Effects of Exposure Misspecification in Log-linear Models for Rates

John W. McDonald and Peter W. F. Smith

Social Statistics and Southampton Statistical Sciences Research Institute
University of Southampton
Southampton, SO17 1BJ
United Kingdom

Keywords: log-linear model, misspecification, measurement error, offset term, rate

Paper prepared for Session 32: Modeling Issues in Statistical Demography at the Annual Meeting of the Population Association of America, Boston, April 1-3, 2004.

1 Introduction

Log-linear models are used in demography and epidemiology to model rates cross-classified by explanatory factors. A rate is defined as the ratio of the number of events of interest to the exposure. For example, for mortality rates, the exposure is total person-years at risk. The rates are not modelled directly, but a table of counts of events of interest and a table of exposures are required. Often, the true exposure is unknown and an estimate of exposure is used. The problem we study in this paper is the effects of misspecification of the exposures on inferences drawn from a log-linear model for rates. Our literature review suggests that the effects of this type of misspecification have not been investigated. In particular, we investigate how this type of rates model misspecification affects parameter estimates, estimated standard errors, confidence intervals, test statistics, etc.

To fix ideas, we first consider a simple example of comparing two rates, taken from Agresti (2002, Problem 9.17). We wish to compare the motor vehicle accident rate of men and women aged 65 to 84, who had a valid driver's license during the study period 1984–1988. The women had 175 accidents and the men had 320. While Agresti provides the person-years of observation, we are interested in what inferences can be drawn without knowing the true exposures or with only partial information, e.g., the ratio of exposures. We then consider 2-way and 3-way tables. We discuss the implications of our results to a 2-way table of lung cancer cases by age in four Danish cities, previously analyzed by Andersen (1977). For a 3-way table, we also consider the situation where one dimension of the table is time, but the exposures are only known at one time point. We discuss the implications of our results for this situation with a 3-way table of fatal fire casualties in the UK for the years 1969 to 1973 cross-classified by age and sex. Here the population at risk for the various age categories for each sex is known only for the Census year 1971.

2 Log-linear Models for Rates

Following the notation of McDonald, Smith and Forster (1999), let $\boldsymbol{Y} = \{Y_i : i = 1, \ldots, n\}$ be a vector of independent counts of events of interest, e.g., deaths, with $Y_i \sim \text{Poisson}(\lambda_i\, e_i)$, where λ_i is the rate for population i and the e_i are *fixed* (non-random), *known* constants termed rate multipliers. Then the log-linear model for rates

$$\log(\lambda_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} \qquad i = 1, \ldots, n,$$

can be written as the log-linear model for expected counts as

$$\log E(Y_i) = \log(\lambda_i\, e_i) = \boldsymbol{x}_i^T \boldsymbol{\beta} + \log e_i \qquad i = 1, \ldots, n,$$

where $\log e_i$ is a known 'offset', $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ is a vector of parameters corresponding to the vector of covariates $\boldsymbol{x}_i^T = (x_{i1}, \ldots, x_{in})$, and T denotes the transpose. An offset is a term in the linear predictor with coefficient set equal to one, rather than estimated.

3 One-way Tables

We first consider the simple example of comparing two rates. We consider testing that the two rates are equal using the Wald test, a conditional test, the score test and the likelihood ratio test. Note that, in general, inferences based on these alternative procedures may yield different results.

Here the data are often presented in two one-way tables, the first containing the counts of the events of interest and the second the exposures or in the following more condensed format with the observed rates.

	population	
	1	2
events	y_1	y_2
exposure	e_1	e_2
observed rate	y_1/e_1	y_2/e_2

In this situation, the log-linear rates model has the form $\log E(Y_1) = \beta_1 + \log e_1$ and $\log E(Y_2) = \beta_1 + \beta_2 + \log e_2$, where $\boldsymbol{x}_1^T = (1, 0)$ and $\boldsymbol{x}_2^T = (1, 1)$ and the hypothesis of equal rates is $H_0 : \beta_2 = 0$.

3.1 Wald Test

It is straightforward to show that the maximum likelihood estimates (mles) are $\widehat{\beta}_1 = \log(y_1/e_1)$ and $\widehat{\beta}_2 = \log(y_2/e_2) - \log(y_1/e_1) = \log(\frac{y_2/e_2}{y_1/e_1}) = \log(\frac{y_2\, e_1}{y_1\, e_2})$. Note that $\exp(\widehat{\beta}_1) = y_1/e_1$, i.e., the observed rate in population 1 and $\exp(\widehat{\beta}_2) = \frac{y_2/e_2}{y_1/e_1}$, i.e., the observed rate ratio (population 2 versus population 1). It is also straightforward to show that the estimated asymptotic standard error of $\widehat{\beta}_2$ is $\sqrt{1/y_1 + 1/y_2}$. Hence, the Wald test statistic for $H_0 : \beta_2 = 0$ is

$$\frac{\log(\frac{y_2\, e_1}{y_1\, e_2})}{\sqrt{\frac{1}{y_1} + \frac{1}{y_2}}}$$

and a 95% confidence interval for β_2 is

$$\log(\frac{y_2\, e_1}{y_1\, e_2}) \pm 1.96 \sqrt{\frac{1}{y_1} + \frac{1}{y_2}}.$$

Note that only the ratio of exposures e_1/e_2 is required to correctly calculate these statistics. Therefore, the length of the confidence interval does not depend at all on the individual exposures and the location only on their ratio! The effect of not knowing the exposures is only to shift the location of the interval, if $e_1/e_2 \neq 1$.

