STATISTICS · TESTS AND MODELS

Categorical Data Analysis and Log-Linear Models for Contingency Knowledge

How log-linear models extend the chi-square test to tables with three or more variables, separate real associations from spurious ones, and support sound conclusions from survey cross-tabulations.

ModQu research article · 1,368 words · 14 references (APA 7th edition)

Most survey data are categorical: a region, an age band, a preferred channel, a yes or a no. The natural summary is a contingency table that counts how many respondents fall into each combination of categories. For two variables, the chi-square test of independence answers the basic question of whether they are related (Pearson, 1900). Real questions rarely stop at two variables, however. A marketing team may want to know whether channel preference is linked to loyalty-programme membership once age is taken into account. **Log-linear models** are the classical framework for answering such questions about multiway tables (Agresti, 2013; Bishop et al., 1975).

Why two-way tables are not enough

A two-way table collapses over every variable that is not shown. That can create associations that do not exist within any subgroup, hide associations that do, or even reverse their direction. Simpson (1951) showed formally how an association present in each subtable can disappear or change sign when the subtables are added together, the effect now known as Simpson's paradox. The cure is to model the full table, so that each association is judged while the other variables are held fixed.

The log-linear model

A log-linear model describes the expected count in each cell of the table. For three variables A, C and L, with categories indexed by *i*, *j* and *k*, the most general model is:

$$\log \mu_{ijk} = \lambda + \lambda_i^{A} + \lambda_j^{C} + \lambda_k^{L} + \lambda_{ij}^{AC} + \lambda_{ik}^{AL} + \lambda_{jk}^{CL} + \lambda_{ijk}^{ACL}$$

The single-variable terms reproduce the totals for each variable. The two-way terms capture associations between pairs of variables, and the three-way term allows an association between two variables to differ across the levels of the third. Including every term gives the saturated model, which reproduces the observed counts exactly and so explains nothing. The analyst's task is to find the simplest model that still fits.

Two conventions keep the models interpretable. First, models are usually **hierarchical**: if a two-way term is included, the corresponding single-variable terms are included too. Second, a hierarchical model can be named by its highest-order terms. The model [AC][AL], for example, contains the age-channel and age-loyalty associations but no channel-loyalty term. It states that channel and loyalty are independent within each age group, a conditional independence statement that a two-way table cannot test.

Log-linear models were developed in the 1960s and 1970s, building on Birch's (1963) results on maximum likelihood estimation for three-way tables. Goodman (1970) set out a systematic approach to analysing interactions among several categorical variables, and Bishop et al. (1975) brought the theory together in a standard reference. Today they are usually fitted as Poisson regressions within the family of generalised linear models, in which the cell counts are the outcome and the categorical variables are the predictors (Nelder & Wedderburn, 1972; McCullagh & Nelder, 1989). Many hierarchical models can also be fitted by iterative proportional fitting, which repeatedly scales a table until it matches the required margins, the algorithm Deming and Stephan (1940) introduced and survey raking still uses.

Reading the parameters

The two-way terms have a direct interpretation. In a table where both variables have two categories, the association term determines the log of the odds ratio between them. A two-way term of zero corresponds to an odds ratio of 1, meaning no association once the other variables in the model are held constant.

Judging the fit

Each unsaturated model implies a set of fitted counts. Two statistics compare these with the observed counts: Pearson's chi-square statistic, X^2, and the likelihood-ratio statistic, G^2, also called the deviance:

$$G^2 = 2 \sum n_{ijk} \log(n_{ijk} \div \hat{\mu}_{ijk})$$

Both are referred to a chi-square distribution with degrees of freedom equal to the number of terms the model leaves out. A small G^2 relative to its degrees of freedom means the model is consistent with the data. Because G^2 is additive across nested models, the difference between two nested models tests the terms that separate them (Agresti, 2013).

Overall fit statistics should be followed by a look at individual cells. Haberman (1973) proposed adjusted residuals that are approximately standard normal when the model holds, so that values beyond about ±2 point to cells the model describes poorly.

A worked example

Consider a hypothetical survey of 1,200 customers classified by age (A: under 40 or 40 and over), preferred shopping channel (C: online or in store) and loyalty-programme membership (L: yes or no).

Age	Channel	Loyalty: yes	Loyalty: no	Membership rate
Under 40	Online	214	206	51.0%
Under 40	In store	86	94	47.8%
40 and over	Online	57	123	31.7%
40 and over	In store	123	297	29.3%

Collapsed over age, 45.2% of online shoppers are loyalty members, compared with 34.8% of in-store shoppers, an odds ratio of about 1.54. A business reading might conclude that moving customers online raises loyalty sign-ups.

Figure 1 shows what a sequence of log-linear models says instead. Mutual independence, [A][C][L], fits very badly. Adding the age-channel association, [AC][L], still leaves a large misfit. The model [AC][AL], which allows age to be related to both channel and loyalty but makes channel and loyalty independent within each age group, fits well: $G^2 = 0.85$ on 2 degrees of freedom, $p = 0.65$. Adding the channel-loyalty term improves the fit by only 0.85 on 1 degree of freedom, which is not significant.

