Quality & Quantity **33:** 45–58, 1999.

On the Problem of Identification in Multiplicative Intensity-Rate Models with Multiple Interactions[1]

GEBRENEGUS GHILAGABER
Department of Statistics, Uppsala University PO Box 513, 751 20 Uppsala, Sweden

Abstract. In this paper we examine a multiplicative intensity model in which a covariate interacts with two other covariates in the same model. We demonstrate, analytically, that in such situations a log-linear parameterization based on two pairs of baseline levels cannot be transformed, uniquely, to the, otherwise equivalent, multiplicative parameterization. We show that the problem lies in an oversight of the *conditional independence* between the two covariates interacting with a common third covariate. As a solution, therefore, we propose an approach that takes due account of such dependence. Our proposed approach uses a common baseline level for the three covariates involved in interaction while estimating the corresponding relative intensities. The issues addressed are illustrated with a demographic data set involving the estimation of rates of transition to parenthood.

Key words: intensity rate, baseline level, multiplicative intensity model, log-linear intensity model, multiple interactions, identification, complete independence, conditional independence.

1. Introduction

In an attempt to investigate sex-differentials in the intensity of first-birth to cohabiting and married Swedish adults, Bernhardt and Bjerén (1990) used a multiplicative intensity model of the type discussed in Breslow and Day (1975) and reviewed in Hoem (1987). The model controls for five sociodemographic variables – *Sex*, *Education*, *Residence*, *Age*, and *Duration*.[3] In their final analysis, the authors found a five-factor two-interaction model that fits the data 'best'. This final model is such that *Education* and *Duration* act independently while *Sex* interacts with both *Residence* and *Age*. The relative intensities resulting from such a model and displayed in their Table 4 (page 13) are as shown in Table I.

We shall postpone details on how the values in the table are obtained to later sections. For the moment it suffices with the interpretation. According to Table I(a), the low-educated (men or women) are about twice (1.91 times) as likely to have first-birth as those with middle-level education when the other covariates are controlled for. Those with high-level education, on the other hand, have about the same intensity (1.03 times) as those with middle-level education.

Similarly, Table I(b) shows that men residing in Värmland (except Torsby) are more than 5 times as likely to have first-birth as men residing in Torsby, while men residing in other parts of Sweden (outside Värmland) have intensity that is

Table I. Relative intensities of first-birth across covariates as shown in Bernhardt and Bjerén (1990)

(a) Relative intensities for *Education*

Education		
Low	Middle	High
1.91	1	1.03

(b) Relative intensities for the interaction between *Sex* and *Residence*

	Residence		
Sex	Torsby	Värmland	Sweden
Male	1	5.11	1.62
Female	2.63	2.17	4.07

(c) Relative intensities for the interaction between *Sex* and *Age*

	Age	
Sex	20–24	25–29
Male	1	0.33
Female	0.63	1.04

only about 1.6 times as high.[4] For women residing in Torsby, the intensity is 2.63 times the intensity of those men residing in Torsby, while women residing in other parts of Sweden are about 4 times as likely to have first child as men residing in Torsby. A similar interpretation applies to the values in Table I(c). All relative risks have been tested and were reported as significant by the authors. The test used is the likelihood ratio test, which is common in the demographic tradition and, as a result, no standard errors of estimates are reported in the original tables.

In the present paper we shall demonstrate that the values presented in Table I are obtained through a misspecified approach and that they provide a less than complete picture of the structural relationship between the covariates involved in interaction. The specific goals of the paper are (1) to suggest how, in a simple manner, the model that generated Table I can be given a mathematical representation applying the log-linear parameterization; (2) to demonstrate that in situations where a covariate interacts with two others, one cannot transform the log-linear model, in a unique manner, to the multiplicative model (both described in the next section); and (3) to present the appropriate tables one should use to convey the empirical results. As we do not intend to re-analyze the original data, which we do not have at hand, we followed the format in the original work in presenting our results. Thus, we do not intend to report standard errors of estimates, as is common in the statistical papers.

