we add all of the coefficients for all of these effects together then we obtain the predicted frequency for that cell. In fact, in the last section where we calculated the coefficients for the two-way crosstabulation we used this equation to calculate the cell frequency for cell A.

MODELS AND MODEL FITTING WITH THREE VARIABLES

Once we have calculated all of the coefficients, we can set about the task of model fitting. The aim, remember, is to find the simplest possible model which estimates cell frequencies that are not significantly different from the observed frequencies.

In a two-way categorical analysis, we test only one model (the no association model, based on the null hypothesis). With three tables, however, the number of possible models that we might test increases because the number of effects increases. In fact there are 19 models of relationships that could fit the data with three variables! The difficulty involved in manually selecting the simplest one of these models that best fits the data is the primary reason for using loglinear analysis.

Table I:.2 Some of the models of relationships between variables that can be fitted to data with three variables

Model	Model description	Corresponding loglinear equation
A*B*C	Interaction (saturated)	$\log(Fa) = \mu + \mu_1^A + \mu_1^B + \mu_1^C + \mu_{11}^{AB} + \mu_{11}^{AC} + \mu_{11}^{BC} + \mu_{11}^{ABC}$
A*B + B*C + A*B	No interaction, pairwise association	$\log(Fa) = \mu + \mu_1^A + \mu_1^B + \mu_1^C + \mu_{11}^{AB} + \mu_{11}^{AC} + \mu_{11}^{BC}$
A*B	Pairwise association between A and B only	$\log(Fa) = \mu + \mu_1^A + \mu_1^B + \mu_1^C + \mu_{11}^{AB}$
A+B+C	No interaction and no association	$\log(Fa) = \mu + \mu_1^A + \mu_1^B + \mu_1^C$

Four of the models that might fit the data in a three-variable loglinear analysis are shown in Table I:2. For each of models that are fitted to the data there is a corresponding loglinear equation.

- The first model that is fitted is the saturated model where all 3 variables interact with each other (commonly written as A*B*C). This model mirrors the extent of the interaction exactly as in the data because it includes the effects of all the terms in calculating the expected value. The fact that all of the effects are included can be seen in the loglinear equation for the model which includes the effects of the grand mean(μ), the main effects of each variables ($\mu_1^A, \mu_1^B, \mu_1^C$), each of the pairwise association effects ($\mu_{11}^{AB}, \mu_{11}^{AC}, \mu_{11}^{BC}$) and the three-variable interaction effect (μ_{11}^{ABC}). Since all of the effects have been carried over into the model table, the cell frequencies are identical to the observed frequencies and the table will obviously fit the observed data perfectly.

- The next model is pairwise association between A and B, B and C and A and C, with no interaction between the three variables. This model is fitted to the data without the effect of the interaction term μ_{11}^{ABC} being included but all the other effects are included.

- The next model is pairwise association between A and B only. To predict the cell values with this model, we include the effect of the association between variables A and B, plus the main effects of each of the three variables.

- The last model in Table I:2 is the model of no association (A+B+C). This includes all of the main effects but does not include the pairwise associations or the interaction term.

Loglinear analysis involves trying to fit these and other models to the data in order to find the best fitting, most parsimonious model. The model fitting procedure in SPSS does this by starting with the saturated model (which fits perfectly). It then tries to simplify by removing the interaction term (i.e. fitting the pairwise association model). If this fits it tries to fit an even simpler model with one of the associations removed. It continues removing the effects until it finds a model that does not fit the data satisfactorily. It then returns to the last model that did fit satisfactorily, and this is our final model.

Selecting the best fitting model using measures of significance

If you are primarily concerned with exploring relationships in the data, it is often convenient to let the computer select the best fitting model automatically. However, if you are testing hypotheses about the relationships between variables, you should choose the best fitting model manually. Your theory might dictate, for example, that a variable, pair of variables or interaction must stay in the model even if this effect is not significant. Similarly, you might decide to choose a slightly less parsimonious model than the one the computer selects because it fits the data substantially better.

To see how manual model fitting works, let us take an example from Chapter 8, where we ended up running a three-way cross tabulation between smoking behaviour and mother's smoking history, controlling for gender. Our results suggested that there may be an interaction between these three variables – in particular, that the relationship between respondents' smoking behaviour and mothers' smoking

behaviour depends upon their gender. It seemed that males were more likely to smoke if their mothers had smoked, but this was not true for females.