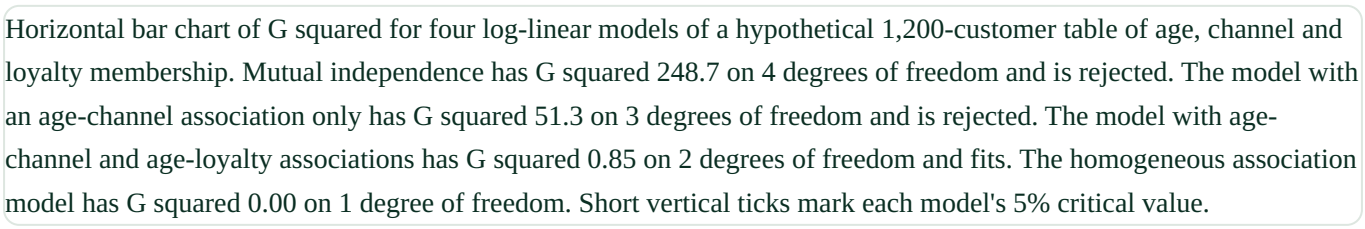


Figure 1. Fit of four hierarchical log-linear models to the age × channel × loyalty table. Models whose G^2 falls below the 5% critical value are consistent with the data. Illustrative data; chart by ModQu.

The explanation lies in age. Younger customers are much more likely to shop online (odds ratio about 5.4) and also more likely to join the programme (odds ratio about 2.3). Within each age group, the channel-loyalty odds ratio is only about 1.1. The apparent channel effect in the two-way table is largely age in disguise.

Log-linear models and logistic regression

When one variable is clearly an outcome, as loyalty membership is here, a logistic regression of that outcome on the other variables gives equivalent answers to the corresponding log-linear model, provided the log-linear model includes all associations among the predictors (Agresti, 2013). Log-linear models remain valuable when no single variable is the outcome, when the structure of association itself is of interest, or when modelling cell counts directly is natural, as in the analysis of mobility, brand-switching or repeated-choice tables.

Practical guidance for survey and business data

- **Account for the survey design.** Clustering and weighting make the usual X^2 and G^2 tests too liberal. Rao and Scott (1981, 1984) developed corrections for two-way and multiway tables that are now standard in survey software.

- **Watch for sparse tables.** Adding variables multiplies the number of cells quickly. Many empty or small cells make the chi-square approximation unreliable and can produce infinite parameter estimates. Combine categories with a substantive rationale or use exact and small-sample methods (Fagerland et al., 2017).
- **Visualise the table.** Mosaic displays, which draw each cell as a tile with area proportional to its count and colour by residual, make patterns of association visible at a glance (Friendly, 1994).
- **Prefer the simplest model that fits** and report its key associations as odds ratios with confidence intervals.

Common mistakes

- **Collapsing too early.** Drawing conclusions from two-way tables when a third variable is related to both.
- **Searching without limits.** With many variables, a mechanical search through hundreds of possible models invites chance findings. Begin from substantive hypotheses.
- **Equating fit with truth.** A model that fits is consistent with the data, not proven. Large samples reject small, unimportant departures; small samples fail to reject large ones.
- **Interpreting associations as causes.** Log-linear models describe structure in the table. Causal claims require design or further assumptions.

Key takeaways

Log-linear models extend the chi-square test from two variables to many. They express the structure of a contingency table as a set of associations, test each one while holding the others fixed, and reveal conditional independence that two-way tables cannot show. Used with design-corrected tests, attention to sparse cells and clear reporting of odds ratios, they turn survey cross-tabulations into reliable evidence about how categorical characteristics really relate.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley.

Birch, M. W. (1963). Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 25*(1), 220–233. https://doi.org/10.1111/j.2517-6161.1963.tb00504.x

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. MIT Press.

Deming, W. E., & Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics, 11*(4), 427–444. https://doi.org/10.1214/aoms/1177731829

Fagerland, M. W., Lydersen, S., & Laake, P. (2017). *Statistical analysis of contingency tables*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315374116

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association, 89*(425), 190–200. https://doi.org/10.1080/01621459.1994.10476460

Goodman, L. A. (1970). The multivariate analysis of qualitative data: Interactions among multiple classifications. *Journal of the American Statistical Association, 65*(329), 226–256. https://doi.org/10.1080/01621459.1970.10481076

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics, 29*(1), 205–220. https://doi.org/10.2307/2529686

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman and Hall. https://doi.org/10.1007/978-1-4899-3242-6

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: General, 135*(3), 370–384. https://doi.org/10.2307/2344614

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Rao, J. N. K., & Scott, A. J. (1981). The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the American Statistical Association, 76*(374), 221–230. https://doi.org/10.1080/01621459.1981.10477633

Rao, J. N. K., & Scott, A. J. (1984). On chi-squared tests for multiway contingency tables with cell proportions estimated from survey data. *The Annals of Statistics, 12*(1), 46–60. https://doi.org/10.1214/aos/1176346391

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 13*(2), 238–241. https://doi.org/10.1111/j.2517-6161.1951.tb00088.x