The paper consists of five sections. In Section 2 we introduce the multiplicative and log-linear parameterizations of the intensity model. In Section 3 we give a mathematical representation of the model used to obtain the results in Table I. Further, we demonstrate, analytically, that when a covariate interacts with two others, the use of a log-linear model with two pairs of baseline levels leads to estimates of intensity rates that are different from those obtained by a multiplicative formulation.[5] In Section 4 we propose an approach based on a single set of baseline level involving the three factors in the interaction, and re-estimate the relative intensities of interest. Section 5 summarizes the paper. To facilitate discussion, we have preferred to present our empirical results within the main body of the paper.

2. Two Parameterizations for the Intensity-Rate

In this section we shall define the intensity rate and describe its multiplicative and log-linear parameterizations. We have discussed such rates and their corresponding models in greater detail elsewhere (Ghilagaber, 1995, 1998). For completeness purposes, we shall describe them here again. To simplify matters we shall begin with a model with two categorical covariates (including the time variable).

Denote by λ_{ij} the rate at which an individual with the j-th level of a categorical covariate experiences an event in duration interval i. A multiplicative structure for such a rate arises when λ_{ij} is obtained from multiplicative contributions of the i-th duration group, say θ_i, and the j-th level of the categorical covariate, say α_j:

$$\lambda_{ij} = \theta_i \alpha_j, \quad 1, \ldots, I, \quad j = 1, \ldots, J, \tag{1}$$

with one of the α_j's, say α_{j_0} fixed to be equal to 1. Equation (1) will, henceforth, be referred to as a multiplicative parameterization of the intensity rate. θ_i is the baseline intensity (value of the intensity rate when $\alpha_j = 1$), while α_j represents the intensity of an individual with level j of the covariate relative to that of an individual with the baseline level j_0.

Let us now define $A_i = \ln \theta_i$, and $B_j = \ln \alpha_j$ so that $\ln \lambda_{ij} = \ln \theta_i + \ln \alpha_j = A_i + B_j$. If we further let

$$\bar{A} = \frac{\sum_{i=1}^{I} A_i}{I}, \ \bar{B} = \frac{\sum_{j=1}^{J} B_j}{J}, \ \Delta = \bar{A} + \bar{B}, \ a_i = A_i - \bar{A}, \ b_j = B_j - \bar{B},$$

we may define a log-linear equivalent of the intensity rate in (1) as

$$\begin{aligned} \ln \lambda_{ij} &= \ln \theta_i + \ln \alpha_j = A_i + B_j = (a_i + \bar{A}) + (b_j + \bar{B}) \\ &= (\bar{A} + \bar{B}) + a_i + b_j = \Delta + a_i + b_j. \end{aligned} \tag{2}$$

Olivier and Neff's (1976) program – LOGLIN, which is commonly used to estimate parameters of log-linear models for contingency tables, yields values of Δ, a_i, and b_j such that

$$\sum_{i=1}^{I} a_i = \sum_{j=1}^{J} b_j = 0.$$

One can, therefore, make use of these values to estimate the intensity rates as

$$\lambda_{ij} = \exp(\Delta + a_i + b_j), \quad i = 1, \ldots, I; \; j = 1, \ldots, J. \tag{3}$$

Equation (3) is a log-linear parameterization of the intensity rate. In other words, (1) and (3) represent two different parameterizations of the same intensity rate λ_{ij}.

Without loss of generality, we may select j_0 to be the first level ($j_0 = 1$). Then, since by design $\alpha_1 = 1$, we have

$$\lambda_{i1} = \theta_i \alpha_1 = \theta_i(1) = \theta_i$$

or, equivalently

$$\theta_i = \theta_i \alpha_1 = \lambda_{i1} = \exp(\Delta + a_i + b_1) \tag{4}$$

and

$$\alpha_j = \lambda_{ij}/\theta_i = \exp(\Delta + a_i + b_j)/\exp(\Delta + a_i + b_1) = \exp(b_j - b_1) \tag{5}$$

as estimates of the baseline and relative intensities, respectively.

More details on procedures for estimating the parameters in (1) and/or (3) may be found in Breslow and Day (1975), Laird and Olivier (1981), Hoem (1987), and Ghilagaber (1995, 1998).