If the exposures are unknown then, for a given test size, 5% say, we can calculate a range of values for the ratio of exposures for which H_0 will be accepted:

$$\frac{y_1}{y_2}\exp(-1.96\sqrt{\frac{1}{y_1}+\frac{1}{y_2}}) \leq \frac{e_1}{e_2} \leq \frac{y_1}{y_2}\exp(1.96\sqrt{\frac{1}{y_1}+\frac{1}{y_2}}).$$

For the Agresti (2002, Problem 9.18) example, with $y_1 = 175$ and $y_2 = 320$, this gives

$$0.416 \leq \frac{e_1}{e_2} \leq 0.599.$$

Hence, provided that e_1, the person years of observation for women, is at least 59.9% of e_2, the person years of observation for men, we would reject H_0 and conclude that the rate for men is greater than that for women. On the other hand, if e_1 is less than 41.6% of e_2 we would conclude that the rate is greater for women. The exposure for women was 17.3 thousand years of observation and for men 21.4 thousand years, so the ratio of exposures is $e_1/e_2 = 17,300/21,400 = .808$, so that the person years of observation for women was 80.8% of that of men.

3.2 Conditional Test

Alternatively, a conditional test may be used to test the equality of rate parameters, i.e., $H_0 : \lambda_1 = \lambda_2$. Here $Y_1 \sim \text{Poisson}(\lambda_1\, e_1)$ and $Y_2 \sim \text{Poisson}(\lambda_2\, e_2)$ and the conditioning is on the total number of events, $Y_1 + Y_2$. The conditional distribution of Y_1, given $Y_1 + Y_2 = t$ is binomial:

$$Y_1 \mid Y_1 + Y_2 = t \sim \text{binomial}\,(t, \frac{\lambda_1\, e_1}{\lambda_1\, e_1 + \lambda_2\, e_2})$$

where the probability of "success" depends only on the fraction of total exposure in population 1, i.e., $e_1/(e_1+e_2)$, when the rates are equal. Equivalently,

the conditional distribution only depends on the ratio of exposures, when the rates are equal:

$$Y_1 \mid Y_1 + Y_2 = t \sim \text{binomial}\,(t, \frac{1}{1 + (\lambda_2/\lambda_1)\,(e_2/e_1)}).$$

3.3 Score Test

Alternatively, a score test may be used to test $H_0 : \beta_2 = 0$. The score (gradient of the log likelihood at the null value of the parameter) is

$$U = Y_1 - (Y_1 + Y_2)\,(\frac{e_2}{e_1 + e_2})$$

and the score variance (minus the curvature of the log likelihood also at the null value) is

$$V = (Y_1 + Y_2)\,(\frac{e_2}{e_1 + e_2})\,(\frac{e_1}{e_1 + e_2}).$$

The score test statistic is U^2/V, which has asymptotically a χ^2_1 distribution under $H_0 : \beta_2 = 0$. For the score test, inference about the rate ratio only depends on the ratio of exposures or equivalently, the fraction of total exposure.

3.4 Likelihood Ratio Test

Note also that the conditional likelihood and profile likelihood of β_2 only depends on the ratio of exposures or equivalently, the fraction of total exposure (Clayton & Hills, 1993). Here the conditioning is on the total number of events and the conditional log likelihood equals the profile log likelihood. These log likelihoods, up to a constant of proportionality, are

$$y_2 \log(\theta\,\frac{e_2}{e_1}) + (y_1 + y_2) \log(1 + \theta\,\frac{e_2}{e_1}),$$

where $\theta = \exp(\beta_2) = \lambda_2/\lambda_1$ is the rate ratio. Hence, the likelihood ratio test statistic for H_0 again only depends on the ratio of exposures (or equivalently, the fraction of total exposure) since $\theta = 1$ under H_0 and $\widehat{\theta} = y_2 e_1/y_1 e_2$ under the unconstrained alternative hypothesis.

