Loglinear models. Online Supplement 5 to Serious stats: A guide to advanced statistics for the behavioral sciences. Basingstoke...

Chapter · January 2012

DOI: 10.13140/2.1.5107.4884

Thom S Baguley

Nottingham Trent University

Online Supplement 5

Loglinear models

This supplement draws primarily on Chapters 4, 7 and 17.

OS5.1 Loglinear models

A loglinear model is a generalized linear model that is closely related to both logistic and Poisson regression. Used in its broadest sense, all models in which Y is an additive function of a set of predictors and in which a logarithmic link function or transformation is employed are loglinear models (Agresti, 1996). This includes both logistic and Poisson regression, but the term loglinear model is most strongly identified with applications in which contingency tables are modeled in terms of an additive function of the logarithm of the counts in each cell. As contingency tables are widely used in medicine, psychology and other disciplines, this kind of loglinear model (like logistic regression) is often learned in isolation, rather than as part of the generalized linear model. This is a shame, because loglinear models are in some ways more restrictive than Poisson or logistic regression (e.g., being unable to incorporate continuous predictors).

The link between a loglinear model and analysis of contingency tables can be demonstrated by considering a two independent group experiment with a dichotomous outcome. The usual analysis in this situation is a χ^2 test of independence, and the data are usually set out as a 2×2 contingency table:

$$\begin{array}{c|c|c|c}
 & A_1 & A_2 & \\
\hline
B_1 & O_{1,1} & O_{1,2} & R_1 \\
\hline
B_2 & O_{2,1} & O_{2,2} & R_2 \\
\hline
 & C_1 & C_2 & n
\end{array}$$

If you assume independence of observations, the probability of any observation falling into a given cell is its row probability multiplied by its column probability:

$$\hat{P}_{ij} = \frac{R_i}{N} \times \frac{C_j}{N} = \hat{P}_i \hat{P}_j \qquad \text{Equation OS5.1}$$

Multiplying this by N produces the expected value of the cell counts under independence (and gives the formula for the expected value of the cells in the χ^2 test of independence).[1] For present purposes, what matters is that there is a multiplicative relationship between expected probabilities of the marginals (row and column totals) and the expected probabilities of the cells. Taking logarithms of both sides would give us an additive formula:

$$\ln\left(\hat{P}_{ij}\right) = \ln\left(\hat{P}_i\hat{P}_j\right) = \ln\left(\hat{P}_i\right) + \ln\left(\hat{P}_j\right) \qquad \text{Equation OS5.2}$$

This in turn implies that a loglinear model can be used to predict cell probabilities from marginal probabilities. This is the core of how a loglinear model of contingency table data works, except that the modeling is done as a form of Poisson regression of the counts. This works because of the links between the Poisson and multinomial distributions; k independent Poisson variables will have a joint distribution that is multinomial. Thinking of it in terms of a Poisson model, there are four counts to be modeled (one for each cell of the 2×2 table). The model can therefore have, at most, four parameters when expressed in count form.

The independence model for a two-way contingency table between categorical variables A and B, with I rows and J columns is:

$$\ln\left(y_{ij}\right) = \lambda + \lambda_i^A + \lambda_j^B \qquad \text{Equation OS5.3}$$

The notation here is that commonly used for loglinear models (e.g., see Agresti, 1996). The superscripts 'A' and 'B' in λ_i^A and λ_j^B are not exponents, they are just used as labels for the row and column effects.[2] Thus A might be the presence or absence of a drug and B might be the subsequent health status of a patient (e.g., healthy versus unhealthy). The number of rows I is the number of categories for the variable A, and the number of columns J is the number of categories for the variable B. This model has three parameters: one for the intercept, one for the row effect and one for column effect. This leaves $4 - 3 = 1$ residual *df*.

The independence model might not be a good fit, and this can be assessed in relation to a model with additional terms. The logical comparator is a model that also includes the interaction between row and column effects (i.e., the $A \times B$ interaction):

$$\ln\left(y_{ij}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{A\times B} \qquad \text{Equation OS5.4}$$

This model will predict the observed counts perfectly and is therefore a saturated model. It has four parameters and therefore $4 - 4 = 0$ residual *df*. Any loglinear model that includes the intercept, all marginal effects (i.e., row and column effects) and all interactions between marginal effects will be a saturated model. Considered as a Poisson regression, a saturated loglinear model has one parameter for each of the counts being modeled; in Equation OS5.4 there are four parameters and four cells to be modeled. The perfect fit is therefore spurious, in that being able to predict four observations from four parameters is trivial. If you are unsure why such perfect fits are trivial, think about a simpler model such as that for a Pearson correlation. With n data points it is always possible to obtain a perfect correlation by correlating a variable with itself (or a linear function of it). The interaction term in a two-way contingency table represents the residuals of the counts after the main effects have been stripped out, so the model being fitted is equivalent to predicting the observed counts from themselves.