The above models and their corresponding methods of estimation can, easily, be extended to the case of more than two covariates and can entertain interactions between covariates. With three factors indexed by i, j, and k, in which the last two factors interact, (1) and (3) may be extended to

$$\lambda_{ijk} = \theta_i \gamma_{jk}, \quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K \tag{6}$$

and

$$\lambda_{ijk} = \exp(\Delta + a_i + b_j + c_k + d_{jk}),$$

$$i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, K \tag{7}$$

respectively, such that

$$\sum_{j=1}^{I} a_i = \sum_{j=1}^{J} b_j = \sum_{k=1}^{K} c_k = \sum_{j=1}^{J} d_{jk} = \sum_{k=1}^{K} d_{jk} = 0.$$

Below, we extend the above three-factor one-interaction model to a five-factor model where a covariate interacts with two others. We shall demonstrate that in such situations it is not possible to, uniquely, transform the log-linear parameterization of the intensity rate (3) into the multiplicative parameterization (1).

3. Multiple Interactions and the Identification Problem

3.1. MATHEMATICAL REPRESENTATION

In accordance with our numerical example, let us consider a set of data involving a duration variable indexed by i and four other covariates indexed respectively, by j, k, r, and s. Assume further, that the factor indexed by k interacts with both the covariates indexed by r and s, while those indexed by i and j are not involved in any interaction. A mathematical representation for the multiplicative model used (but not explicitly mentioned) in generating Table I is given as

$$\lambda_{ijkrs} = \theta_i \beta_j \rho_{kr} \psi_{ks} \tag{8}$$

while the corresponding log-linear model is given by

$$\ln \lambda_{ijkrs} = \Delta + a_i + b_j + c_k + d_r + e_s + f_{kr} + g_{ks} \tag{9}$$

or, equivalently,

$$\lambda_{ijkrs} = \exp\{\Delta + a_i + b_j + c_k + d_r + e_s + f_{kr} + g_{ks}\}. \tag{10}$$

Estimates of the terms on the right hand side of (9), for the data in Bernhardt and Bjerén (1990) are shown in Table II.

The relative intensities as shown in Table I, are calculated by the usual approach. For the factor indexed by j, for instance, we have

$$\beta_j = \frac{\lambda_{ijkrs}}{\lambda_{ij_o krs}} = \exp(b_j - b_{j_0}). \tag{11}$$

The corresponding estimates of β_j (with $j_0 = 2$), are shown in Table I(a).

For the interacting factors, Bernhardt and Bjerén (1990) implicitly use the following expressions for the relative intensities:

$$\rho_{kr} = \exp\{(c_k - c_{k_0}) + (d_r - d_{r_0}) + (f_{kr} - f_{k_0 r_0})\}, \tag{12}$$

Table II. Estimates of the parameters in the log-linear model for the data set in Bernhardt and Bjerén (1990)

(a) Main effects

Factor (covariate)	Level	Symbol	Estimate
Grand mean effect		Δ	-3.716
Duration	0–6	a_1	0.557
	7–18	a_2	-0.243
	19–36	a_3	0.073
	37–60	a_4	-0.206
	61–120	a_5	-0.180
Education	Low	b_1	0.422
	Middle	b_2	-0.226
	High	b_3	-0.197
Sex	Male	c_1	-0.172
	Female	c_2	0.172
Residence	Torsby	d_1	-0.393
	Värmland	d_2	0.326
	Sweden	d_3	0.067
Age	20–24	e_1	0.152
	25–29	e_2	-0.152

(b) Interaction effects between Sex and Residence

	Residence		
Sex	Torsby	Värmland	Sweden
Male	$f_{11} = -0.312$	$f_{12} = 0.600$	$f_{13} = -0.288$
Female	$f_{21} = 0.312$	$f_{22} = -0.600$	$f_{23} = 0.288$

