

## Loglinear analysis

### *An alternative loglinear example with medical data*

People with a particular variant of the FTO gene (the abbreviation derives from an observation of fused toes and other abnormalities in mice following a deletion in a homologous area) have increased risk of being obese (30% increase with one copy; 70% increase with two copies). There is also a causal relationship between obesity and Type 2 Diabetes. One hundred adults with and 100 without the FTO variant are recruited and presence or absence of Diabetes Type 2 is recorded, together with the presence or absence of obesity, based on a body mass index (BMI) cutoff of 30. The data appear in Table 12.1.

### *Entering the data in SPSS*

We have already warned you that factor levels must be given numerical codes if you want to do a loglinear analysis in SPSS, so don't code the diabetics and non-diabetics as D and N, or obesity as Yes and No. Once you have the data represented with suitable codes, there are two ways to present it to SPSS. If you already have the cell frequencies, you can enter these in the SPSS data window as in Table 12.3, which we have set out for the data in Table 12.1.

Table 12.1  
*Frequencies for three factors each at two levels*

	Layer 1: not obese				Layer 2: obese		
	diabetic	not diabetic	totals		diabetic	not diabetic	totals
	1	2			1	2	
variant FTO 1	5	9	14	variant FTO 1	29	10	39
normal FTO 2	17	54	71	normal FTO 2	49	27	76
<b>totals</b>	<b>22</b>	<b>63</b>	<b>85</b>	<b>totals</b>	<b>78</b>	<b>37</b>	<b>115</b>

Table 12.2  
*Two-way table showing FTO variant by diabetes*

	diabetic 1	not diabetic 2	totals
variant FTO 1	34	19	53
normal FTO 2	66	81	147
<b>totals</b>	100	100	200

Table 12.3  
*Presenting a frequency table to SPSS (med.loglinear.sav)*

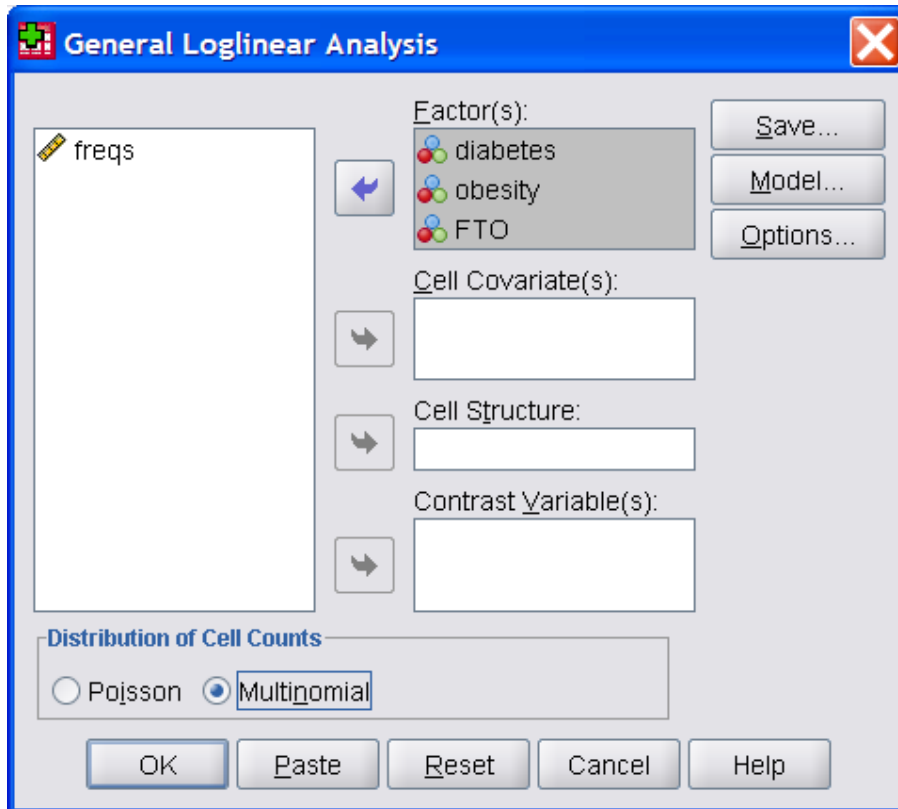
diabetes	obesity	FTO	freqs
1	1	1	5
1	1	2	17
1	2	1	29
1	2	2	49
2	1	1	9
2	1	2	54
2	2	1	10
2	2	2	27

The first three columns show all eight combinations of factor levels, and the column of frequencies gives the cell frequency for each combination of factor levels. To use these frequencies we select **Data** from the menu bar, then **Weight Cases**. A small dialog box opens; click the radio button **Weight cases by**, then enter FREQS into the **Frequency Variable** box using the arrow. When you save the data file, the weighting will be saved with it, so you only need to do this the first time you use the frequency table.

If instead of the layout in Table 12.3 you have each case listed with the levels of each factor for the case, you can enter the data in the usual way, with each case making one row, and each factor or variable making one column. SPSS will calculate the cell frequencies when the analysis is requested.

### *A loglinear model: requesting the analysis in SPSS*

Whichever way the data were entered, select **Analyze** from the menu bar, then **Loglinear**, then **General...**, and SPSS Dialog Box 12.1 appears. Enter the factors using the arrow, and click the radio button to select **Multinomial**. The other boxes are left empty as we only have factors in this problem.



