

14. Analysing Categorical Data

Log-linear analysis

In clinical investigations we often have response and explanatory variables that are both categorical. For example ill / not ill as response variable and immunised / not immunised as explanatory variable. The categories here are nominal. There is no ordering between them. Sometimes the categories could be ordered, and we say that the variable is ordinal. For example survived; survived with deficits; died.

In the case of categorical data one is commonly looking for association between two variables. The χ^2 test is one example. Usually the χ^2 test is performed for a 2×2 contingency table. Even though the test is still valid for larger tables, one can run into difficulties with interpretation. All that a significant χ^2 test tells us is that the pattern of data as depicted in the table could not arise by chance. In a 2×2 contingency table the presence or absence of association between the two variables is often clear from inspection alone. The formal statistical test merely confirms (or refutes) it. In the case of complicated contingency tables involving several variables a more robust form of analysis is the log-linear analysis.

Recall that the χ^2 test involves entering the frequency counts for the two categorical variables in rows and columns together with the marginal totals (i.e. totals for each row and each column), as well as the full overall total. From these totals the expected frequency for each cell is calculated. Then $\chi^2 = \sum (\text{Observed frequency} - \text{Expected})^2 \div \text{Expected}$. (Recall also that the probability of the joint occurrence of two independent events is the product of their separate probabilities). A log-linear model is best thought of as a model for the expected frequencies in a contingency table. But it is more than just an alternative form of the χ^2 test. Its strength lies in that it can be extended to quite complicated contingency tables involving several variables.

In a 2×2 contingency table the probability of an individual occupying a given cell is the product of the marginal totals, since they represent the respective main effects probabilities. Log-linear analysis is based on the fact that the logarithm of a product is the sum of the individual logarithms of the individual terms in the product. In other words $\log(p \times q) = \log p + \log q$. To put it in the statistical jargon, the logarithm of the cell frequencies is a linear function of the logarithms of the components.

In log-linear analysis tables are formed that contain one-way, two-way, and higher order associations. The logarithm of the cell frequency is estimated by means of a linear equation (function in mathematical terminology). The log-linear model so developed starts with all the one-way, two-way, and higher order associations. The aim is to construct a model such that the cell frequencies in a contingency table are accounted for by the minimum number of terms. This is done by a process of backward elimination. What this means is that one begins with the maximum number of terms, and then drops a term in each round. Statisticians refer to it as the backward hierarchical method.

In practice, one commences the analysis by including all the variables. This is referred to as the saturated model. It can usually be expected to predict the cell frequencies perfectly. Then the highest order interaction is removed, and its effect on how closely the model can now predict the cell frequencies is noted. This process of progressive elimination is continued. Each time a variable is removed a statistical test is performed to determine whether the accuracy of prediction falls to an extent such that the component most recently eliminated should be one of the components of the final model. At each stage the assessment of

goodness-of-fit is made by means of a statistic known as the likelihood ratio. The final model includes only the associations necessary to reproduce the observed frequencies.

A comparison of the observed and expected frequencies for each cell using the likelihood ratio makes the evaluation of the final model. In the same way as in the case of χ^2 test, small expected frequencies can lead to loss of power. It is recommended that all expected frequencies should be greater than 1, and not more than 20% should be less than 5.

We take the following example to illustrate how log-linear analysis works:

In a hospital accident and emergency service 176 subjects who attended for acute chest pain were enrolled in a study. Of these 71 had abnormal electrocardiograms and in the case of 105 it was normal. Of those with abnormal electrocardiograms, 57 were overweight as judged by their body mass index, and 14 were normal. By comparison out of the 105 subjects with normal electrocardiograms 40 were overweight and 65 normal.

In the first group of 71 subjects with abnormal electrocardiograms, out of the 57 overweight subjects 47 were smokers and 10 non-smokers. Amongst the 14 with normal weights 8 were smokers and 6 non-smokers.

In the second group of 105 with normal electrocardiograms out of the 40 overweight subjects 25 were smokers and 15 non-smokers. Amongst the 65 with normal weights 35 were smokers and 30 non-smokers.

The investigators wish to assess the contribution that overweight and smoking make to coronary artery disease.

The data is presented below:

ECG	BMI	SMOKE	COUNT
1	1	1	47
1	1	2	10
1	2	1	8
1	2	2	6
2	1	1	25
2	1	2	15
2	2	1	35
2	2	2	30

ECG 1= Abnormal
 2= Normal
BMI 1= Overweight
 2= Normal weight
Smoke 1= Smoker
 2= Non-smoker

We first perform a simple cross-tabulation to check whether the frequencies per each cell are adequate to allow log-linear analysis.