4 Two-way Tables

Now consider a random two-way table, $Y = \{Y_{ij}\}$ of independent counts of events of interest, with $Y_{ij} \sim \text{Poisson}(\lambda_{ij}\, e_{ij})$, where λ_{ij} is the rate and e_{ij} are known exposures for the ij cell of the $R \times C$ table, where $i = 1, \ldots, R$ and $j = 1, \ldots, C$. The multiplicative rates model (no interaction model) corresponds to the hypothesis that $\lambda_{ij} = c \times a_i \times b_j$. As

$$\lambda_{ij} = c \times a_i \times b_j = c \times a'_i \times b'_j$$

for $a'_i = k \times a_i$ and $b'_j = b_j/k$, we require parameter constraints for identifiability. Set $a_1 = 1$ and $b_1 = 1$ which implies that $\lambda_{11} = c$ and $\lambda_{ij} = \lambda_{11} \times a_i \times b_j$ and thus c is interpreted as the rate in the (1,1) cell, which is called the baseline or reference cell. These parameter constraints are termed baseline or corner-point constraints. The multiplicative rates model (no interaction model) corresponds to all cross-product ratios of rates equalling one. For example, for a 2×2 table, $\lambda_{11}\lambda_{22}/\lambda_{12}\lambda_{21} = 1$.

Commonly, inference about a two-way table of counts is made conditional on either the table total, one or both of the marginal totals, using respectively, the multinomial, product-multinomial and hypergeometric distributions. As is well-known, for a two-way table of multinomially distributed counts y_{ij}, the additive log-linear model without offset corresponds to the hypothesis of independence. In this situation, the mles of the cell means m_{ij} can be expressed in closed form

$$\widehat{m}_{ij} = y_{i+}\, y_{+j}/y_{++},$$

where + denotes summation over the corresponding subscript. Also, closed form mles of the row and column main effect parameters can be written as logarithms of the ratios of the row and column marginal totals respectively. The exact expression depends on the parameterization of the log-linear model. It is little known that for the additive log-linear rates model where the exposures have a multiplicative form, i.e., $e_{ij} = e_{i+}\, e_{+j}/e_{++}$, closed form mles of the parameters exist (Hoem, 1995). Closed form mles of the cell means m_{ij} and cell rates λ_{ij} are

$$\widehat{m}_{ij} = y_{i+}\, y_{+j}/y_{++}$$

and

$$\widehat{\lambda}_{ij} = (y_{i+}\, y_{+j}/y_{++})/(e_{i+}\, e_{+j}/e_{++})$$

(see the Appendix for further details). Closed form mles of the row and column main effect parameters are logarithms of the ratios of the row and column marginal rates respectively. Again, the exact expression depends on the parameterization of the log-linear model. The asymptotic standard error of the row and column main effects for baseline parameter constraints is the square root of the reciprocals of the marginal totals of the event counts.

4.1 2×2 Table

Now consider a 2×2 table and the saturated log-linear model with offset $\log e_{ij}$,

$$\log E(Y_{ij}) = \beta + \beta_i^1 + \beta_j^2 + \beta_{ij}^{12} + \log e_{ij} \qquad i = 1, 2;\ j = 1, 2, \qquad (1)$$

where the $\{\beta_i^1\}$ are the main effects for factor 1, the $\{\beta_i^2\}$ are the main effects for factor 2 and the $\{\beta_{ij}^{12}\}$ are the two-factor effects or interactions between factors 1 and 2. For identifiability, we set $\beta_1^1 = \beta_1^2 = \beta_{i1}^{12} = \beta_{1j}^{12} = 0$. The sufficient statistics are y_{++} for β, y_{i+} for β_i^1 and y_{+j} for β_j^2. The null hypothesis of $H_0 : \beta_{ij}^{12} = 0$ corresponds to the multiplicative rates model (no interaction model). When the exposures also have a multiplicative form, the mles under $H_0 : \beta_{ij}^{12} = 0$ are

$$\widehat{\beta} = \log[(y_{1+}\, y_{+1}/y_{++})/(e_{1+}\, e_{+1}/e_{++})],$$

$$\widehat{\beta_2^1} = \log[\, y_{2+}/e_{2+}] - \log[\, y_{1+}/e_{1+}] = \log[\, (y_{2+}/e_{2+})/(\, y_{1+}/e_{1+})]$$

and

$$\widehat{\beta_2^2} = \log[\, y_{+2}/e_{+2}] - \log[\, y_{+1}/e_{+1}] = \log[\, (y_{+2}/e_{+2})/(\, y_{+1}/e_{+1})]$$

The estimated asymptotic standard error of $\widehat{\beta_2^1}$ is $\sqrt{1/y_{1+} + 1/y_{2+}}$ and of estimated asymptotic standard error of $\widehat{\beta_2^2}$ is $\sqrt{1/y_{+1} + 1/y_{+2}}$. It is important to note that all these closed form parameter estimates only depend on the marginal counts of events and marginal exposures, i.e., the marginal rates, and the mles of the main effects depend only on the ratios of the marginal exposures. Also, the estimated asymptotic standard errors only depend on the marginal counts of events and not the exposures.