The saturated model, although of little interest in its own right, is a natural standard of comparison for the independence model. Deviance for a loglinear model can be calculated using

Equation 17.15. The likelihood ratio test of the independence model is therefore the difference in deviance between the saturated model (D_S) and the independence model (D_M). This is another way of arriving at the likelihood chi-square statistic as $G^2 = D_M - D_S$. As the deviance of the saturated model is always zero, the comparison of the independence model with the saturated model is in this case also the goodness-of-fit for the model (i.e., $G^2 = D_M$). The connection between loglinear models for contingency tables and the χ^2 test of independence extends to the problems of approximating discrete outcomes with a continuous distribution. When data are sparse, inferences from loglinear models can be problematic (see Chapter 17). Agresti (1996) considers some alternative (e.g., exact approaches) for sparse tables, though the ideal solution is to obtain a larger sample. An even more serious problem is if independence of counts is violated. Lack of independence is a major cause for concern, but can sometimes be dealt with using the solutions suggested for Poisson and logistic regression (e.g., for dealing with overdispersion or repeated measures).

Example OS5.1 The table below sets out a two-way contingency table for the dream data first introduced in Example 17.1.

	Ordinary	*Scary*	
No magical suggestion	25	1	*26*
Magical suggestion	14	7	*21*
	39	*8*	*47*

Rather than calculate a χ^2 by hand, it will be useful to work with the parameters estimated from a fitted model. The parameter estimates will come out slightly differently depending on how the categories represented by the rows (no magical suggestion versus magical suggestion) and columns (ordinary versus scary dream) are coded. The two obvious choices are dummy coding and effect coding. For effect coding[3] the coefficients of the independence model $\ln(y_{ij}) = \hat{\lambda} + \hat{\lambda}_i^{Group} + \hat{\lambda}_j^{Dream}$ fitted in R are:

Intercept ($\hat{\lambda}$)	2.1727
No magical suggestion ($\hat{\lambda}^{Group = -1}$)	0.1068
Magical suggestion ($\hat{\lambda}^{Group = 1}$)	−0.1068
Ordinary dream ($\hat{\lambda}^{Dream = -1}$)	0.7921
Scary dream ($\hat{\lambda}^{Dream = 1}$)	−0.7921

As this is a form of loglinear model, the expected or predicted counts under independence can be obtained by exponentiation. The expected count under independence of ordinary dreams in the no suggestion group is:

$$e^{2.1727+0.1068+0.7921} = e^{3.0716} \approx 21.58$$

This will be identical to the expected values of the chi-square test of independence:

$$E_{1,1} = \frac{R_1 \times C_1}{N} = \frac{39 \times 26}{47} \approx 21.58$$

We could get different, but equivalent, parameter estimates using dummy coding. This model, fitted in SPSS, gives:

Intercept ($\hat{\lambda}$)	1.274
No magical suggestion ($\hat{\lambda}^{Group\,=\,1}$)	0.214
Ordinary dream ($\hat{\lambda}^{Dream\,=\,1}$)	1.584

Here the intercept represents the estimate for the group coded zero on both dummy variables. The prediction for ordinary dreams in the no magical suggestion group is unchanged at $e^{1.274+0.214+1.584} = e^{3.072} \approx 21.58$. The parameter estimate for the group has been shifted by a constant to reflect the difference in reference category. The effects of the row and column marginals (main effects by analogy to ANOVA) are double the size of the effect estimates under effect coding (because they are the estimated effect of a one-unit increase in the predictor on the log odds).

The deviance for the independence model and therefore the likelihood ratio chi-square test statistic is: $G^2(1, N = 47) = 7.67$, $p = .006$, identical to the value in Example 17.1. The independence model can be rejected in favor of a model that includes the group × dream interaction term. The saturated model is therefore a better fit to the observed counts than the independence model: the rate of scary dreams seems to differ between the two groups. Of course the interaction term requires an extra parameter with $(r-1)(c-1) = 1$ *df*. Even taking this into account (e.g., using AIC or BIC), the saturated model including the two-way interaction term is preferred. Note that the *df* of a χ^2 test of independence is therefore the 1 *df* difference between the residual *df* of the independence model (residual $df = 1$) and the saturated model (residual $df = 0$).

Predicted counts from the saturated model can also be obtained. The outcome may be slightly different – depending on your software (because of how it deals with residuals equal to zero in a perfectly fitted model). SPSS fits parameter estimates for the saturated model by adding 0.5 to each count. The deviance is exactly zero and need not be computed. For the dummy coded model (adding 0.5 to each cell) the parameter estimates are:

Intercept ($\hat{\lambda}$)	2.015
No magical suggestion ($\hat{\lambda}^{Group\,=\,1}$)	−1.609
Ordinary dream ($\hat{\lambda}^{Dream\,=\,1}$)	0.659
Interaction ($\hat{\lambda}^{Group\times Dream\,=\,1}$)	2.174

The predicted value of ordinary dreams in the no magical suggestion condition is now $e^{2.105+(-1.609)+0.659+2.174} = e^{3.239} \approx 25.51$. This (allowing for rounding error) is equal to the observed count of 25 plus the 0.5 SPSS adds to deal with saturation. The interaction term $e^{2.174} =$ gives the estimated *OR* for the saturated model. Because SPSS adds 0.5 to each cell, this is equal to the Haldane estimator for the *OR* (see Section 7.4.6). The *OR* can be estimated instead directly from the observed values as $(7/14)/(1/25) = 12.5$. Thus the odds of a scary dream are 12.5 times higher in the magical suggestion than the no suggestion condition. The Haldane estimator gives a more conservative estimate (probably too conservative) when some cells have small values. Other software, such as R, fits the model by another route and will predict the exact count of 25 for this cell. A final option would be to use logistic regression. This gives the interaction effect as 2.526 and $e^{2.526} = 12.5$.