Cross-tabulation

Control: SMOKING = 1

Rows: ECG	Columns: BMI		
	1	2	All
1	47 34.43	8 20.57	55 55.00
2	25 37.57	35 22.43	60 60.00
All	72 72.00	43 43.00	115 115.00

Chi-Square = 23.503, DF = 1, P-Value = 0.000

Control: SMOKING = 2

Rows: ECG	Columns: BMI		
	1	2	All
1	10 6.56	6 9.44	16 16.00
2	15 18.44	30 26.56	45 45.00
All	25 25.00	36 36.00	61 61.00

Chi-Square = 4.151, DF = 1, P-Value = 0.042

Cell Contents --
 Count
 Exp Freq

From the above results we infer that among both smokers and non-smokers there is an association between being overweight and an abnormal electrocardiogram. How much is the extent of the interaction between an abnormal electrocardiogram, smoking and being overweight?

This question is better answered by log-linear analysis as shown below:

[In SPSS Statistics → Loglinear. Then click on Model selection to open Model Selection Loglinear Analysis dialogue box.]

176 cases will be used in the analysis.

FACTOR Information

Factor	Level	Label
BMI	2	
ECG	2	
SMOKING	2	

***** H I E R A R C H I C A L L O G L I N E A R *****

DESIGN 1 has generating class

BMI*ECG*SMOKING

Note: For saturated models .500 has been added to all observed cells.

This value may be changed by using the CRITERIA = DELTA subcommand.

The Iterative Proportional Fit algorithm converged at iteration 1.
 The maximum difference between observed and fitted marginal totals is .000
 and the convergence criterion is .250

(1). Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
BMI	1				
ECG	1				
SMOKING	1	47.5	47.5	.00	.00
SMOKING	2	10.5	10.5	.00	.00
ECG	2				
SMOKING	1	25.5	25.5	.00	.00
SMOKING	2	15.5	15.5	.00	.00
BMI	2				
ECG	1				
SMOKING	1	8.5	8.5	.00	.00
SMOKING	2	6.5	6.5	.00	.00
ECG	2				
SMOKING	1	35.5	35.5	.00	.00
SMOKING	2	30.5	30.5	.00	.00

Goodness-of-fit test statistics

Likelihood ratio chi square = .00000 DF = 0 P = 1.000
 Pearson chi square = .00000 DF = 0 P = 1.000

(2)

***** H I E R A R C H I C A L L O G L I N E A R *****

Tests that K-way and higher order effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
3	1	1.389	.2386	1.420	.2334	4
2	4	44.530	.0000	46.724	.0000	2
1	7	69.822	.0000	68.727	.0000	0

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	3	25.292	.0000	22.004	.0001	0
2	3	43.142	.0000	45.304	.0000	0
3	1	1.389	.2386	1.420	.2334	0

***** H I E R A R C H I C A L L O G L I N E A R *****

(3)

Backward Elimination (p = .050) for DESIGN 1 with generating class

BMI*ECG*SMOKING

Likelihood ratio chi square = .00000 DF = 0 P = 1.000

If Deleted Simple Effect is	DF	L.R. Chisq	Change	Prob	Iter
BMI*ECG*SMOKING	1		1.389	.2386	4

Step 1

The best model has generating class

BMI*ECG
 BMI*SMOKING
 ECG*SMOKING

Likelihood ratio chi square = 1.38856 DF = 1 P = .239

If Deleted Simple Effect is	DF	L.R. Chisq	Change	Prob	Iter
BMI*ECG	1		27.631	.0000	2
BMI*SMOKING	1		3.080	.0792	2
ECG*SMOKING	1		3.505	.0612	2

Step 2

The best model has generating class

BMI*ECG
 ECG*SMOKING

Likelihood ratio chi square = 4.46886 DF = 2 P = .107

If Deleted Simple Effect is	DF	L.R. Chisq	Change	Prob	Iter
BMI*ECG	1		32.094	.0000	2
ECG*SMOKING	1		7.968	.0048	2

***** H I E R A R C H I C A L L O G L I N E A R *****

Step 3

The best model has generating class

BMI*ECG
 ECG*SMOKING

Likelihood ratio chi square = 4.46886 DF = 2 P = .107

***** H I E R A R C H I C A L L O G L I N E A R *****

The final model has generating class

BMI*ECG
 ECG*SMOKING

The Iterative Proportional Fit algorithm converged at iteration 0.
 The maximum difference between observed and fitted marginal totals is .000
 and the convergence criterion is .250