(c) Interaction effects between Sex and Age at start of union

	Age	
Sex	20–24	25–29
Male	$g_{11} = 0.406$	$g_{12} = -0.406$
Female	$g_{21} = -0.406$	$g_{22} = 0.406$

$$\psi_{ks} = \exp\{(c_k - c_{k_0}) + (e_s - e_{s_0}) + (g_{ks} - g_{k_0 s_0})\}. \tag{13}$$

The corresponding estimates of ρ_{kr} and ψ_{ks} (with $k_0 = r_0 = s_0 = 1$) are shown in Tables I(b) and I(c), respectively. Thus, for instance, the value 5.11 in Table I(b) is obtained as $5.11 = \exp\{(-0.172 + 0.326 + 0.600) - (-0.172 - 0.393 - 0.312)\}$, while the value 2.63 in the same table is obtained as $2.63 = \exp\{(0.172 - 0.393 + 0.312) - (-0.172 - 0.393 - 0.312)\}$. The corresponding entries in Table I(c) are

$0.33 = \exp\{(-0.172 - 0.152 - 0.406) - (-0.172 + 0.152 + 0.406)\}$, and $0.63 = \exp\{(0.152 + 0.172 - 0.406) - (0.152 - 0.172 + 0.406)\}$.

3.2. THE IDENTIFICATION PROBLEM

It is now time to have a closer look at the two parameterizations for the intensity rate. Recall that the baseline intensity is given by

$$\theta_i = \lambda_{ij_0k_0r_0s_0} = \exp\{\Delta + a_i + b_{j_0} + c_{k_0} + d_{r_0} + e_{s_0} + f_{k_0r_0} + g_{k_0s_0}\}. \tag{14}$$

Thus, if we substitute the expressions in (11)–(14) for the corresponding parameters on the right hand of (8), we have

$$\begin{aligned}\lambda_{ijkrs} &= \theta_i\beta_j\rho_{kr}\psi_{ks} \\ &= \exp\{\Delta + a_i + b_{j_0} + c_{k_0} + d_{r_0} + e_{s_0} + f_{k_0r_0} + g_{k_0s_0}\} \times \exp(b_j - b_{j_0}) \times \\ &\quad \times \exp\{(c_k - c_{k_0}) + (d_r - d_{r_0}) + (f_{kr} - f_{k_0r_0})\} \times \\ &\quad \times \exp\{(c_k - c_{k_0}) + (e_s - e_{s_0}) + (g_{ks} - g_{k_0s_0})\},\end{aligned} \tag{15}$$

which, after simplification, reduces to

$$\lambda_{ijkrs} = \exp\{\Delta + a_i + b_j + c_k + d_r + e_s + f_{kr} + g_{kr} + g_{ks} + (c_k - c_{k_0})\}. \tag{16}$$

The expression in (16) is different from that in (10) because (16) contains the additional term $(c_k - c_{k_0})$ in the exponent.[6]

If we redefine

$$\rho^*_{kr} = \exp\{\phi(c_k - c_{k_0}) + (d_r - d_{r_0}) + (f_{kr} - f_{k_0r_0})\} \tag{17}$$

and

$$\psi^*_{ks} = \exp\{(1 - \phi)(c_k - c_{k_0}) + (e_s - e_{s_0}) + (g_{ks} - g_{k_0s_0})\} \tag{18}$$

for some unspecified ϕ, then we get

$$\begin{aligned}\lambda^*_{ijkrs} &= \theta_i\beta_j\rho^*_{kr}\psi^*_{ks} = \exp\{\Delta + a_i + b_{j_0} + c_{k_0} + d_{r_0} + e_{s_0} + f_{k_0r_0} + g_{k_0s_0}\} \times \\ &\quad \times \exp(b_j - b_{j_0}) \exp\{\phi(c_k - c_{k_0}) + (d_r - d_{r_0}) + (f_{kr} - f_{k_0f_0})\} \times \\ &\quad \times \exp\{(1 - \phi)(c_k - c_{k_0}) + (e_s - e_{s_0}) + (g_{ks} - g_{k_0s_0})\} \\ &= \exp\{\Delta + a_i + b_j + c_k + d_r + e_s + f_{kr} + g_{ks}\},\end{aligned} \tag{19}$$