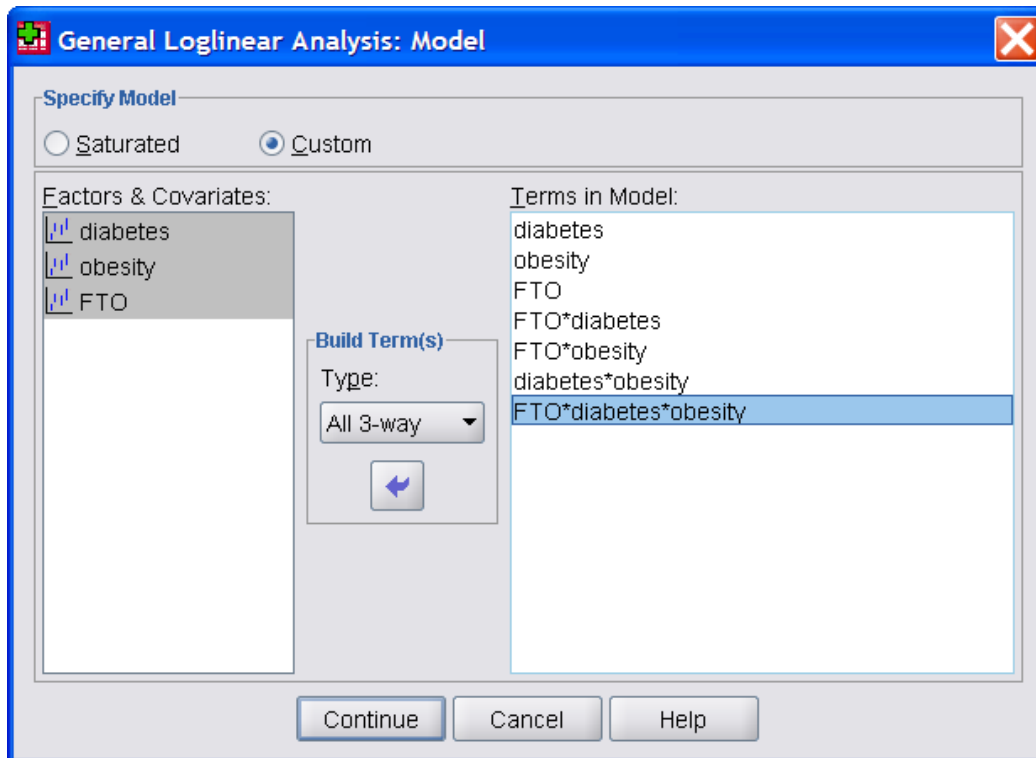
*SPSS Dialog Box 12.1. Starting a loglinear analysis*

The multinomial model is the one to choose if we took one or more random samples and classified the cases according to one or more variables or factors. This is usually how data are obtained. We took two samples, one each of diabetics and non-diabetics. We could have used exactly the same methods if we had taken one sample and classified the cases as diabetic or non-diabetic as well as on the other two variables,

but the sampling method used does affect which terms have to be included in any model.

Click the **Model** button to get SPSS Dialog Box 12.2 and define the model. The *saturated* model includes all main effects and all interactions. This is equivalent to estimating a separate probability for every cell. Usually we are interested in a model with fewer terms than this, and then we must click the **Custom** radio button. You can add terms one at a time or in groups using the arrow and the drop-down menu below it. For instance we can select **Main effects** on the menu, select all three factors and click the arrow to enter all main effects. Then select **All 2-way** from the menu, select all the factors again and click the arrow to get all two-factor interactions entered. In Dialog Box 12.2 you can see all the terms of the saturated model entered. We could just have used the **Saturated** radio button, but we shall want a series of **Custom** models later, and it is convenient to start with the list of terms as we have them here.

The model corresponding to complete independence of the three factors is that with only the main effects. However, we are usually interested in models that fall somewhere between complete independence and the saturated model.



*SPSS Dialog Box 12.2. defining the model*

Click **Continue** and then the **Options** button from Dialog Box 12.1. We can accept all the defaults here except that we want the **Estimates** of the effects, so click that box in the **Display** group. Now click **Continue** and **OK** to get the analysis.

*The loglinear analysis: understanding the output for the saturated model*

The output for the saturated model begins with a warning: all residuals are zero so no charts will be created. This is because we have estimated separate probabilities for every cell in the table, so the model is a perfect fit with no residuals.

Next a data information table (not shown here) tells us that 8 cases are accepted but 200 weighted cases are used in the analysis (this is because we entered our data as frequencies), 8 cells are defined and no cases are rejected because of missing data.

Then there is a list of the variables (factors) and the number of levels for each. Three

more tables follow which are not shown here as none of them is very interesting for the saturated model. The convergence information table shows that 5 iterations out of a possible 20 were sufficient: the saturated model is a perfect fit so the process of fitting is not difficult. The goodness of fit tests show that the fit is perfect: both goodness of fit statistics are zero. The table of cell counts and residuals again just repeats the information that the saturated model is a perfect fit: all cell frequencies fitted by the model are just the actual cell frequencies, so all residuals are zero.

Next comes a list of the parameters, shown as SPSS Output 12.1. At the bottom of the list is a note saying that all those parameters with a superscript  $b$  are set to zero because they are redundant. This is because, as in ANOVA, all effects are measured from an overall mean, so when the probability is known for one of only two levels, the value for the other level is known as well. We show how to calculate expected cell values from the parameter estimates later. Other notes at the bottom of the table remind us that we used the multinomial model and all main effects and interactions are included. Each of the main effects and interactions has one non-zero term (estimate) in the model (because each of the three factors has just two levels), and for each of these terms a  $Z$  value, a confidence interval and a significance level is given in the table.