Using loglinear modelling, we can now analyze these patterns of association much more rigorously. Instead of the three-way smoke*mumsmoke*gender relationship, for example, we might want to test whether the three pairwise associations of smoke*mumsmoke, smoke*gender and mumsmoke*gender fit the data without the need for the interaction term.

To determine whether the cell frequencies predicted by each model adequately fit the observed data, we use a test, not dissimilar to the Chi Square test, called the log-likelihood ratio (often referred to as G^2). Just as with chi square, we can derive the statistical significance from its value once the number of degrees of freedom is taken into account.

A perfectly fitting model would have a G^2 of 0 and 100 per cent significance (because the model and the data are identical). On the other hand, a high G^2 and a low level of significance (for example, p = 0.01) represents a poor fitting model and should therefore be rejected. What we want is the simplest model that is not significantly different from the observed data. By 'not significantly different' what is conventionally meant is a model with a significance level of 0.05 or more.

Table I:3 shows how well a range of models fit the observed relationship between respondents' smoking behaviour, mothers' smoking behaviour and gender. The models are ordered in a hierarchy starting with the most complex model first (the saturated model), followed by successively simpler models where the model cell frequencies are estimated using fewer and fewer effects:

.

- The best fitting model is obviously the saturated model (G^2 of 0 and 100 per cent significance).

- However, the simpler model that excludes the interaction effect also fits the data satisfactorily with a significance level of 0.150.

- If we simplify the model further by removing the association between mothers' smoking behaviour and respondents' gender (model 3), the fit of the model actually improves.

- Just meeting the conventional criteria of a fitting model are the alternative two-way associations (models 4 and 5). However, these models are no simpler than model 3, and the fit is worse.

- Models 6 to 8 predict the cell frequencies using the parameters for one of the pairwise associations and the main effect for the third variables. These all fail to predict the observed data accurately ($p < 0.05$). Similarly, the no interaction model (model 9), where only the main effects of the three variables are included, also fails to fit the data satisfactorily ($p < 0.05$).

Table I:3: A table showing the goodness of fit of a range of models fitted to the observed data of the relationship between respondents' smoking behaviour, mothers' smoking history and gender

Model	DF	G^2	Significance
1. A*B*C	0	0	1
2. A*B + A*B + B*C	1	2.1	0.150
3. A*C + A*B	2	3.4	0.181
4. A*B + B*C	2	5.9	0.051
5. A*C + B*C	2	5.9	0.051
6. A*C + B	3	7.9	0.049
7. A*B + C	3	7.9	0.049
8. B*C + A	3	10.5	0.016
9. A + A + C	4	12.3	0.015

Key:
A = Respondents' smoking behaviour
B = Mothers' smoking behaviour
C= Respondents' gender

Now we have to make a choice over which model to choose. There are six that meet the level of significance we set, or are very close to it. Models 6 and 7 are the simplest models, using the fewest parameters to predict the cell frequencies, but the fit of these models is not very good. Models 4 and 5 are slightly more complex in that they include two pairwise associations. In spite of this, the fit of these models is

barely any better than models 6 and 7. Model 3, on the other hand, is no more complex than models 4 and 5, and only slightly more complex than models 6 and 7, yet it fits the data much better than any of these models (G^2 of 3.4 with 2 df and 0.181 per cent significance). Therefore, model 3 is the best fitting model using the fewest set of parameters.

Our conclusion, therefore, is that a model including just two associations (between respondents' smoking behaviour and the respondents' gender, and between respondents' smoking behaviour and mothers' smoking behaviour) fits the data satisfactorily. In contrast in Chapter 8, we do not need gender to mediate the relationship between the respondents' smoking behaviour and their mothers' smoking behaviour..

Residuals

It is important when fitting models to remember that that the log-likelihood ratio only measures the overall fit of the expected cell counts with the observed cell counts. Therefore, it is always a good idea to examine the cell-by-cell residuals to see how well the model fits for each cell, for there may be one or two cells where the model fits the data poorly. In SPSS these residuals are given in standardized form and, in general, the convention is that any residual larger than 2 suggests the expected model cell counts do not fit the observed data well.