which is identical to λ_{ijkrs} in (10). Once $\phi \neq 1$ is chosen, however, it is impossible to partition the contribution of $(c_k - c_{k_0})$, in a unique way, among the two exponents in (17) and (18). In other words, the relative intensities ρ^*_{kr} and ψ^*_{ks} become unidentified in the sense that they cannot be represented in a unique way. As a result, it is

not possible to, uniquely, transform the log-linear parameterization of the intensity rate λ_{ijkrs} to the relative intensities (multiplicative) format. Thus, unless proper care is taken to use the correct formulae in estimating the relative intensities, the two parameterizations can lead to different estimates of the intensity rate.

4. Proposed Solution

In our attempt to look for a solution, we have made use of graphical models but we have chosen not to present the details here.[7] From such models it was obvious that the main drawback in the approach used by Bernhardt and Bjerén (1990) lies in failing to recognize the fact that the two covariates interacting with the same covariate are only conditionally (and not completely) independent. Such oversight has led to ignoring the factor *Age* while computing the effects of the interaction between *Sex* and *Residence*, and to ignoring the factor *Residence* in the computation of interaction effects between *Sex* and *Age*. Such error was done because of the tacit but unwarranted assumption that *Residence* and *Age* are completely independent.

Having exposed methodological drawbacks of previous works, and having explained why this is so, we shall now proceed to the last and important purpose of the paper – presenting the appropriate tables one should use to convey the empirical results concerning sex-differentials in first-birth intensities. Our suggested solution is to use a common baseline level for the three factors involved in interaction and to re-compute the relative intensities of interest. In this way, we take due account of the conditional independence between the two covariates interacting with a common third covariate.

Denote by μ_{krs} the relative intensity at the k-th level of the covariate *Sex*, r-th level of the covariate *Residence*, and s-th level of the covariate *Age*. Further, let θ_i and β_j represent, as before, the baseline intensity and the relative intensity of *Educational level* j, respectively. Our multiplicative intensity model will then be given by

$$\lambda_{ijkrs} = \theta_i \beta_j \mu_{krs} \tag{20}$$

and the relative intensity of interest, μ_{krs}, is given by

$$\begin{aligned} \mu_{krs} &= \frac{\lambda_{ijkrs}}{\lambda_{ijk_0r_0s_0}} \\ &= \exp\{(c_k - c_{k_0}) + (d_r - d_{r_0}) + (e_s - e_{s_0}) + (f_{kr} - f_{k_0r_0}) + \\ &\quad + (g_{ks} - g_{k_0s_0})\}. \end{aligned} \tag{21}$$

If we were to compute the joint effect of the three covariates involved in interaction using the mathematical representation corresponding to Table I (with two pairs of baseline levels), such effect would have been obtained as a product of the two factors (relative intensities) in (15):

Table IIIa. Relative intensities μ_{krs} with a common baseline level for *Sex*, *Residence*, and *Age*

		Residence		
Sex	Age	Torsby	Värmland	Sweden
Male	20–24	1	5.11	1.62
	25–29	0.33	1.67	0.53
Female	20–24	1.17	0.97	1.81
	25–29	1.94	1.60	3.00

Table IIIb. Sex profiles of relative intensities across Residence

	Residence		
Sex	Torsby	Värmland	Sweden
Male	1	5.11	1.62
Female	1	0.83	1.55

Table IIIc. Sex profiles of relative intensities across Age

	Age	
Sex	20–24	25–29
Male	1	0.33
Female	1	1.66

Table IIId. First-birth intensities of women relative to that of men across Residence and Age.