Parameter Estimates<sup>c,d</sup>

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	3.314 <sup>a</sup>					
[diabetes = 1]	.588	.238	2.471	.013	.122	1.054
[diabetes = 2]	0 <sup>b</sup>					
[obesity = 1]	.684	.234	2.924	.003	.226	1.142
[obesity = 2]	0 <sup>b</sup>					
[FTO = 1]	-.963	.363	-2.654	.008	-1.674	-.252
[FTO = 2]	0 <sup>b</sup>					
[diabetes = 1] * [FTO = 1]	.445	.431	1.033	.302	-.399	1.290
[diabetes = 1] * [FTO = 2]	0 <sup>b</sup>					
[diabetes = 2] * [FTO = 1]	0 <sup>b</sup>					
[diabetes = 2] * [FTO = 2]	0 <sup>b</sup>					
[obesity = 1] * [FTO = 1]	-.784	.505	-1.552	.121	-1.774	.206
[obesity = 1] * [FTO = 2]	0 <sup>b</sup>					
[obesity = 2] * [FTO = 1]	0 <sup>b</sup>					
[obesity = 2] * [FTO = 2]	0 <sup>b</sup>					
[diabetes = 1] * [obesity = 1]	-1.724	.363	-4.744	.000	-2.436	-1.012
[diabetes = 1] * [obesity = 2]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 1]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 2]	0 <sup>b</sup>					
[diabetes = 1] * [obesity = 1] * [FTO = 1]	.144	.740	.195	.846	-1.307	1.595
[diabetes = 1] * [obesity = 1] * [FTO = 2]	0 <sup>b</sup>					
[diabetes = 1] * [obesity = 2] * [FTO = 1]	0 <sup>b</sup>					
[diabetes = 1] * [obesity = 2] * [FTO = 2]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 1] * [FTO = 1]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 1] * [FTO = 2]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 2] * [FTO = 1]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 2] * [FTO = 2]	0 <sup>b</sup>					

a. Constants are not parameters under the multinomial assumption. Therefore, their standard errors are not calculated.

b. This parameter is set to zero because it is redundant.

c. Model: Multinomial

d. Design: Constant + diabetes + obesity + FTO + diabetes \* FTO + obesity \* FTO + diabetes \* obesity + diabetes \* obesity \* FTO

*SPSS Output 12.1. Estimates of the parameters (terms in the model)*

We see that the Z values (column 4) for the third order interaction term and the terms for OBESITY\*FTO and DIABETES\*FTO are all less than 2 in absolute value, which suggests that the corresponding model terms may not be needed (1.96 is the 5% critical value for Z). We could try omitting all these three terms from the model, but we will use a more systematic approach and begin by omitting just the three-factor interaction. The final two tables in the output window (not shown here) give the correlations and covariances of the estimates.

## Selecting a reduced model

### Removing the three-factor interaction

If we return to SPSS Dialog Box 12.2 we can remove the three-factor interaction from the model using the arrow. We repeat the analysis and look through the output until we find the Goodness of fit Statistics (shown in SPSS Output 12.2(a)). The less familiar one based on the likelihood ratio is sometimes called the Goodman statistic, and it has the advantage that it can be partitioned in a way that enables us to test hypotheses about terms in a hierarchical set of models. Like the Pearson statistic, it has an approximate  $\chi^2$  distribution if the model is a good fit. In practice there is usually little difference between the two values. In our example, using either of these  $\chi^2$  tests, we would not reject the hypothesis that the model is a good fit.

#### (a) Goodness of fit of model with all two-factor interaction terms

Goodness-of-Fit Tests<sup>a,b</sup>

	Value	df	Sig.
Likelihood Ratio	.017	1	.897
Pearson Chi-Square	.017	1	.896

a. Model: Multinomial

b. Design: Constant + diabetes + obesity + FTO + diabetes \* FTO + obesity \* FTO + diabetes \* obesity

#### (b). Goodness of fit of model with FTO\*OBESITY and DIABETES\*OBESITY

Goodness-of-Fit Tests<sup>a,b</sup>

	Value	df	Sig.
Likelihood Ratio	1.979	2	.372
Pearson Chi-Square	1.999	2	.368

a. Model: Multinomial

b. Design: Constant + diabetes + obesity + FTO + obesity \* FTO + diabetes \* obesity

#### SPSS Output 12.2. Goodness of fit of two models

### Removing a two-factor interaction

Now we would like to see if the model can be further simplified by removing one or more of the two-factor interactions. Looking at the new list of parameter estimates (not shown here), we find that the two-factor interaction with the smallest  $Z$  value is

that for FTO\*DIABETES, so we try removing this one first. (In SPSS Output 12.1, the FTO\*DIABETES interaction is also the two-factor interaction with the smallest Z value.) So, omitting FTO\*DIABETES, using the arrow in SPSS Dialog Box 12.2, we search down the output for the Goodness of fit Statistics and find SPSS Output 12.2(b). Again we would not reject the null hypothesis that the fit is good. The Z values for all remaining terms exceed 2.

### *The conditional independence model*

This is the model for the conditional independence of FTO and DIABETES within each level of OBESITY. If we try further reductions by removing either of the other two-factor interactions we find that the goodness of fit statistics cause us to reject the models as poorly fitting, so we settle for the model with all main effects and the interactions FTO\*OBESITY and DIABETES\*OBESITY. In this model both DIABETES and FTO are each associated with OBESITY.

We now have two models that both fit well. We are usually interested in finding a model with fewer terms than the saturated one, but which still fits well. In fact, we usually only want to use a more complex model if including the extra term(s) provides a significant improvement in the fit. Equivalently, we only want terms in the model if they differ significantly from zero. We now show how to test hypotheses about terms in a nested set of hierarchical models where a simpler model contains a subset of the terms from a complex model with which it is being compared.