OS5.1.1 Interpreting *k*-way contingency tables

If a loglinear model is restricted to two-way contingency tables it makes sense to talk about the marginals as being row and column effects. But loglinear models for two-way tables are not very useful (as most of what they deliver can be achieved using the χ^2 test of independence). Loglinear models are most useful for exploring k-way tables where $k > 2$. For such data it is a good idea to ditch talk of rows and columns. ANOVA terminology is often used in its place. The marginal effects are termed main effects of predictors. Interaction effects can be added to the model until a saturated model is achieved: this is a model in which all k-way and lower-order interactions are present (in addition to the main effects and the intercept). Each interaction effect incorporates non-independent effects of two or more main effects. Thus a four-way contingency table that adds a $B \times C \times D$ interaction estimates the expected counts for a model in which the effects of B and C depend on categories defined by the variable D. Although the terminology is borrowed from ANOVA there are some important differences.

Loglinear models are slightly unusual in two ways. First, it is trivial to fit the observed data perfectly by including all the terms for a saturated model. This has led to the convention of fitting models hierarchically (see Chapter 14), either by automatic selection methods such as backwards elimination or by fitting the saturated models and removing interaction terms.[4] The objective is to find a plausible model for the data that is simpler than the saturated model. The saturated model is, in effect, a prediction of the cell counts using the observed cell counts. A good model should explain the data as well, or almost as well, as the saturated model. Null hypothesis significance tests (NHSTs) for the models are therefore goodness-of-fit tests for the simpler models relative to either the saturated model or some other competitor. Nested models can be tested by comparing differences in deviance. For non-nested models, separate comparisons can be made with the saturated model. As it is not sensible to fit interaction effects without also including lower-order effects (interactions, main effects and the intercept), non-nested models only arise infrequently. Specifically, they arise when comparing models with only some of the possible terms at a particular level (e.g., with only one two-way interaction for a three-way table). Information criteria such as AIC_C, AIC or BIC provide alternatives for non-nested models, and are increasingly preferred over NHSTs because they favor models with fewer parameters.

The second way in which loglinear models are unusual is that there is no formal distinction between predictors and outcomes among the categorical variables in the model. This characteristic is also true of the χ^2 test of independence. The cell counts are the response and, in a technical sense, all categorical variables act as predictors. This can confuse people, because it will often appear as though one of the categorical predictors is acting as an outcome variable. For instance, a researcher might collect data on preferences for different types of drink as a function of gender and several different health promotion interventions (e.g., children might be offered a choice of a low-sugar or high-sugar carbonated drink). It is natural to think of the type of drink as the outcome and the interventions and gender as predictors. This is not technically a property of the model, but can be a useful way to interpret it (though from time to time it might mislead).

OS5.1.2 Interpreting a three-way contingency table

The loglinear regression equation for k categorical predictors depends on the complexity of the model that is fitted. Assuming a saturated model – with parameters for all effects – the equation becomes cumbersome when k is large because of the number of possible terms. The key points can be illustrated clearly in terms of a three-way contingency table: a model with three categorical variables A, B and C. This model is sufficiently complex to demonstrate how a loglinear model would work for a realistic application. In fitting the model, it is common to work backward from the saturated model. However, in explaining what's going on it is easier to work forward from the intercept-only model.

The intercept-only model estimates one parameter, the constant λ:

$$\ln\left(y_{ijk}\right) = \hat{\lambda} \qquad \text{Equation OS5.5}$$

The subscripts i, j and k refer to the I, J and K categories within each categorical variable. If $I = J = K = 2$ this is a $2 \times 2 \times 2$ contingency table. The intercept-only model is equivalent to fitting the grand mean of the cell count to all cells. The grand mean is the total N divided by the number of cells ($I \times J \times K$). For a $2 \times 2 \times 2$ table $\lambda = N/8$. The intercept-only model is sometimes called an *equiprobability model* because it assumes an observation has an equal probability of falling into any of the $I \times J \times K$ cells. Equiprobability models are only rarely of theoretical interest. The marginals for one or more categorical variables are, in most studies, either fixed by design or unrepresentative of the population of interest. This would make an equiprobability model of little practical value; if you deliberately sampled equal numbers of males and females it would be uninteresting that observations fell evenly between the two categories.

The next model to consider is one that includes all main effects. This is the independence model. For a three-way contingency table it is:

$$\ln\left(y_{ijk}\right) = \hat{\lambda} + \hat{\lambda}_i^A + \hat{\lambda}_j^B + \hat{\lambda}_k^C \qquad \text{Equation OS5.6}$$

Each of the main effects represents the differences in marginal counts for the variables A, B and C. If the marginal counts within each of the categorical variables A, B or C are identical, these parameters take the value zero. If any of the categories within A, B or C are more numerous than the others, this will be reflected in non-zero main effects. Main effects in loglinear models therefore differ from ANOVA because they often reflect structural characteristics of the sample (e.g., the fact that the size of the experimental and control groups differ). However, if one of the categorical variables lends itself to being interpreted as an outcome variable (e.g., C is passing or failing a test) the main effect for this category may have a meaningful interpretation. If so, it may be of theoretical interest to compare the independence model to a model without that term (e.g., without $\hat{\lambda}_k^C$).