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
BMI	1				
ECG	1				

SMOKING	1	47.0	44.2	2.85	.43
SMOKING	2	10.0	12.8	-2.85	-.79
ECG	2				
SMOKING	1	25.0	22.9	2.14	.45
SMOKING	2	15.0	17.1	-2.14	-.52
BMI	2				
ECG	1				
SMOKING	1	8.0	10.8	-2.85	-.86
SMOKING	2	6.0	3.2	2.85	1.60
ECG	2				
SMOKING	1	35.0	37.1	-2.14	-.35
SMOKING	2	30.0	27.9	2.14	.41

Goodness-of-fit test statistics

Likelihood ratio chi square =	4.46886	DF = 2	P = .107
Pearson chi square =	4.88270	DF = 2	P = .087

Interpreting the Output

(1). The output commences with information about the number of cases, the factors and their levels.

A hierarchical model is being fitted. In a hierarchical model it is sufficient to list the highest order terms. This is called “generating class” of the model.

(2). Test for K-way and higher order effects

The likelihood ratio chi-square with no parameters and only the mean is 69.822. The value for the first order effect is 44.530. The difference $69.822 - 44.530 = 25.292$ is displayed on the first line of the next table. The difference is a measure of how much the model improves when first order effects are included. The significantly small *P* value (0.0000) means that the hypothesis of first order effect being zero is rejected. In other words there is a first order effect.

Similar reasoning is applied now to the question of second order effect. The addition of a second order effect improves the likelihood ratio chi-square by 43.142. This is also significant. But the addition of a third order term does not help. The *P* value is not significant.

In log-linear analysis the change in the value of the likelihood ratio chi-square statistic when terms are removed (or added) from the model is an indicator of their contribution. We saw this in multiple linear regression with regard to R^2 . The difference is that in linear regression large values of R^2 are associated with good models. Opposite is the case with log-linear analysis. Small values of likelihood ratio chi-square mean a good model.

(3). Backward elimination ($p = .050$)

The purpose here is to find the unsaturated model that would provide the best fit to the data. This is done by checking that the model currently being tested does not give a worse fit than its predecessor.

As a first step the procedure commences with the most complex model. In our case it is

BMI * ECG * SMOKING. Its elimination produces a chi-square change of 1.389, which has an associated significance level of 0.2386. Since it is greater than the criterion level of 0.05, it is removed.

The procedure moves on to the next hierarchical level described under step 1. All 2 – way interactions between the three variables are being tested. Removal of BMI * ECG will produce a large change of 27.631 in the likelihood ratio chi-square. The *P* value for that is highly significant (prob = 0.0000). The smallest change (of 3.080) is related to the BMI * SMOKING interaction. This is removed next. And the procedure continues until the final model which gives the second order interactions of BMI * ECG and ECG * SMOKING.

Each time an estimate is obtained it is called iteration. The largest difference between successive estimates is called convergence criterion.

At the end the programme provides a table of observed and expected cell count, the residuals (Observed – Expected), and the standardised residuals (Residual $\div \sqrt{\text{Expected cell count}}$). This table helps us to assess how well the model fits the data. If the model fits the data well, the residuals should be small and without any identifiable pattern. Also if the model fits well the standardised residuals should have a normal pattern. Values of standardised residuals >1.96 or <-1.96 suggest discrepancies.

Goodness-of-fit test statistic checks how well the model fits the data. It is based on Pearson's Chi-square statistic $\chi^2 = \sum (\text{Observed} - \text{Expected})^2 \div \text{Expected}$.

An alternative statistic is the likelihood ratio chi-square

$$G^2 = 2 \sum \text{Observed} \ln (\text{Observed} \div \text{Expected}).$$

Small values of chi-square statistics indicate a good model. For large sample sizes both types of chi-square statistics are equivalent.

We conclude that being overweight and smoking have each a significant association with an abnormal cardiogram. However, in this particular group of subjects being overweight is more harmful.

We could have inferred this by calculating the odds ratio when we performed the cross tabulation. The odds ratio calculation is shown below:

	Cardiogram abnormal (ECG 1)	Cardiogram normal (ECG 2)
Overweight (BMI 1)	47	25
Normal weight (BMI2)	8	35
	Odds Ratio = 8.225	
	Cardiogram abnormal (ECG 1)	Cardiogram Normal (ECG 2)
Smoker (Smoking 1)	10	15
Non-Smoker (Smoking 2)	6	30
	Odds ratio = 2	

Comment

To perform a multi-way frequency analysis tables are formed that contain the one-way, two-way, three-way, and higher order associations. The log-linear model starts with all of the one-, two-, three-, and higher-way associations, and then eliminates as many of them as possible while still maintaining an adequate fit between expected and observed cell frequencies. In log-linear modelling the full model that includes all possible main effects and interactions fits

the data exactly, with zero residual deviance. One then assesses whether a less full model fits the data adequately by comparing its residual deviance with the full model.