### *Testing the hypothesis that the three-factor interaction effect is zero*

The saturated model contains all main effects and interactions. The reduced model with all two-factor interactions is nested within it since it contains a subset of these

parameters. To test the hypothesis that the three-factor interaction is zero, find the difference between the  $\chi^2$  statistic for the likelihood ratio for the reduced model with all two-factor interactions and that for the saturated model. For the saturated model, the goodness of fit statistics are zero (since the fit is perfect), and there are no degrees of freedom for them. So the difference between  $\chi^2$  statistics is  $0.017 - 0 = 0.017$  (SPSS Output 12.2(a)). This is approximately  $\chi^2$  with degrees of freedom equal to the difference between the degrees of freedom for the  $\chi^2$  in the two models ( $1 - 0 = 1$ ). Since our value is much less than the critical value of 3.84 for a test at the 5% level, we do not reject the null hypothesis that the three-factor interaction is zero.

*Testing the hypothesis that the FTO by DIABETES interaction effect is zero*

The model without the FTO\* DIABETES interaction is nested within that with all two-factor interactions, since it contains a subset of the parameters in the model with all two-factor interactions. To test the hypothesis that the FTO\* DIABETES interaction is zero, find the difference between the  $\chi^2$  statistic for the likelihood ratio for the model without this term (SPSS Output 12.2(b)) and the model with all two-factor interactions (SPSS Output 12.2(a)). This is  $1.979 - 0.017 = 1.962$ , and is approximately  $\chi^2$  with degrees of freedom equal to the difference between the degrees of freedom for the  $\chi^2$  in the two models ( $2 - 1 = 1$ ). Since our value is much less than the critical value of 3.84 for a test at the 5% level, we do not reject null hypothesis that the FTO\* DIABETES interaction is zero.

*Testing several interactions together*

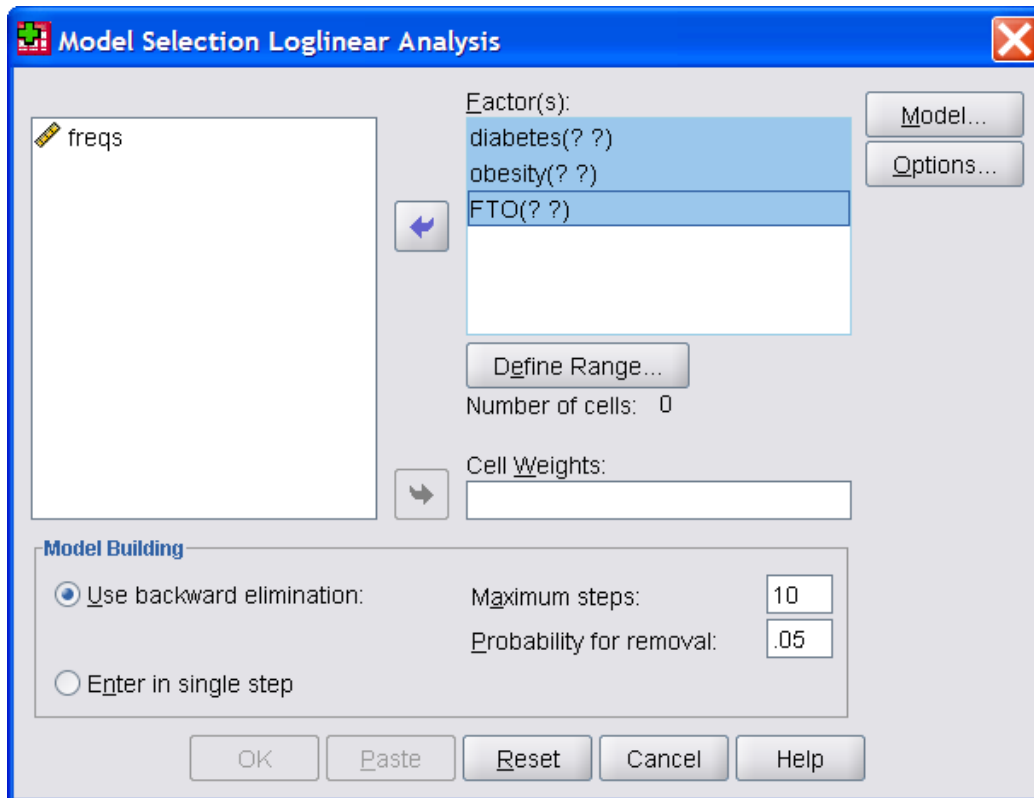
It is possible to test a hypothesis that several parameters are zero. For instance, we could test the hypothesis that the three-factor interaction and the FTO\* DIABETES

interaction are both zero. To do this, use the  $\chi^2$  statistic for the likelihood ratio for the model without these two terms (SPSS Output 12.2(b)) and for the saturated model (which is zero). This gives us  $1.979 - 0 = 1.979$  on  $2 - 0 = 2$  degrees of freedom, so we do not reject the hypothesis. We could have gone to our selected model in one step instead of two.

### *Automating model selection*

#### *Requesting a backward elimination loglinear analysis*

You can get SPSS to automate the process of working down from the saturated model to a reduced model that fits well. Use the menu commands **Analyze**, **Loglinear** and **Model Selection** to get SPSS Dialog Box 12.3. After using the arrow to enter the factors we must define the range for each. Selecting each factor in turn, click the **Define Range** button, fill in the **Minimum** and **Maximum** values (ours are all 1 and 2) for that factor in the small dialog box that opens.



SPSS Dialog Box 12.3. Getting SPSS to do the model selection

Make sure the radio button **Use backward elimination** is selected for **Model Building**.