If the independence model is a poor fit, then one or more interaction terms are required. The model with all possible two-way interactions is:

$$\ln\left(y_{ijk}\right) = \hat{\lambda} + \hat{\lambda}_i^A + \hat{\lambda}_j^B + \hat{\lambda}_k^C + \hat{\lambda}_{ij}^{A\times B} + \hat{\lambda}_{ik}^{A\times C} + \hat{\lambda}_{jk}^{B\times C} \qquad \text{Equation OS5.7}$$

Each interaction term incorporates a specific departure from independence. A two-way interaction in a loglinear model implies that the probability of an observation falling into a cell is different from that obtained by multiplying the marginal probabilities of the two categorical variables (collapsing over any other categories). $\hat{\lambda}_{ij}^{A\times B}$ thus represents a difference in the observed counts from the independence model; the probability of observations falling into categories defined by B differs across categories defined by A. If A were gender and B were treatment condition (e.g., placebo or drug), this would suggest that the number of males and females in the different treatment conditions differed. If C defined the outcome of treatment (e.g., success or failure) a $B \times C$ interaction would indicate that the proportion of successes differed between drug and placebo conditions. For this type of situation (where C can readily be interpreted as an outcome variable) two-way interactions in a loglinear model have an interpretation similar to that of main effect effects in ANOVA. They indicate a difference in outcomes between categories or groups. Models with only some of the two-way effects can be particularly useful in testing theoretical predictions. Agresti (1996) shows how a two-way model dropping a term such as $\hat{\lambda}_{ij}^{A\times B}$ assumes independence of A and C, controlling for the potential A-B and B-C associations. It follows that in Equation OS5.7, where all two-way terms are present, the interaction between any two variables is independent of the other (e.g., $A \times C$ is independent of B). This means that the odds ratio for any two variables is the same for all values of the third. Agresti (*ibid.*) calls this a *homogeneous association model*. In practical terms it implies that you can split the three-way table into separate $A \times B$, $A \times C$ and $B \times C$ tables to be interpreted independently.

The saturated model for a three-way table takes the form:

$$\ln\left(y_{ijk}\right) = \hat{\lambda} + \hat{\lambda}_i^A + \hat{\lambda}_j^B + \hat{\lambda}_k^C + \hat{\lambda}_{ij}^{A\times B} + \hat{\lambda}_{ik}^{A\times C} + \hat{\lambda}_{jk}^{B\times C} + \hat{\lambda}_{ijk}^{A\times B\times C} \qquad \text{Equation OS5.8}$$

This equation differs only in the addition of the three-way interaction term. A comparison of the saturated model with all two-way effects indicates whether the three-way interaction should be included in the model. Needing to include a $\hat{\lambda}_{ijk}^{A\times B\times C}$ term implies that the homogeneous association implied by the model with all two-way effects is rejected; the odds ratios between two variables (e.g., A and B) differ between categories defined by the third variable. In the example where A is gender, B is treatment group and C is success of treatment, this would imply that the difference in outcomes for the two treatment groups differs for males and females (e.g., the drug is more effective for males than females). This 'feels' like the interpretation of a two-way interaction in ANOVA (because C acts like an outcome measure).

In a $2 \times 2 \times 2$ (and in general for 2^k) contingency table each parameter has exactly 1 *df* (being coded by one indicator variable). What happens if a categorical variable has more than two categories? For each main effect (e.g., $\hat{\lambda}_i^A$) this simply requires an additional parameter for each extra category. As you might expect, it takes $J-1$ *df* (for $J-1$ indicator variables) to represent J categories. Interaction effects require additional indicator variables and use up addition *df*. A two-way interaction for variables with I and J categories requires $(I-1)(J-1)$ *df*. A three-way interaction requires $(I-1)(J-1)(K-1)$ *df*. As the number of categories increases, excluding interaction terms creates increasingly parsimonious models. The number of counts for a three-way model is $I \times J \times K$. This is maximum number of parameters that can be fitted. Fitting all these parameters exhausts the *df* and defines the saturated model. The more categories there are per variable, the greater the parsimony offered by dropping the highest-order interaction terms (e.g., in a $3 \times 3 \times 4$ table the homogeneous association model releases $2 \times 2 \times 3 = 12$ residual *df* relative to the saturated model).