4.2 Effects of Omitting a Covariate in Poisson Models

Note that for a $R \times C$ table and more generally, the estimates of the effects of interest and of their standard errors are unaffected when a covariate is removed from the Poisson multiplicative rates model when the exposures have a multiplicative form, e.g., the estimate of the effect of factor 1 is unaffected by whether factor 2 is in the model or not (and vice versa). Petersen and Deddens (2000) proved this result for the case of balanced Poisson distributed data if no offset variable is included. Proceeding to analyze the effect of factor 1 without taking into account the presence of factor 2 is model misspecification since factor 2 is assumed to influence the rate. Here we have harmless model misspecification, see Hoem (1995).

There is a close analogy with analysis of variance with the same number of observations in each cell of the table (balanced data). It is well known that for balanced data the test sum of squares and parameter estimates are based on an orthogonal decomposition of the data vector. In this situation, we do not need to adjust for factor 1 in order to estimate the effect of factor 2 and to calculate the test sum of squares that factor 1 has no effect. For the unbalanced case, where we have unequal numbers of observations in the cells of the table, we must usually adjust for the effect of factor 1 in order to assess the effect of factor 2 and to calculate the test sum of squares to test that factor 2 has no effect. There is one exception to this, the case of proportional numbers of observations, where the cell numbers in any two rows (or columns) are proportional. In this case, the decomposition of the data vector is again orthogonal (Scheffe, 1959) as the multiplicative cell counts are equivalent to independent factors 1 and 2. In the analysis of rates, multiplicative exposures have a role quite analogous to the situation of multiplicative cell counts in ANOVA and in this situation, we have harmless model misspecification.

Petersen and Deddens (2000) note that in linear models, when a covariate is omitted from the model, that the sum of squares for that effect is pooled with the error. If the F-statistic for the omitted effect is greater than one, the results in a larger standard error for the remaining parameter estimates. Hence, in the linear model case, if the data are balanced, the estimated parameter estimate will be the same as that in the full model, but the standard error will be changed. In contrast, the estimates of effects of interest *and* their standard errors are unaffected when a covariate is removed from a Poisson multiplicative rates model when the exposures have a multiplicative

form. One obvious use of this result is the situation when only the marginal exposures are known, but the analyst is willing to assume a multiplicative form for the exposures. In this situation, the actual cell exposures are not needed for the analysis.

4.3 Wald Test

Testing goodness of fit of the multiplicative rates model corresponds to testing $H_0 : \beta_{ij}^{12} = 0$ for model (1) or equivalently $\lambda_{11}\lambda_{ij}/\lambda_{1j}\lambda_{i1} = 1$ for all ij. It is straightforward to show that the mle of β_{ij}^{12} is

$$\widehat{\beta_{ij}^{12}} = \log[\frac{(y_{11}/e_{11})(y_{ij}/e_{ij})}{(y_{1j}/e_{ij})(y_{i1}/e_{i1})}] = \log[\frac{y_{11}\, y_{ij}}{y_{1j}\, y_{i1}}] - \log[\frac{e_{11}\, e_{ij}}{e_{ij}\, e_{i1}}].$$

Note that this estimate is the log of the crossproduct ratio of the observed rates or equivalently, the log of the crossproduct ratio of the observed counts minus the log of the crossproduct ratio of the exposures. Note also that $\exp(\widehat{\beta_{ij}^{12}})$ is the crossproduct ratio of the observed rates. It is also straightforward to show that the estimated asymptotic standard error of $\widehat{\beta_{ij}^{12}}$ is $\sqrt{1/y_{11} + 1/y_{ij} + 1/y_{1j} + 1/y_{i1}}$. Hence, the Wald test statistic for $H_0 : \beta_{ij}^{12} = 0$ is

$$\frac{\log[\frac{y_{11}\, y_{ij}}{y_{1j}\, y_{i1}}] - \log[\frac{e_{11}\, e_{ij}}{e_{ij}\, e_{i1}}]}{\sqrt{1/y_{11} + 1/y_{ij} + 1/y_{1j} + 1/y_{i1}}}$$

and a 95% confidence interval for β_{ij}^{12} is

$$(\log[\frac{y_{11}\, y_{ij}}{y_{1j}\, y_{i1}}] - \log[\frac{e_{11}\, e_{ij}}{e_{ij}\, e_{i1}}]) \pm 1.96\, \sqrt{1/y_{11} + 1/y_{ij} + 1/y_{1j} + 1/y_{i1}}.$$

Note that only the crossproduct ratio of exposures is required to correctly calculate these statistics. Therefore, the length of the confidence interval does not depend at all on the individual exposures, and the location only on their crossproduct ratio! If the exposures have a multiplicative form, the log crossproduct ratio of the exposures term in the Wald statistic is zero and hence, the Wald test does not depend on the exposures! Hence, in this case, no further information on the exposures is required. If the exposures

have a multiplicative form, there is no effect of exposure misspecification on the Wald test. If the exposures do not have a multiplicative form, then the length of the Wald-test based confidence interval remains unchanged, but the location depends on the log crossproduct ratio of the cell exposures. The effect of not knowing the exposures is only to shift the location of the confidence interval. Similarly to the one-way table case, if the exposures are unknown then, for a given test size, 5% say, we can calculate a range of values for the crossproduct ratio of exposures for which H_0 will be accepted.