*The backward elimination analysis: understanding the model evaluation output*

The process of selection will start from the saturated model, then try removing the three-factor interaction. The goodness of fit will be checked, then the hypothesis that the third order interaction is zero will be tested. If the fit is good and the hypothesis not rejected, the two-factor interaction which contributes least to the goodness of fit will be removed. The fit will again be checked and the hypothesis that the interaction is zero will be tested. The process continues until no more terms can be removed from the model without significantly worsening the fit. The process is exactly the same as

the one we went through in the previous two sections, and the result is the same. The history is shown in a table, the first in SPSS Output 12.3. The *generating class* just means the set of highest order interactions in the model: because models are hierarchical, the presence of any interaction implies the presence of all related lower order interactions and main effects. The changes in goodness of fit statistics are also shown for each model as the interaction effects are removed one at a time. For instance, if you look at the row of the table for Step 1, Deleted Effect 2, you see the change in  $\chi^2$  of 1.962 which we derived in the previous section. For the original saturated model and for the final one we have a small table like the second one shown in SPSS Output 12.3, which is for the final model, the same one that we arrived at using **Analyze, Loglinear and General....**

Step Summary						
Step <sup>a</sup>		Effects	Chi-Square <sup>c</sup>	df	Sig.	Number of Iterations
0	Generating Class <sup>b</sup>	diabetes*obesity*FTO	.000	0	.	
	Deleted Effect 1	diabetes*obesity*FTO	.017	1	.896	4
1	Generating Class <sup>b</sup>	diabetes*obesity, diabetes*FTO, obesity*FTO	.017	1	.896	
	Deleted Effect 1	diabetes*obesity	31.694	1	.000	2
	2	diabetes*FTO	1.962	1	.161	2
	3	obesity*FTO	4.052	1	.044	2
2	Generating Class <sup>b</sup>	diabetes*obesity, obesity*FTO	1.979	2	.372	
	Deleted Effect 1	diabetes*obesity	35.569	1	.000	2
	2	obesity*FTO	7.927	1	.005	2
3	Generating Class <sup>b</sup>	diabetes*obesity, obesity*FTO	1.979	2	.372	

- a. At each step, the effect with the largest significance level for the Likelihood Ratio Change is deleted, provided the significance level is larger than .050.
- b. Statistics are displayed for the best model at each step after step 0.
- c. For 'Deleted Effect', this is the change in the Chi-Square after the effect is deleted from the model.

**Convergence Information<sup>a</sup>**

Generating Class	diabetes*obesity, obesity*FTO
Number of Iterations	.000
Max. Difference between Observed and Fitted Marginals	.000
Convergence Criterion	.250

- a. Statistics for the final model after Backward Elimination.

*SPSS Output 12.3. Backward elimination history and model summary for the final model in an automated loglinear model fitting*

There is also a table of expected values for the final model and we discuss this in the next section.

*The backward elimination analysis: the expected values table*

Observed and expected values are given in the Cell Counts and Residuals table shown in SPSS Output 12.4(a). Here we use the version obtained from **Analyze, Loglinear and General...**, where we asked for adjusted and deviance residuals. You can see that DIABETES is used for the layers: this is because it was entered first in the list of factors (SPSS Dialog Box 12.1). You can control the layout of the table of expected values by altering the order in which factors are entered. Enter OBESITY first if you want it to form the layers like in Table 12.1.

*(a) Observed and expected values for well-fitting model*

**Cell Counts and Residuals<sup>a,b</sup>**

			Observed		Expected		Residual	Standardized Residual	Adjusted Residual	Deviance
diabetes	obesity	FTO	Count	%	Count	%				
diabetes	not obese	FTO variant	5	2.5%	3.624	1.8%	1.376	.730	.919	1.794
		FTO normal	17	8.5%	18.376	9.2%	-1.376	-.337	-.919	-1.627
	obese	FTO variant	29	14.5%	26.452	13.2%	2.548	.532	1.074	2.309
		FTO normal	49	24.5%	51.548	25.8%	-2.548	-.412	-1.074	-2.229
no diabetes	not obese	FTO variant	9	4.5%	10.376	5.2%	-1.376	-.439	-.919	-1.601
		FTO normal	54	27.0%	52.624	26.3%	1.376	.221	.919	1.670
	obese	FTO variant	10	5.0%	12.548	6.3%	-2.548	-.743	-1.074	-2.131
		FTO normal	27	13.5%	24.452	12.2%	2.548	.550	1.074	2.314

a. Model: Multinomial

b. Design: Constant + diabetes + obesity + FTO + diabetes \* obesity + obesity \* FTO

*(b) Parameter list and estimates for well-fitting model*

Parameter Estimates<sup>c,d</sup>

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	3.197 <sup>a</sup>					
[diabetes = 1]	.746	.200	3.736	.000	.355	1.137
[diabetes = 2]	0 <sup>b</sup>					
[obesity = 1]	.766	.223	3.439	.001	.330	1.203
[obesity = 2]	0 <sup>b</sup>					
[FTO = 1]	-.667	.197	-3.387	.001	-1.053	-.281
[FTO = 2]	0 <sup>b</sup>					
[diabetes = 1] * [obesity = 1]	-1.798	.318	-5.652	.000	-2.421	-1.174
[diabetes = 1] * [obesity = 2]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 1]	0 <sup>b</sup>					
[diabetes = 2] * [obesity = 2]	0 <sup>b</sup>					
[obesity = 1] * [FTO = 1]	-.956	.353	-2.713	.007	-1.647	-.265
[obesity = 1] * [FTO = 2]	0 <sup>b</sup>					
[obesity = 2] * [FTO = 1]	0 <sup>b</sup>					
[obesity = 2] * [FTO = 2]	0 <sup>b</sup>					