Example OS5.2 Appleton *et al.* (1996) describe data taken from a 20-year follow-up study of 1314 women from the North-East of England. Among the variables were smoking status at the start of the study (smoker or non-smoker), age and survival at 20 years. Subsequent examples will refer to this data set as the smoking data. An obvious, but ultimately unwise, analysis is to use a χ^2 test of independence to compare the proportions of smokers and non-smokers who survived for 20 years. The raw data set out in contingency table form are:

	Non-smoker	*Smoker*	
Dead	230	139	*369*
Alive	502	443	*945*
	732	*582*	*1314*

The test of independence for this contingency table is statistically significant, $\chi^2(1, N=1314) = 9.12$, $p = .0025$. A similar result is obtained from the likelihood ratio test, G^2 $(1, N=1314) = 9.2$, $p = .0024$, AIC = 1048.5. Given the sample size, statistical significance is not surprising, but the direction of the effect is. Of the smokers 443/582 = .761 (76.1%) are alive after 20 years, but only 502/732 = .686 (68.6%) of the non-smokers survived. Although the difference is small the null hypothesis of independence can be rejected by this NHST or by other methods (e.g., ΔAIC relative to the independence model is 1055.7 − 1048.5 = 7.2).

This analysis ignores some of the available information (the age of the women in the sample at the start of the study). Exact ages are not available, but Appleton *et al*. report separate contingency tables for each of seven age groups. The percentage survival for smokers and non-smokers at each age group are summarized in Table OS5.1. Looking carefully at Table OS5.2 it should be possible to see that the smokers have lower percentage survival in every age group except 75+ and 25–34. This suggests the opposite pattern to that observed for the 2 × 2 table.

A loglinear analysis can help make sense of this discrepancy. The data set out by age group can be described in terms of a three-way contingency table defined by *smoking* (smoker versus non-smoker), *age* (from 18–24 through to 75+) and *survival* (alive or dead).

Table OS5.1 Percentage survival and sample size by age group for the smoking data

	Smokers		**Non-smokers**	
Age group	**Survival (%)**	***n***	**Survival (%)**	***n***
18–24	96.4	55	98.4	62
25–34	97.6	124	96.8	157
35–44	87.2	109	94.2	121
45–54	79.2	130	84.6	78
55–64	55.7	115	66.9	121
65–74	19.4	36	21.7	129
75+	0.0	13	0.0	64

With only three predictors, it makes sense to take a hierarchical approach. We'll start by fitting a model with three-way and lower effects: smoking * age * survival. The model of the cell counts is:

$$\ln\left(y_{ijk}\right) = \hat{\lambda} + \hat{\lambda}_i^{smoking} + \hat{\lambda}_j^{age} + \hat{\lambda}_k^{survival} + \hat{\lambda}_{ij}^{smoking \times age} + \hat{\lambda}_{ik}^{smoking \times survival} + \hat{\lambda}_{jk}^{age \times survival} + \hat{\lambda}_{ijk}^{age \times smoking \times survival}$$

In practice, there is no real need to fit this model, as it is a saturated model of all the predictors and has 2 × 7 × 2 = 28 parameters (one for each count being modeled). Residual deviance and *df* will both be zero. R reports AIC as 190.2.

The next step is to determine whether the model with all two-way interactions is a reasonable fit relative to the saturated model. Dropping the three-way smoking × age × survival interaction releases $(2-1)(7-1)(2-1)=6$ *df* with residual deviance of 2.38. The two-way model is therefore not a poorer fit according to a deviance test, G^2 $(6, N=1314)=2.38$, $p=.88$. The AIC is 180.6 and the AIC reduction of 9.6 suggests the two-way model is to be preferred ($LR_{AIC}=121.5$ in favor of the simpler model).

Only one further term is worth dropping from the model to test its effect. There is no need for a statistical test to decide to keep the age × survival term in the model; it can be assumed *a priori* that the older you are at the start of the study the less likely you are to survive 20 years. The age × smoking term also needs to stay in the model. If the smokers are older (or younger) than non-smokers in the sample at the start of the study, this is a confounding variable that needs to be controlled for. The term that needs testing is survival × smoking. Dropping this term provides a test of the hypothesis that survival differs between smokers and non-smokers. Keeping age × smoking and smoking × survival in the model ensures that the confounding influence of age is accounted for. Dropping survival × smoking frees up 1 *df* and increases deviance by 5.95 and AIC by 3.94. This indicates a worse fit with the likelihood ratio test, $G^2=5.95$, $p<.05$, or with AIC ($LR_{\mathrm{AIC}}=7.2$ in favor of the model including the interaction).

Because age has seven categories, the model with all two-way interactions has a large number of parameters (20). Rather than interpret all 20, we'll focus on a few key parameters:

Term	λ	*SE*
Intercept	0.246	0.595
Smoker	0.297	0.253
Age (45–54)	2.203	0.617
Survival	3.860	0.594
Smoker × age (45–54)	0.565	0.236
Survival × age (45–54)	−2.113	0.612
Survival × smoker	−0.427	0.177

Positive parameters indicate higher cell counts. As being alive and being a smoker are coded one and because the reference category for age is 18–24, the intercept is the predicted value of λ for non-smokers in the youngest age group who are deceased. This is $e^{0.246}=1.28$ and is close to the observed count of one. The smoker main effect merely indicates the relative counts of smokers and non-smokers (there are more non-smokers in the sample and so it is slightly positive). In a balanced design this parameter would be zero. The various age coefficients indicate the numbers at each age group in the sample. Here the 45–54 age group is fairly numerous and has a large coefficient. A high coefficient for the survival main effect just indicates that many more of the sample survived for 20 years than did not. The smoking × age coefficients indicate greater or fewer numbers smoking at each age group, and so the positive coefficient for the 45–54 group reflects the disparity between numbers of smokers (130) and non-smokers (78) in this subgroup. The survival × age coefficient captures the different survival prospects of older and younger people. As the 45–54 age group is at the older end of the sample, its members' survival prospects are worse than the reference category (the 18–24-year-olds). Putting this together, the predicted count for a 45–54-year-old smoker alive after 20 years is:

$$\hat{Y}=e^{0.246+0.297+2.203+3.860+0.565-2.113-0.427}=e^{4.631}=102.6$$

The observed value is 103.