	Residence		
Age	Torsby	Värmland	Sweden
20–24	1.17	0.19	1.12
25–29	5.89	0.96	5.66

$$\begin{aligned}\rho_{kr}\psi_{ks} &= \frac{\lambda_{ijkrs}}{\lambda_{ijk_0r_0s_0}} \\ &= \exp\{2(c_k - c_{k_0}) + (d_r - d_{r_0}) + (e_s - e_{s_0}) + \\ &\quad +(f_{kr} - f_{k_0r_0} + (g_{ks} - g_{k_0s_0})\} \\ &= \mu_{krs}\exp(c_k - c_{k_0})\end{aligned}$$

so that,

$$\mu_{krs} = \rho_{kr}\psi_{ks}\exp\{-(c_k - c_{k_0})\}. \tag{22}$$

If we solve for the absolute intensity, we get

$$\lambda^{**}_{ijkrs} = \lambda_{ijk_0r_0s_0}\rho_{kr}\psi_{ks} = \theta_i\beta_j\mu_{krs}\exp(c_k - c_{k_0}),$$

which is different from our formulation in (20). In other words, the use of two pairs of baseline levels, instead of one common for the three factors involved in the interaction, will inflate the real intensity rate by a factor of $\exp(c_k - c_{k_0})$. In the empirical example this quantity is equal to $\exp\{0.172 - (-0.172)\} = 1.41$, which implies that estimated intensities for females are inflated by a factor of 1.41.

Let us now proceed to the re-estimation of the relative intensities of interest. The relative risks for the noninteracting factor *Education* (Table I(a)) and those of the baseline intensities remain unchanged.

For the factors involved in interaction, we get the $(2 \times 2 \times 3)$ table of relative intensities, μ_{krs}, obtained by straight application of Equation (21) and shown in Table III(a).

Further, we can compute the following relative intensities of interest:

1. Sex-profiles of intensities across Residence (Table III(b)):

$$\begin{aligned}\frac{\lambda_{ijkrs}}{\lambda_{ijkr_0s}} &= \exp\{(d_r - d_{r_0}) + (f_{kr} - f_{kr_0})\} \\ &= \text{Effect of Residence on men's first-birth intensity (when } k = 1) \\ &= \text{Effect of Residence on women's first-birth intensity} \\ &\quad \text{(when } k = 2).\end{aligned} \tag{23}$$

Thus, men living in Värmland (except Torsby) are more than 5 times as likely to get their first birth relative to those still living in Torsby, while those living in other parts of Sweden have rates that are only 1.62 times that of men living in Torsby. For women, the corresponding relative intensities are 0.83 and 1.55, respectively. These intensities control for Age, Education and Duration. It is worth reminding the reader that these values can be obtained from Table I(b), by dividing the entries in the second and third columns by those in the first.

2. Sex profiles of intensities across Age (Table III(c)):

$$\begin{aligned}\frac{\lambda_{ijkrs}}{\lambda_{ijkrs_0}} &= \exp\{(e_s - e_{s_0}) + (g_{ks} - g_{ks_0})\} \\ &= \text{Effect of Age on the men's first-birth intensity (when } k = 1) \\ &= \text{Effect of Age on women's first-birth intensity (when } k = 2). \quad (24)\end{aligned}$$

In other words, men forming a union at older ages (25–29 years) have first-birth intensity that is only 0.33 times that of men forming a union at younger ages, net of the effects of Education and Residence and at all union durations. For women, on the other hand, the corresponding relative intensity is 1.66 – forming a union at higher ages increases the intensity by 66% relative to that of younger ages. Again, these values can also be obtained from Table I(c), by dividing the entries in the second column by those in the first.

3. Sex-differentials of intensities across Residence and Age (Table III(d)):

$$\frac{\lambda_{ijkrs}}{\lambda_{ijk_0rs}} = \exp\{(c_k - c_{k_0}) = (f_{kr} - f_{k_0r}) + (g_{ks} - g_{k_0s})\}. \quad (25)$$

The intensity of first birth for women is higher or lower than that of men depending on which age-group and/or residence one refers to. For instance, the relative intensity is about six times for women who initiate their union at older ages and live either in Torsby or outside the rest of Värmland. For those living in Värmland, the intensity for the two sex's is almost the same for those initiating the union at older ages. Women living in Värmland that have initiated their union early have only about one-fifth as high a rate as men in the same (age, residence) group. Women initiating their union early (20–24 years) and living in either Torsby or other parts of Sweden (outside Värmland) have a rate that is over 10% higher than that of their male counterparts.