In our example, the three-way association tested was between category of electrocardiogram, body mass index, and smoking. It got eliminated because it was found not significant. After that a two-way association (type of electrocardiogram and body mass index; type of electrocardiogram and smoking) was tested. The two-way association was found significant.

As we have seen the purpose of multi-way frequency analysis is to test for association among discrete variables. Once a preliminary search for association is completed by simple 2×2 contingency tables a model is fitted that includes only the associations necessary to reproduce the observed frequencies.

In the above example, we have a data set with a binary response variable (Electrocardiogram abnormal/normal) and explanatory variables that are all categorical. In such a situation one has a choice between using logistic regression and log-linear modelling. For performing logistic regression rearrangement of the data is needed so that for each variable we have a column of 1's and 0's.

Other differences from logistic regression are:

1. There is no clear demarcation between outcome and explanatory variables in log-linear models.
2. Logistic regression allows continuous as well as categorical explanatory variables to be included in the regression analysis.

We now look at the result of logistic regression for the same data:

Binary Logistic Regression

Link Function: Logit

Response Information

Variable	Value	Count
ECG	1	71 (Event)
	0	105
Total		176

Logistic Regression Table

Predictor	Coef	StDev	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-1.9523	0.3854	-5.07	0.000			
BMI	1.7960	0.3641	4.93	0.000	6.03	2.95	12.30
SMOKING	0.6954	0.3754	1.85	0.064	2.00	0.96	4.18

Log-Likelihood = -100.890

Test that all slopes are zero: G = 35.599, DF = 2, P-Value = 0.000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	1.420	1	0.233
Deviance	1.389	1	0.239
Hosmer-Lemeshow	1.420	2	0.492

Table of Observed and Expected Frequencies:

(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

Value	Group				Total
	1	2	3	4	
1					
Obs	6	8	10	47	71
Exp	4.5	9.5	11.5	45.5	
0					
Obs	30	35	15	25	105
Exp	31.5	33.5	13.5	26.5	
Total	36	43	25	72	176

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	4650	62.4%	Somers' D	0.49
Discordant	1020	13.7%	Goodman-Kruskal Gamma	0.64
Ties	1785	23.9%	Kendall's Tau-a	0.24
Total	7455	100.0%		

Recall that in logistic regression we are interested in the odds of events. The logistic model is about logarithms of odds, so that $\log_e \{P/1-P\} = \alpha + \beta_1 X_1 + \beta_2 X_2 \dots$ and so on. Hence the coefficients are measures of change in log odds associated with a unit change in the explanatory variable.

The programme calculates the odds ratios (actually these are the adjusted odds ratios) and lists them under the heading of the same name. For BMI the adjusted odds ratio is 6.03 and for smoking it is 2.00. These values are close to the crude odds ratio calculated from cross tabulation.

The details of the rest of the output are not explained here. Suffice it to say that it is a well fitting model. We can conclude that in this group of subjects those with an abnormal cardiogram were 6 times more likely to be overweight and twice more likely to be smokers compared to those with normal cardiograms.

Is there any interaction between being overweight and smoking? We can check for this by introducing an interaction term in the regression.

The results of the analysis including an interaction term are shown below:

Logistic Regression Table

Predictor	Coef	StDev	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-1.6094	0.4472	-3.60	0.000			
BMI	1.2040	0.6055	1.99	0.047	3.33	1.02	10.92
SMOKING	0.1335	0.5946	0.22	0.822	1.14	0.36	3.67
BMI*Smok	0.9032	0.7626	1.18	0.236	2.47	0.55	11.00

Log-Likelihood = -100.196

Test that all slopes are zero: G = 36.987, DF = 3, P-Value = 0.000

The result of the second regression does not require further explanation and are not discussed.

Comment

The advantage of logistic regression is that it makes less demand on computer resources in terms of memory and time. A second advantage is that logistic regression can deal with continuous as well as categorical explanatory variables. Log-linear analysis involves modelling the relationships between the explanatory variables, and this may be of interest in some instances. Secondly, the log-linear approach works well for categorical response variables with more than two categories. The classical form of logistic regression does not work well in such situations, though more recent versions of some packages (e.g. Minitab) can address this issue.