- a. Constants are not parameters under the multinomial assumption. Therefore, their standard errors are not calculated.
- b. This parameter is set to zero because it is redundant.
- c. Model: Multinomial
- d. Design: Constant + diabetes + obesity + FTO + diabetes \* obesity + obesity \* FTO

*SPSS Output 12.4. Observed and expected values, and parameter list and estimates for well-fitting model*

*The backward elimination analysis: parameter estimates*

The table of estimates for the terms in the model with FTO and DIABETES conditionally independent is shown in SPSS Output 12.4(b). These parameter estimates are analogous to the sizes of effects in ANOVA. Their relationship to the expected frequencies is hard to comprehend if you do not have a mathematical background and it will be okay to skip over the rest of this section and rejoin us at 'Residuals and their plots'. For those readers who are strong on mathematics or are masochists, however, we include an attempt to show how expected frequencies are obtained from the parameter estimates.

*Obtaining expected frequencies from parameter estimates*

To see how the expected values relate to the parameter values, look first in SPSS Output 12.4(a) at the cell where all factors are at level 2 (no diabetes present, obese, normal FTO). The expected count for this cell is given as 24.452. From the list of parameter values for the model (see note d below SPSS Output 12.4(b) for the model

summary) we see that for this cell (all terms that include a factor at level 2), all parameter values in Output 12.4(b) are zero (because they are redundant, indicated by the superscript <sup>b</sup>). That leaves the overall mean (the constant), which has the value 3.197, so the expected value for this cell is the natural log of the constant; i.e.,  $\exp(3.197) = 24.5$ . If you want to check this on your calculator, look for a button labelled 'ln' and above it will probably be  $e^x$  in a colour indicating that it is a *second function* on the calculator. Select the second function mode, then ln, and enter the value (on some calculators, the sequence will be different; e.g., enter the value, select mode (sometimes labelled 'inverse'), then select ln).

If OBESITY = 1 and DIABETES and FTO are both 2, only the main effect for OBESITY and the constant (the non-redundant parameters) are needed, so the expected value is  $\exp(3.197 + 0.766) = 52.6$ , as indicated in row 4 within Table 12.4. The other expected values and the parameter combinations needed are also shown in Table 12.4.

Table 12.4  
*Parameter combinations to give expected values*

DIABETES	OBESITY	FTO	exp of these terms are computed	expected frequency
1	1	1	$3.197+0.746+0.766-0.667-0.956-1.798$	3.6
1	1	2	$3.197+0.746+0.766-1.798$	18.4
1	2	1	$3.197+0.746-0.667$	26.5
1	2	2	$3.197+0.746$	51.6
2	1	1	$3.197+0.766-0.667-0.956-$	10.4
2	1	2	$3.197+0.766$	52.6
2	2	1	$3.197-0.667$	12.6
2	2	2	3.197	24.5

### *Residuals and their plots*

The residuals are the differences between the observed and expected cell frequencies. Adjusted residuals and deviance residuals are both derived from the raw residuals because they each have the useful property that, for a well fitting model, they will be approximately Normal provided the sample size is not too small. So, plotting either of

these against the expected values should give a shapeless cloud of points if the model fits well. Patterns in the residuals may give a clue about how a model does not fit. In a well-fitting model, none of the residuals should be much larger than the rest (in absolute value). Any residual with an absolute value larger than about 3 (roughly corresponding to the 1% value for a standard Normal distribution) suggests that the model does not provide an adequate fit for that point.

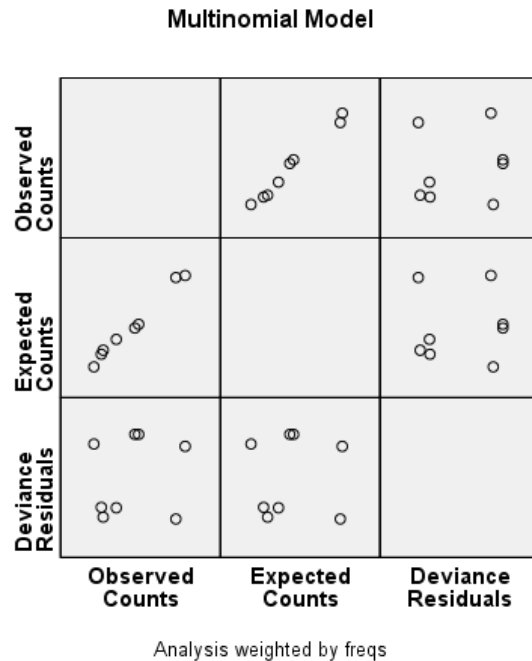
The Q-Q (quantile-quantile) plot, not shown here, will be approximately a straight line if the residuals are standard Normal random variables, which should be the case for a well fitting model.

Raw, adjusted and deviance residuals are displayed (and can be saved if required, using the **Save** button in SPSS Dialog Box 12.1). The deviance residuals are calculated from the contribution each cell makes to the Likelihood Ratio goodness of fit statistic. This makes them particularly useful if you are trying to find the cell(s) where the fit is bad. If you want the deviance residuals rather than the adjusted residuals plotted, click the box in the **Options** dialog box.

#### *Residuals and their plots for the conditional independence model*

The residuals can be seen in SPSS Output 12.4(a) (none had an absolute value above 2.55), and we show in SPSS Output 12.5 one of the plots provided to assist in identifying any lack of fit. Only three of the small graphs are needed as the other three are just mirror images. The observed and expected counts form a nearly straight line: the expected values are all close to the observed values. The plot of the residuals against expected counts shows no pattern or exceptional values. The plot of residuals

against observed is similar since observed and expected values are close for this model.