4.4 Conditional Test

The exact conditional distribution for testing goodness of fit of this model is given by (2) and is a multivariate noncentral hypergeometric distribution (McCullagh and Nelder, 1989, Section 7.3.4), with noncentrality parameter defined by the offset. For a 2×2 table, this is a univariate distribution which can be written as

$$\begin{aligned} f(y_{11} \,|\, Y_{1+} = y_{1+}, Y_{+1} = y_{+1}, Y_{++} = y_{++}) & \\ \propto & \frac{(\lambda_{11}\, e_{11})^{y_{11}} (\lambda_{12}\, e_{12})^{y_{12}} (\lambda_{21}\, e_{21})^{y_{21}} (\lambda_{22}\, e_{22})^{y_{22}}}{y_{11}!\, y_{12}!\, y_{21}!\, y_{22}!} \\ = & \frac{(\lambda_{11}\, e_{11})^{y_{11}} (\lambda_{12}\, e_{12})^{(y_{1+}-y_{11})} (\lambda_{21}\, e_{21})^{(y_{+1}-y_{11})} (\lambda_{22}\, e_{22})^{(y_{++}-y_{1+}-y_{+1}+y_{11})}}{y_{11}!(y_{1+}-y_{11})!(y_{+1}-y_{11})!(y_{++}-y_{1+}-y_{+1}+y_{11})!} \\ \propto & \frac{(\psi\, \phi)^{y_{11}}}{y_{11}!(y_{1+}-y_{11})!(y_{+1}-y_{11})!(y_{++}-y_{1+}-y_{+1}+y_{11})!}. \end{aligned} \tag{2}$$

with noncentrality parameter $\psi\, \phi$, where $\psi = \lambda_{11}\lambda_{22}/\lambda_{12}\lambda_{21}$ and $\phi = e_{11}e_{22}/e_{12}e_{21}$, the cross-product ratio of the rates and exposures respectively; see also Gart (1975, 1978).

For $\phi = 1$, the conditional mle of ψ is the value of ψ, say $\widehat{\psi}$, which makes the conditional expectation of y_{11} exactly equal to its observed value, i.e.,

$$y_{11} = E(Y_{11} \mid y_{1+}, y_{+1}, y_{++}, \widehat{\psi}).$$

This equation is a polynomial in ψ. Exact $1-\alpha$ conditional limits for ψ may also be obtained by using this conditional distribution, say (ψ_L, ψ_U). For

$\phi = e_{11}e_{22}/e_{12}e_{21} \neq 1$, the conditional mle is

$$\frac{\widehat{\psi}}{e_{11}e_{22}/e_{12}e_{21}}$$

and exact conditional limits for ψ are

$$(\frac{\psi_L}{e_{11}e_{22}/e_{12}e_{21}}, \frac{\psi_U}{e_{11}e_{22}/e_{12}e_{21}})$$

(Gart, 1978). Again, the conditional mle and conditional confidence limits only depend on the crossproduct ratio of the exposures. In contrast to the Wald-based confidence interval where the length did not depend on the crossproduct ratio, the conditional limits are scaled by the crossproduct ratio of the exposures.

4.5 Likelihood Ratio Test

The likelihood ratio test of goodness of fit of the multiplicative rates model is

$$L^2 = 2\sum_{ij} y_{ij} \log(y_{ij}/\widehat{m}_{ij})$$

As shown in the Appendix, $\widehat{m}_{ij}$ only depends on the crossproduct ratio of the exposures. Therefore, when the exposures have a multiplicative form, all these crossproduct ratios are one and $\widehat{m}_{ij} = y_{i+}\, y_{+j}/y_{++}$. In this case, the likelihood ratio test of the goodness of fit of the multiplicative rates model is

$$L^2 = 2\sum_{ij} y_{ij} \log(y_{ij}/(y_{i+}\, y_{+j}/y_{++})).$$

Note also that in this case, the Pearson goodness of fit test statistic as well as the Pearson and deviance residuals do not depend on the exposures.