Returning to the initial question, is it the case the smokers live longer? This simple answer is that the smokers in this sample do live longer; however, the loglinear model indicates that this not because they are smokers. Smoking is associated with worse life expectancy: the smoking × survival coefficient is negative. At -0.427 this is equivalent to a decrease of $100 \times (1 - e^{-0.427})$% or a 35% decrease in survival. The 95% CI for the counts is [0.46, 0.92], equivalent to a decrease in survival of between 8% and 64%. The original χ^2 analysis not only failed to detect this effect, but it suggested the reverse. This phenomenon is known as *Simpson's paradox*.[5] The apparent paradox occurs when collapsing data over one category (in this case age group) obscures or distorts the relationship between other variables (in this case smoking and survival). The smokers live longer because the original sample contained a higher proportion of older non-smokers than smokers (e.g., 314 versus 164 in the three oldest categories). Simpson's paradox is surprisingly common. If it seems familiar, this may be because it is related both to Lord's paradox and to suppression in multiple regression (Lord, 1967; Darlington, 1968). Tu *et al*. (2008) argue that all three are examples of a broader phenomenon (a *reversal paradox*) arising for continuous (suppression), categorical (Simpson's paradox) or a mixture of continuous and categorical variables (Lord's paradox). Ignoring a theoretically important variable can distort or reverse the direction of an effect.

Appleton *et al*. (1996) note that this analysis probably underestimates the impact of smoking on mortality. They point out that the apparent bias in sampling younger non-smokers probably arose because of the relative scarcity of older smokers in the population. Because they smoked, a greater proportion would have died young or have been too ill to enroll in the study. A final statistical concern is that the survival data appear to be underdispersed ($\hat{\varphi} = 0.40$). In an underdispersed model the *SE*s will be too large and inferences too conservative. This could be corrected (e.g., using a quasipoisson model) and would lead to increased confidence in the conclusions regarding smoking and survival.

OS5.1.3 When should you choose a loglinear model?

For two-way tables the χ^2 test of independence or a CI for the *OR* will often be sufficient. For three-way, four-way or higher-order tables a loglinear model is preferable to traditional χ^2 analyses and provides a richer set of hypotheses to test or models to compare. Nevertheless, it is probably also sensible to consider two alternative approaches: logistic regression and Poisson regression. All three approaches will be identical under certain circumstances (e.g., for a 2 × 2 table). It is worth discussing these connections in more detail.

Logistic regression and loglinear models may be equivalent when all predictors in the logistic regression are categorical. If continuous predictors are present a loglinear model is not applicable. Not all loglinear models are equivalent to logistic regressions with categorical predictors. A logistic regression with k categorical predictors is equivalent to a loglinear model in which one of the predictors is dichotomous and can be considered the outcome. This categorical predictor is the variable that defines success or failure in the logistic regression. Fitting a loglinear model with a saturated model of the remaining predictors (those not defining the dichotomous outcome) produces models that are exact equivalents to a logistic regression model. For instance, in a model with three dichotomous categorical variables: gender (G), treatment (T) and mortality (M), mortality can be modeled as a dichotomous outcome in a logistic regression:

$$\ln\left(\frac{P(M=1)_i}{P(M=0)_i}\right)=b_0+b_1G_i+b_2T_i$$

This would be equivalent to the following loglinear model:

$$\ln\left(y_{ijk}\right)=\lambda+\lambda_i^G+\lambda_j^T+\lambda_k^M+\lambda_{ij}^{G\times T}+\lambda_{ik}^{G\times M}+\lambda_{jk}^{T\times M}$$

Provided the same coding scheme is employed, the coefficients and deviance of the fitted model should match. The key distinction is that a loglinear model provides a more flexible set of hypotheses to test than a logistic regression. In the latter it is not possible to model the interrelationships between predictors.

For this reason logistic regression should be preferred when there is an unambiguous dichotomous outcome variable. If the 'outcome' is not dichotomous, multinomial logistic regression can be used (and will be equivalent to a loglinear model with a saturated model of the categorical predictors from the multinomial logistic regression). With respect to Poisson regression, a loglinear model has fewer advantages. The main one is the clear link to analysis of two-way contingency tables. As contingency tables are familiar to many researchers, it may be convenient to present a Poisson model with categorical predictors as a loglinear model. There are also some extensions to loglinear models that make them useful for particular applications (e.g., see Agresti, 1996). On the other hand, Poisson regression is more versatile, can incorporate continuous predictors and be readily extended to cope with zero-inflated and overdispersed data.