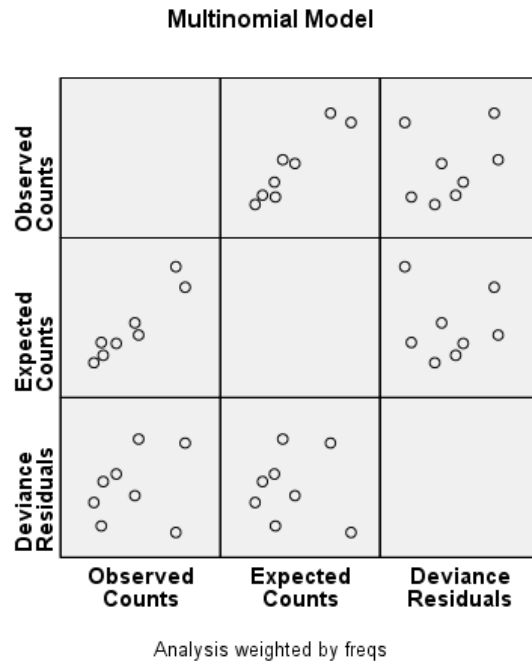


*SPSS Output 12.5. Graphical diagnostics for well-fitting model*

*Residuals and their plots for an alternative (badly fitting) model*

For comparison, we show the same graphs for the model with only one two-factor interaction, OBESITY\*DIABETES, in SPSS Output 12.6. This model did not fit well according to the goodness of fit statistics ( $\chi^2$  for the likelihood ratio was 9.906 on 3 degrees of freedom). Here we see that the observed and expected points form a less good approximation to a straight line, as we expect since this model was not a good fit. The residuals against expected counts do not, in this case, show a pattern that is helpful in identifying the problem, but we can at least see that we don't have just one or two large residuals which would indicate that the fit was much worse for one or two cells than for the rest. In fact, using the parameter estimates from the saturated

model (SPSS Output 12.1) we can easily see how to improve it: just try adding the term that has the largest Z value that is not already in the model (i.e., FTO\*OBESITY).



*SPSS Output 12.6. Graphical diagnostics for poorly-fitting model*

### *Collapsing a table*

If a variable in a three dimensional table is at least conditionally independent of one of the other two variables at each level of the third, then we can collapse the table over that variable and it will not affect the interaction between the other two variables.

In our example, FTO and DIABETES are conditionally independent at each level of OBESITY, so we could collapse over either FTO or DIABETES and see the interaction between the other two. Collapsing over FTO gives us Table 12.5(a), and collapsing over DIABETES gives Table 12.5(b).

Table 12.5

*(a) Collapsing Table 11.1 over FTO*

	diabetes 1	no diabetes2	totals
not obese 1	22	63	85

<b>obese</b>	<b>2</b>	78	37	115
<b>totals</b>		100	100	200

*(b) Collapsing Table 11.1 over DIABETES*

		<b>variant FTO 1</b>	<b>normal FTO 2</b>	<b>totals</b>
<b>not obese 1</b>		14	71	85
<b>obese 2</b>		39	76	115
<b>totals</b>		53	147	200

Both the tables in Table 12.5 show significant association between the row and column variables (the  $\chi^2$  values are 34.39 and 5.95 on one degree of freedom). This is why our attempt to collapse the table over OBESITY when we first described the experiment was misleading: the table can be collapsed over either of the other two variables (which are conditionally independent within levels of OBESITY) but not over OBESITY, which is associated with both the other variables.

There is another situation where a table can be collapsed. If a two-way table has independent row and column variables, then, if there are more than two categories for the row or the column, categories can be combined and the row and column variables will still be independent.

### *Measures of association and size of effects*

#### *The odds ratio for 2 x 2 tables*

In a 2\*2 table with frequencies  $a$  and  $b$  in the first row,  $c$  and  $d$  in the second, the odds of the levels of the row variable are  $a/c$  and  $b/d$  for the two levels of the column variable. The ratio of these odds,  $(a/c)/(b/d) = ad/bc$  will be close to 1 in the case of independence, and the stronger the association between row and column variables, the further from 1 this ratio will be. If we rearranged the category order in the rows or

columns, the odds ratio would be inverted (e.g., 0.06 would become  $1/0.06 = 1.67$ ). It is just as valid to look at either version. For Tables 12.5(a) and 12.5(b), both with significant association between row and column variables, the values of  $ad/bc$  are 0.17 and 0.38. The association between DIABETES and OBESITY (0.17) is stronger than between FTO and OBESITY (0.38). For the two tables in Table 12.1, neither of which show a significant association between row and column variables, the values are 1.76 and 1.60 (their reciprocals are 0.57 and 0.62, which are closer to 1 than 0.17 and 0.38). This odds ratio for a 2\*2 table is one of several possible measures of association. The odds ratio cannot be applied to tables with more than two categories in rows or columns. Other measures, which can be applied to a two-way table with more than two categories in one or both variables, are based on the  $\chi^2$  statistic for the table.

For tables in more than two dimensions, measures of association can only be calculated for pairs of variables within levels of the third variable, or combinations of levels of the third and any other variables (i.e., compound variables). Only if the table turns out to be collapsible down to two dimensions can a single measure of association tell us all we want to know.