4.6 Measurement Error

The cell exposures may be subject to measurement error, e.g., the cell exposures may be underestimated. For example, suppose we wish to compare the motor vehicle accident rates of men and women aged 65-74 and 75-84, who

had a valid driver's license during the study period 1984–1988. The data in the form of 2×2 tables are a table of the number of accidents by age and sex and the corresponding table of exposures by age and sex. If the (1,1) cell exposure is underestimated by 28%, the (1,2) cell exposure by 20%, the (2,1) cell exposure by 10% and the (2,2) cell exposure by 0%, then

$$(.72\, e_{11})(1.0\, e_{22})/(.8\, e_{12})(.9\, e_{21}) = e_{11}\, e_{22}/e_{12}\, e_{21}.$$

Here, the crossproduct ratio of exposures measured with error equals the crossproduct ratio of true exposures, so inferences are unaffected. Again, proportions are important, not absolute values.

4.7 Danish Lung Cancer Data

In the early 1970s, a study on the potential health effects of air pollution in the Danish city of Fredericia, which was dominated by a large fertilizer plant, was carried out by comparing the number of lung cancer cases during 1968-1971, by age in Fredericia with three other cities close to Fredericia and of about the same size (Andersen, 1990). These counts are presented in Table 1. These cities are denoted city 1 (Fredericia), city 2, city 3 and city 4. Andersen (1977) presents the bivariate distribution of the number of inhabitants by city and age (Table 2), while Andersen (1990) only presents the marginal number of inhabitants for each age group and for each city!

The null hypothesis of interest is the rate of getting lung cancer is the same in all four cities for each age group. If we assume that the age distribution is the same in all four cities,

$$e_{ij}/e_{+j} = e_{i+}/e_{++}$$

we have a multiplicative form for the exposures. A likelihood ratio test of the hypothesis of homogeneous age distributions for each city yields a $L^2 = 134.83$ on 15 degrees of freedom, so that the hypothesis of homogeneity is rejected. This is not too surprising given the large "sample size" of 26 408. Table 3 presents the percentage age distribution for each city. These age distributions are similar. Table 6 presents the adjusted residuals for the model of homogeneity. Adjusted residuals have asymptotically a standard normal distribution when the model is true. The largest adjusted residual

in absolute value, 8.96, corresponds to the (1,1) cell, i.e., age group 40-54 in city 1, while the fourth smallest adjusted residual in absolute value, 0.53, corresponds to the (6,4) cell, i.e., age group 75+ in city 4. The crossproduct ratios of exposures relative to the (1,1) cell and relative to the (6,4) cell are presented in Tables 4 and 5 respectably. How far these ratios are from the null value of one, which corresponds to homogeneity, is one measure of how similar these age distributions are. Using the (1,1) cell as the baseline cell is probably the worst case as this cell has the largest adjusted residual, while using the (6,4) cell as the baseline cell is almost the best case as it has the fourth smallest adjusted residual. Hence, in Table 6 these ratios look closer to 1 with 9 ratios less than 1 and 6 ratios greater than 1, rather than in Table 5 with all 15 ratios greater than 1.

Table 1: Number of lung cancer cases during 1968-71 for four Danish cities by age.

age	city 1	city 2	city 3	city 4	total
40-54	11	13	4	5	33
55-59	11	6	8	7	32
60-64	11	15	7	10	43
65-69	10	10	11	14	45
70-74	11	12	9	8	40
75+	10	2	12	7	31
total	64	58	51	51	224

Table 2: Number of inhabitants for four Danish cities by age.

age	city 1	city 2	city 3	city 4	total
40-54	3 059	2 879	3 142	2 520	11 600
55-59	800	1 083	1 050	878	3 811
60-64	710	923	895	839	3 367
65-69	581	834	702	631	2 748
70-74	509	634	535	539	2 217
75+	605	782	659	619	2 665
total	6 264	7 135	6 983	6 026	26 408

Table 3: Percentage age distributions for four Danish cities.

age	city 1	city 2	city 3	city 4	total
40-54	48.33	40.35	44.99	41.82	43.93
55-59	12.77	15.18	15.04	14.57	14.43
60-64	11.33	12.94	12.82	13.92	12.75
65-69	9.28	11.69	10.05	10.47	10.41
70-74	8.13	8.89	7.66	8.94	8.40
75+	9.66	10.96	9.44	10.27	10.09

Table 4: Adjusted residuals for model of homogeneity

age	city 1	city 2	city 3	city 4
40-54	8.96	-7.12	2.10	-3.75
55-59	-4.28	2.10	1.68	0.35
60-64	-3.85	0.55	0.20	3.11
65-69	-3.36	4.15	-1.13	0.19
70-74	-0.88	1.75	-2.58	1.75
75+	-1.30	2.85	-2.12	0.53

Table 5: Crossproduct ratios of exposures relative to the (1,1) cell

age	city 1	city 2	city 3	city 4
40-54				
55-59		1.438	1.278	1.332
60-64		1.381	1.227	1.434
65-69		1.525	1.176	1.318
70-74		1.323	1.023	1.275
75+		1.373	1.060	1.242