Table I:4: A crosstabulation showing the standardized residuals for the model fitted to the observed data

Smoked in last week? * MUMSMOKE * gender

gender				MUMSMOKE No	MUMSMOKE Yes	Total
female	Smoked in last week?	no	Count	59	35	94
			Expected Count	59.6	34.4	94.0
			Residual	-.6	.6	
			Std. Residual	-.1	.1	
		yes	Count	41	27	68
			Expected Count	35.3	32.7	68.0
			Residual	5.7	-5.7	
			Std. Residual	1.0	-1.0	
	Total		Count	100	62	162
			Expected Count	94.9	67.1	162.0
male	Smoked in last week?	no	Count	50	28	78
			Expected Count	49.4	28.6	78.0
			Residual	.6	-.6	
			Std. Residual	.1	-.1	
		yes	Count	41	49	90
			Expected Count	46.7	43.3	90.0
			Residual	-5.7	5.7	
			Std. Residual	-.8	.9	
	Total		Count	91	77	168
			Expected Count	96.1	71.9	168.0

The standardized residuals for Table I:4 refer to the pairwise association model – A*C+A*B – that was found to be the best fitting model. We can see that all of the model cell counts fall safely within 2 standardized residuals of the observed counts. However, we can also see that the model expects more male smokers with mothers who do not smoke to be smokers than the observed data shows, and it also expects fewer female smokers with mothers who do not smoke to smoke than actually do smoke. This suggests that there is *some* evidence that gender does affect the association between respondents' smoking behaviour and that of their mothers. However, this effect is not large enough to warrant the rejection of the simplified model 3 (in Table I:3) in favour of model 1 (where the effect of gender on the association between respondent smoking and mother's smoking is included).

Measuring the strength of effects

We saw in Chapter 8 how the results of a Chi Square test are influenced by sample size. The same is true of the G^2 test in loglinear modelling. Had our sample size been twice as large, model 3 would not have fitted the data adequately, and the only model that we could have chosen would have been the model of interaction (confirming our speculation in Chapter 8).

With a large sample size, even relatively small effects (for example, weak associations or weak interactions) enter the model. One way of dealing with this problem is to report the coefficients for each of the effects as well as the best fitting model. The coefficients tell us the impact on the cell frequencies of each effect and are not affected by sample size. For our example the coefficients are as follows:

Grand mean (μ) = 3.686

Main effect for whether the respondent smokes(μ_1^A) = 0.0324

Main effect for whether the mother smoked (μ_1^B) = 0.1655

Main effect for whether gender (μ_1^C)= -0.0255,

The association effect between respondent smoke and mum smoke (μ_{11}^{AB}) = 0.1067

The association effect between respondent smoke and gender(μ_{11}^{AC}) = 0.1214

The association effect between mum smoke and gender(μ_{11}^{BC}) = 0.0665

The interaction between all three variables (μ_{111}^{ABC}) = -0.0804

What each of these coefficients is doing is informing us by how much each of the effects is increasing or decreasing the predicted cell value. The larger the coefficient, the larger the impact of the effect on the cell frequencies. For example, the grand mean logged cell value is 3.686; adding the coefficient main effect of variable A increases the predicted cell value by log 0.0324; adding the coefficient main effect of variable B increases the predicted cell value by 0.1655; and so on.

Comparing the relative strength of these effects, we can see in our example that the most important effect (disregarding the grand mean) in predicting the cell frequency is

the main effect for smoking behaviour of mothers (0.1654), followed by the association between respondents' smoking behaviour and gender (0.1214). We can also see why the effects for the association with gender and the three-variable interaction were both left out of the model, for the coefficients are both relative small (0.0665 and –0.0804 respectively).

MODELLING MORE THAN THREE VARIABLES

Fitting models with more than three variables follows exactly the same procedure as we have covered so far. The only difference is that you will have more models to select from.

For example, with four variables the most complex model (i.e. saturated model) is the four way interaction model – A*B*C*D**.** Less complex models include three three-way interactions (A*B*C + A*B*D + B*C*D) and two three-way interactions with an association (e.g. A*B*C + A*B*D + B*C). As you can see, with four variables there are many, many possible models to choose from. However, the logic of finding the simplest model that fits satisfactorily follows what has been described earlier.

RUNNING A LOGLINEAR PROCEDURE IN SPSS

The loglinear procedure can be found in the '*Analyze*' menu bar under 'Loglinear'. There are three loglinear procedures that you can choose from but you are likely to find the 'Model Selection' procedure the most useful. To run the procedure you must select the categorical variables you wish to analyze and define the range of the values for each of them. Unless you select specific models that you wish to fit to the data, when you run the procedure it will automatically select the best fitting model for you. If you wish to examine the coefficients, these can be obtained by ticking the box next to 'Parameter estimates' in the 'Options' dialog box.