Example OS5.3 The surgical checklist analysis, run earlier as a Poisson or negative binomial regression (with mortality or mortality rates as the outcome), could also be run as loglinear model. This would be a Poisson model with cell count (alive or dead) as the outcome. The predictors would be the *hospital*, *mortality* (alive or dead) and *time* (pre or post checklist). In this model the effects of key interest are the time × mortality and the hospital × time × mortality interactions. The model with all two-way effects (AIC = 240.0) is marginally preferred to the three-way model (AIC = 240.6). The three-way model suggests that the checklist may be more effective at preventing deaths in some hospitals than others. The model is also overdispersed, but less so than the earlier models ($\hat{\varphi}=1.47$). Just to be safe, results for a quasipoisson model will be reported, rather than the usual Poisson model. The estimate of the interaction effect is 0.64. This suggests that mortality was $e^{0.64}=1.9$ times higher before the checklist was introduced than after its introduction. The profile likelihood 95% CI for the time × mortality interaction in the quasipoisson model (on the untransformed count scale) is [1.12, 3.30]. So a conservative estimate would be a 10% or 11% reduction in surgical deaths after the introduction of the checklist (as $1/1.12=0.89$).

This analysis could also be run as a logistic regression with mortality as a binary outcome and 7688 cases. The loglinear model is easier to set up (because it has only 32 cases) but should give identical results. The loglinear model here is superior to the earlier Poisson regression, though this stems from the way that the model structures the data. It isn't an intrinsic property of the analysis (as both are forms of Poisson regression).

OS5.2 R code for Online Supplement 5

OS5.2.1 Loglinear models for two-way tables (Example OS5.1)

Example OS5.1 shows how to re-analyze the dream data as a loglinear model. This can be done using a loglinear modeling function such as `loglm()` in `MASS`. The `table()` function from the base package can be used to turn the categorical predictor `group` and the dichotomous outcome `scary` into a contingency table:

```
group <- c(rep(0,26), rep(1,21))
scary <- c(1, rep(0,25), rep(1,7), rep(0,14))

sub09 <- table(group, scary)
```

The `loglm()` can take either dimension names of the table or integers that stand as short-hand for the 1st, 2nd, 3rd or greater dimensions of the table. The following models are equivalent, and fit a saturated model to a two-way contingency table. No outcome (*Y*) variable is required (because loglinear models treat the cell counts as outcomes).

```
library(MASS)
loglm(~ 1*2, sub09)
loglm(~ group * scary, sub09)
```

A model with main effects only, dropping the two-way interaction, tests the independence of the two predictors and reports both the Pearson and likelihood or deviance chi-square.

```
loglm(~ group + scary, sub09)
```

The parameters for predictors under effect coding are:

```
loglm(~ group*scary, sub09)$param
```

It is possible to fit a loglinear model using Poisson regression also. In this case the counts and predictors need to be specified as vectors. This also makes it easy to switch to dummy coding:

```
count <- as.vector(sub09)
gp <- c(0,0,1,1)
sc <- c(0,1,0,1)

glm(count ~ gp + sc, family=poisson)
```

Fitting the model using different functions may (as here) produce different loglikelihood, AIC and so forth (because the arbitrary constant has changed). However, the inferences should be identical. For instance, compare differences in AIC for these models:

```
AIC(glm(count ~ gp * sc - gp:sc, family=poisson)) -
  AIC(glm(count ~ gp * sc, family=poisson))

loglm(~ group + scary, sub09)[[1]]-2
```

The coefficients for the dummy coded saturated model in count and odds ratio form are:

```
glm(count ~ gp*sc, family=poisson)$coef
exp(glm(count ~ gp*sc, family=poisson)$coef)
```

The CIs are also easy to obtain this way:

```
confint(glm(count ~ gp*sc, family=poisson))
exp(confint(glm(count ~ gp*sc, family=poisson)))
```

While this does demonstrate the basics of loglinear modeling, only rarely will it be useful to model a two-way table in this way. Loglinear modeling is best reserved for tables with three or more categorical variables.

OS5.2.2 Modeling counts in three-way tables (Examples OS5.2 and OS5.3)

The data in Example OS5.2 are from Appleton *et al*. (1996). These can be loaded in from the SPSS data file `smoking.sav`.

```
library(foreign)
smoking <- read.spss('smoking.sav', to.data.frame=TRUE)
```

The structure of the data frame lends itself to an analysis using Poisson regression. The likelihood ratio test and AIC for models with and without the two-way interaction are given by `drop1()` or `anova()`.

```
glm(count ~ smoker*survival, family=poisson, data=smoking)

drop1(glm(count ~ smoker*survival, family=poisson,
  data=smoking), test='Chisq')
```

A better analysis includes the age group as a predictor in a three-way model. The following commands fit a saturated three-way model, a model with all two-way effects or lower. Age group is coded by the numbers one to seven, so it is necessary to specify that it should treated as a factor.

```
smok.sat <- glm(count ~ smoker * survival * factor(age_group),
  family = poisson, data=smoking)

smok.2way <- glm(count ~ (smoker + survival +
  factor(age_group))^2, family=poisson, data=smoking)
```