Table 6: Crossproduct ratios of exposures relative to the (6,4) cell

age	city 1	city 2	city 3	city 4
40-54	1.242	0.904	1.171	
55-59	0.932	0.976	1.123	
60-64	0.866	0.871	1.002	
65-69	0.942	1.046	1.045	
70-74	0.966	0.931	0.932	
75+				

Assuming the age distributions are the same for the cities, we can test the multiplicative rates (no interaction) model by using the likelihood ratio test on the lung cancer counts (ignoring the exposures). The $L^2 = 20.67$ on 15 degrees of freedom with p-value = 0.148, so the hypothesis of multiplicative rates is not rejected. Andersen (1990) notes that the tests of city effects and age effects can be carried out using the marginal distributions of the lung cancer cases by city and by age respectively. In the contingency table literature, this is called collapsability, i.e., when a test can be carried out using lower dimensional marginals. Hence, the test of the hypothesis that the rate of lung cancer is the same in the four cities can be carried out using the marginal distribution of lung cancer cases by city and the marginal distribution of inhabitants by city. The fraction of total exposure in each city specifies the multinomial probabilities used for the likelihood ratio test in the one-way table of lung cancer cases by city. Here $L^2 = 3.5$ on 3 degrees of freedom with p-value = 0.32, so the hypothesis of equal expected lung cancer rates in the four cities is not rejected at the 5% level.

Note that one, perhaps surprising, implication of this collapsability result, is that if a multiplicative exposure structure holds, then model screening for an additive log-linear rates model for an arbitrary number of factors only involved tests in one-way tables, one for each factor!

4.8 Multiplicative Exposures

Multiplicative exposure structures may be common. If suicides were classified by month of suicide and sex of suicide, then exposure is known for one margin of the table, namely, the number of days in each month and this would be same for each sex, so that a multiplicative exposure structure holds. If one factor is time (year), then the population at risk over time typically may change slowly, so that a multiplicative exposure structure holds approximately. This is exemplified by our next example of a three-way table.

5 Three-way Tables

Table 7 presents the number of fatal fire casualties in the UK during 1969-73 by sex and age. Table 8 presents the population at risk by sex and age for 1971. This information is available for census year 1971, but not for the other years. One way of dealing with the missing population data for the other years is to assume that the population at risk by age and sex for the other years is the same as the 1971 figures. Alternatively, the population at risk over time may change slowly, so that a multiplicative exposure structure holds approximately. If the total population were growing at 1% per year for each age and sex category, then a type of multiplicative exposure structure would hold over time, namely, the exposures as a function of i (time), j (age) and k (sex) would factorize as $e_{ijk} = e_i\, e_{jk}$. By thinking of the joint age-sex factor as a new factor, then we may think of the three-way table as a restructured two-way table (time factor by joint age-sex factor) and our results for two-way tables apply. For example, one could study effects of time (year) using the time marginal distribution.

Table 7: Number of fatal fire casualties in the UK during 1969-73 by sex and age.

	males				
age	1969	1970	1971	1972	1973
<15	105	96	85	104	105
15-44	126	115	110	162	161
45-64	75	92	84	107	106
65-74	48	62	58	60	76
75+	59	76	78	74	75

	females				
age	1969	1970	1971	1972	1973
<15	83	70	71	90	79
15-44	43	44	68	59	74
45-64	77	67	66	94	99
65-74	91	55	69	92	76
75+	142	145	116	191	167

Table 8: UK population in 1971 (millions) by sex and age.

age	males	females
<15	6.9	6.5
15-44	10.9	10.6
45-64	6.5	6.8
65-74	2.0	2.9
75+	0.8	1.8

References

Agresti, A. (2002). *Categorical Data Analysis, Second Edition.* Wiley.

Andersen, E. B. (1977). Multiplicative Poisson models with unequal cell rates. *Scandinavian Journal of Statistics*, **4**, 153–158.

Andersen, E. B. (1990). *The Statistical Analysis of Categorical Data.* Springer-Verlag.

Clayton, D. and Hills, M. (1993). *Statistical Models in Epidemiology.* Oxford University Press.

Gart, J. J. (1975). The Poisson distribution: the theory and application of some conditional tests. In *Statistical Distributions in Scientific Work*, Vol. 2, G. P. Patil et al. (eds), 125–140. Dordrecht, Holland: Reidel Publishing Company.

Gart, J. J. (1978). The analysis of ratios and cross-product ratios of Poisson variates with application to incidence rates. *Communications in Statistics – Theory and Methods* **7**, 917–937.

Hoem, J. M. (1995). Harmless omission in the standardization of demographic rates. *European Journal of Population*, **11**, 313–322.

McCullagh, P., and J.A. Nelder. 1989. *Generalized Linear Models. Second Edition.* Chapman and Hall.

McDonald, J. W., Smith, P. W. F. and Forster, J. J. (1999). Exact tests of goodness of fit of log-linear models for rates. *Biometrics*, **55**, 620–624.