#### *Using parameter estimates and odds ratios to indicate the size of an effect*

The odds ratio suggests how we can use the parameter estimates to indicate the size of an effect. We will look at the odds of variant FTO for each level of DIABETES within level 2 of OBESITY, which is the level with a zero parameter in SPSS Output 12.4(b). For a person who is DIABETES 1 (has diabetes) or DIABETES 2 (does not have diabetes) and OBESITY 2 (obese) the odds of being FTO 1 (variant) against FTO 2 (normal) can be obtained from the ratios of expected values ( $26.45/51.55 = 0.51$  for those with

diabetes or  $12.55/24.45 = 0.51$  for those without diabetes, see SPSS Output 12.4(a). This value can also be obtained from the main effect for FTO level 1, which is  $-0.667$  (see SPSS Output 12.4(b)). The odds ratio we want from this main effect is  $\exp(-0.667) = 0.51$ . The odds ratio of variant FTO is the same for the two levels of DIABETES within a level of OBESITY because of the conditional independence of FTO and DIABETES within levels of OBESITY. To get this odds ratio from the list of parameter estimates in SPSS Output 12.4(b) we only need the main effect of FTO because we used level 2 of OBESITY, which has a zero parameter estimate, and we are finding the odds ratio within each level of DIABETES.

At the other level of OBESITY, the odds of variant FTO will again be the same for both levels of DIABETES, the value being  $0.20$  this time. We can get this from the ratios of expected values in SPSS Output 12.4(a) (i.e.,  $3.62/18.38$  and  $10.38/52.62$ ), as we did for level 2 of OBESITY. Alternatively, we can use the main effect for FTO and the interaction for FTO and OBESITY from SPSS Output 12.4(b). We need the main effect for FTO since that is the factor for which we are calculating an odds ratio. We need the interaction term for FTO and OBESITY for level 1 of OBESITY (the non-zero level) because it is the strength of the interaction that the odds ratio measures. We are doing an odds ratio for each level of DIABETES so we need no term for DIABETES. So the odds ratio for variant FTO within level 2 of OBESITY is  $\exp(-0.667-0.956) = 0.20$  for either level of DIABETES, as we got from the ratios of expected values.

OBESITY interacts with both FTO and DIABETES, so the odds of being OBESITY 1 are not the same for both levels of DIABETES at a given the level of FTO. For FTO 2 (parameter zero) and DIABETES 2 (parameter zero) for instance, the odds of being OBESITY 1 are

$52.624/24.452 = 2.15$  from the ratio of expected values in SPSS Output 12.4(a).

Alternatively, we can get it from the main effect for OBESITY found in SPSS Output 12.4(b),  $\exp(0.766) = 2.15$ . For FTO 2 and DIABETES 1 we need the main effect of OBESITY (that's the one we want the odds for) and the interaction of OBESITY and DIABETES,  $\exp(0.766-1.798) = 0.36$ , which is the same as we get from the ratio of expected values,  $18.376/51.548 = 0.36$ . These ratios reflect the fact that many more without diabetes than with diabetes are not obese.

In the previous paragraph we considered normal FTO (FTO 2). Even if we look at those who are variant FTO (FTO 1) we still find many more without diabetes than with diabetes at OBESITY 1 (not obese), and this is reflected in the odds ratios of OBESITY 1 for the two levels of DIABETES. So we can now take FTO 1, DIABETES 2, and find the odds ratio of OBESITY 1 from the expected values,  $10.376/12.548 = 0.83$ . Using the parameter estimates from SPSS Output 12.4(b) we need the main effect for OBESITY since we are finding its odds ratio, and the interaction for OBESITY and FTO to find it at FTO 1,  $\exp(0.766-0.956) = 0.83$ , the same as we got from the expected values. For DIABETES 1 the odds are  $3.624/26.452 = 0.14$  or  $\exp(0.766-0.956-1.798) = 0.14$ . This time we needed the interactions for OBESITY and FTO and for OBESITY and DIABETES since we want the odds of OBESITY 1 at the non-zero parameter levels of both FTO and DIABETES. Once again the odds ratio for OBESITY 1 (not obese) is much higher for non-diabetics than diabetics even though we considered only those with variant FTO.

### *A general solution for using parameter estimates to calculate odds*

If you compare the results in the previous two paragraphs with Table 12.4 you can see that reference to this table enables any required odds to be calculated. Just identify the two relevant rows, and use the exp of the terms that are only in one of the rows. For

instance, to get the odds of being OBESITY 1 for FTO 1 and DIABETES 2, we would need rows 5 and 7 (where FTO is 1, DIABETES is 2, and OBESITY is 1 in row 5, and FTO is 1, DIABETES is 2, and OBESITY is 2 in row 7) to give us the odds for OBESITY 1. The terms 0.766 and -0.956 are in row 5 (OBESITY 1) but not row 7, so we need  $\exp(0.766 - 0.956) = 0.83$ , as in the previous paragraph.

We recognise that the computation of effect sizes that we have demonstrated in this section is far from straightforward. We have included more computational detail than we have in general throughout the book because we think that, in this case, the effort of trying to follow the detailed examples offers the best hope at arriving at an intuitive understanding of what is going on in broad terms. If it is too much for you, however, don't despair. Many researchers have conducted and published loglinear analyses without understanding the relationship between odds ratios, parameter estimates and effect size. If you do succeed in grasping the essentials of this relationship, you will probably be ahead of the pack. These considerations remind us that relationships among three variables are not simple unless all the three are independent.

As we noted at the beginning of this section, if the odds ratio  $ad/bc$  in the  $2 \times 2$  table is equal to 1, or equivalently if the  $\log(\text{odds})$  is zero, then the rows and columns are independent, so testing the hypothesis of independence is equivalent to testing that the odds ratio is 1 or the  $\log(\text{odds})$  is zero. However, looking at the odds for one value of a factor as we have done in this section is one way to gain more insight into the magnitude of the interactions, or the scale of the departures from independence.