The `drop1()` function can compare the fits of these two models or they can be compared directly using AIC.

```
drop1(smok.sat, test='Chisq')
AIC(smok.sat) - AIC(smok.2way)
```

There is little reason to keep the three-way term, but dropping any of the two-way terms results in a worse fit.

```
drop1(smok.2way, test='Chisq')
```

The parameter estimates for the model can be easily listed using `summary()`.

```
summary(smok.2way)
```

The CI (on the count scale) for the smoking by survival interaction is given by:

```
exp(confint(smok.2way, 10))
```

(Although R reports several warnings about the CIs, the profile CIs are very similar to the Wald CIs and it would be reasonable to report either.) Checking the quasipoisson fit shows that the data appear to be underdispersed:

```
smok.q <- glm(count ~ (smoker + survival +
  factor(age_group))^2, family = quasipoisson, data=smoking)

summary(smok.q)
```

Example OS5.3 demonstrated that the checklist data analyzed in several earlier examples could also be considered as a three-way loglinear model. This requires a restructured data set to be loaded:

```
checklist.tab <- read.csv('checklist_tab.csv')
```

The three-way and two-way models can then be fitted.

```
mort.3way <- glm(count ~ factor(hospital)*post*survival,
  data=checklist.tab, family=poisson)

mort.2way <- glm(count ~ (factor(hospital) + post +
  survival)^2, data=checklist.tab, family=poisson)
```

The model with only two-way effects provides a pretty good fit given that it has seven fewer parameters, but none of the two-way effects should be dropped:

```
drop1(mort.3way, test='Chisq')
drop1(mort.2way, test='Chisq')
```

To allow for overdispersion, either a negative binomial regression or a quasipoisson regression could be fitted:

```
mort.2way.q <- glm(count ~ (factor(hospital) + post +
  survival)^2, data=checklist.tab, family=quasipoisson)

mort.2way.nb <- glm.nb(count ~ (factor(hospital) + post +
  survival)^2, data=checklist.tab)
```

Both models produce similar parameter estimates though the quasipoisson, but not the negative binomial, produces more conservative CIs for the effect of the checklist on surgical deaths:

```
exp(confint(mort.2way, 25))
exp(confint(mort.2way.q, 25))
exp(confint(mort.2way.nb, 25))
```

OS5.2.3 R packages

R-core members, DebRoy, S., Bivand, R., *et al*. (2011). *foreign*: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase. R package version 0.8-42.

Venables, W. N., and Ripley, B. D. (2002) *MASS*: Modern Applied Statistics with S. (4th edn) Springer: New York.

OS5.3 Notes on SPSS syntax for Online Supplement 5

OS5.3.1 Loglinear models

SPSS can fit loglinear models via several routes (and often the generalized linear model commands for Poisson regression will be a sensible choice). The `GENLOG` command provides a specific loglinear model command. The syntax below fits a saturated model to the smoking data:

SPSS data file: `smoking.sav`

```
GENLOG age_group smoker survival
  /MODEL=POISSON
  /PRINT=FREQ
  /PLOT=NONE
  /CRITERIA=CIN(95).
```

This can be compared to a simpler model such as this one that includes all two-way or simpler effects by adding a `/DESIGN` subcommand.

```
GENLOG age_group smoker survival
  /MODEL=POISSON
  /PRINT=FREQ RESID
  /PLOT=NONE
  /CRITERIA=CIN(95)
  /DESIGN age_group*smoker age_group*survival smoker*survival.
```

The deviance for this model is the residual deviance (the change in deviance from the saturated model) and a test of the three-way interaction effect.

OS5.4 Notes

1. You may wish to refresh your memory about the χ^2 of independence at this point, as it may make later material easier to follow (see Chapter 4).
2. A notation more consistent with that used earlier would be $\ln\left(Y_{ij}\right) = b_0 + b_1 A_{1i} + b_2 B_{2j}$. The notation for loglinear models is more compact. This is useful for models involving many interaction terms.
3. These parameter estimates were obtained using either a sigma-restricted or overparameterized model that produce coefficients equivalent to effect and dummy coding respectively.
4. As loglinear models tend to have relatively few categorical predictors, automatic selection methods are not as dangerous as for regressions with many predictors. Hierarchical comparison of a subset of theoretically interesting models (e.g., using information criteria) is nevertheless the preferred approach here.
5. Another victory for Stigler's law; it was observed earlier by both Karl Pearson and Yule.

OS5.5 References

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: Wiley.

Appleton, D. R., French, J. M., and Vanderpump, M. P. J. (1996) Ignoring a Covariate: An Example of Simpson's Paradox, *The American Statistician*, 50, 340–1.

Darlington, R. B. (1968) Multiple Regression in Psychological Research and Practice, *Psychological Bulletin*, 69, 161–82.

Lord, F M. (1967) A Paradox in the Interpretation of Group Comparisons. *Psychological Bulletin*, 68, 304–5.

Tu, Y.-K., Gunnell, D. J. and Gilthorpe, M. S. (2008) Simpson's Paradox, Lord's Paradox, and Suppression Effects are the Same Phenomenon: The Reversal Paradox. *Emerging Themes in Epidemiology*, 5, 2.