Scheffe, H. (1959). *The Analysis of Variance.* Wiley.

Appendix

Consider the multiplicative rates model for a 2×2 table. The fitted rates are the positive solutions to the equation:

$$\frac{\widehat{\lambda}_{11}\,\widehat{\lambda}_{22}}{\widehat{\lambda}_{12}\,\widehat{\lambda}_{21}} = 1$$

or equivalently

$$\frac{\widehat{m}_{11}\,\widehat{m}_{22}}{\widehat{m}_{12}\,\widehat{m}_{21}} = \frac{e_{11}\,e_{22}}{e_{12}\,e_{21}}$$

subject to the constraints that the fitted counts add up to the observed marginal counts, i.e., $\widehat{m}_{11} + \widehat{m}_{12} = y_{1+}$, $\widehat{m}_{21} + \widehat{m}_{22} = y_{2+}$, $\widehat{m}_{11} + \widehat{m}_{21} = y_{+1}$ and $\widehat{m}_{12} + \widehat{m}_{22} = y_{+2}$. With these constraints, the equation to be solved is

$$\frac{\widehat{m}_{11}\,\widehat{m}_{22}}{\widehat{m}_{12}\,\widehat{m}_{21}} = \frac{(y_{11}+\delta)\,(y_{22}+\delta)}{(y_{12}-\delta)\,(y_{21}-\delta)} = \frac{e_{11}\,e_{22}}{e_{12}\,e_{21}}$$

where $\widehat{m}_{11} = y_{11} + \delta$, $\widehat{m}_{12} = y_{12} - \delta$, $\widehat{m}_{21} = y_{21} - \delta$ and $\widehat{m}_{22} = y_{22} + \delta$. This is a quadratic in δ and has a "closed form" solution. Note that the solution only depends on the crossproduct ratio of the exposures.

When the exposures have a multiplicative form,

$$\widehat{\lambda}_{ij} = (y_{i+}\,y_{+j}/y_{++})/(e_{i+}\,e_{+j}/e_{++})$$

satisfies

$$\frac{\widehat{\lambda}_{11}\,\widehat{\lambda}_{22}}{\widehat{\lambda}_{12}\,\widehat{\lambda}_{21}} = 1$$

as

$$\frac{[(y_{1+}\,y_{+1}/y_{++})/(e_{1+}\,e_{+1}/e_{++})][\,(y_{2+}\,y_{+2}/y_{++})/(e_{2+}\,e_{+2}/e_{++})]}{[(y_{1+}\,y_{+2}/y_{++})/(e_{1+}\,e_{+2}/e_{++})]\,[(y_{2+}\,y_{+1}/y_{++})/(e_{2+}\,e_{+1}/e_{++})]} = 1$$

and

$$\widehat{m}_{ij} = (y_{i+}\,y_{+j}/y_{++})$$

satisfies the constraint that the fitted counts add up to the observed marginal counts.

Now consider the multiplicative rates model for a $R \times C$ table, the fitted rates are the positive solutions to the $(R-1)(C-1)$ equations:

$$\frac{\widehat{\lambda}_{11}\,\widehat{\lambda}_{ij}}{\widehat{\lambda}_{1j}\,\widehat{\lambda}_{i1}} = 1 \quad \text{for} \quad i = 2, \ldots, R;\ j = 2, \ldots, C$$

or equivalently

$$\frac{\widehat{m}_{11}\,\widehat{m}_{ij}}{\widehat{m}_{1j}\,\widehat{m}_{i1}} = \frac{e_{11}\,e_{ij}}{e_{1j}\,e_{i1}} \quad \text{for} \quad i = 2, \ldots, R;\ j = 2, \ldots, C$$

subject to the constraints that the fitted counts add up to the observed marginal counts. When the exposures have a multiplicative form,

$$\widehat{\lambda}_{ij} = (y_{i+}\,y_{+j}/y_{++})/(e_{i+}\,e_{+j}/e_{++})$$

satisfies

$$\frac{\widehat{\lambda}_{11}\,\widehat{\lambda}_{ij}}{\widehat{\lambda}_{1j}\,\widehat{\lambda}_{i1}} = 1$$

as

$$\frac{[(y_{1+}\,y_{+1}/y_{++})/(e_{1+}\,e_{+1}/e_{++})][\,(y_{i+}\,y_{+j}/y_{++})/(e_{i+}\,e_{+j}/e_{++})]}{[(y_{1+}\,y_{+j}/y_{++})/(e_{1+}\,e_{+j}/e_{++})]\,[(y_{i+}\,y_{+1}/y_{++})/(e_{i+}\,e_{+1}/e_{++})]} = 1$$

and

$$\widehat{m}_{ij} = (y_{i+}\,y_{+j}/y_{++})$$

satisfies the constraint that the fitted counts add up to the observed marginal counts. Hence, closed form mles exist for the multiplicative rates model when the exposures have a multiplicative form.