The values in Table III(d) relate to all levels of *Education* and *Duration*. In contrast to those in Tables III(b) and III(c), the values in Table III(d) cannot be obtained from Table I. Here is one of the main drawbacks of the approach that generated Table I. Thus, while the main goal of the investigation that resulted in Table I was to study sex-differentials across covariates, the procedure used yields a less than complete picture of the initial goal. Note also that in the earlier work the results concerning interaction effects are presented in two separate tables with (6 + 4 = 10) cells, while our results are presented in a single table with ($2 \times 2 \times 3 = 12$) cells. One consequence is that it is impossible to compare the two tables entirely.

5. Concluding Remarks

The issue of whether survival models and methods appropriate for one setting can also be used for the analysis of data in another setting has been explored earlier.

One conclusion from previous works is that computer algorithms developed for one setting can often be exploited in another. In particular, LOGLIN, a program developed for log-linear modeling of data in multidimensional contingency tables, has been used to estimate the parameters of a multiplicative intensity model in problems involving grouped survival data.

In the present paper we have explored, in greater detail, the close relationship between the parameterizations in multiplicative and log-linear models. Our exploration indicates that there are situations where this close relationship might lead to erroneous results if proper care is not taken in transforming estimates from one parameterization to the other.

One such situation is when a multiplicative model involves a factor that interacts with two others in the same model. We have demonstrated that, in such situations, the traditional approach of using a model with two baseline levels suffers from drawbacks which we are tempted to call *semi-Simpson's paradox* – a phenomenon that occurs when one incorrectly marginalizes (collapses) over a conditioning covariate.[8] As a consequence, it is impossible to transform the intensity rate from log-linear parameterization into the simpler relative intensities format. This happens because the relative intensities related to the interacting factors are unidentified in the sense that they cannot be expressed in a unique way.

In our attempt to look for a solution, we have made use of graphical models. Such modeling has shown that the covariate interacting with the two others *separates* these two factors in the sense that the latter two are not completely independent but only conditionally independent given the separating covariate. By doing so we have pointed out that the drawback in the earlier approach lies in the unwarranted assumption of complete independence between the two factors, each interacting with another third factor.

Taking into account the conditional independence implied, in our case, that we used a common baseline level for the three factors involved in interaction and recomputed the relative intensities of interest.

We have used a demographic data set to illustrate the issues addressed. Our empirical example shows that failure to take the conditional independence into account might lead to estimates of relative intensities that are inflated by a factor that is a function of the effects of the covariate interacting with the two other covariates. Moreover, the results are difficult to interpret.

A general lesson to learn from our analyses here is that is that it is dangerous to analyze a model with multiple interactions solely by inspecting its separate pairs of two-way (first-order) interactions.

Notes

1. Research Report 1998-1, Department of Statistics, Uppsala University, Uppsala, Sweden.
2. The idea of the work presented in this paper was suggested by Professor Jan M. Hoem at Stockholm University, and I have benefited much from his guidance and initial formulation. Professors Karl-Gustav Jöreskog and Reinhold Bergström; and Assoc. Professors Eva Bernhardt

and Johan Bring have read and contributed towards improvement of the paper. Earlier versions have been presented at the summer school on "Recent Trends in Data Analysis", Ultuna, 3–6 June 1997; and at one of the regular seminars at the Department of Statistics, Uppsala University. Participants at these seminars, and in particular Professors Harry Khamis and Rolf Sundberg made some valuable comments. I am grateful to all.

3. The *Sex* variable represents the gender of the respondent, and has two levels – Male and Female. *Education* is measured by the years of education after primary school. It has three levels – Low (0.5–2 years), Middle (2.5–5 years), and High (5.5–8 years). *Residence* is a time-varying covariate and has three levels – Torsby, Värmland, and Sweden. These levels are described further in footnote 4. *Age* refers to the respondent's age (in years) at the time of forming a union (marriage or cohabitation). It has two levels – 20–24 and 25–29. *Duration* represents the length (in months) of the union. It is the time variable and has five levels – 0–6 months, 7–18 months, 19–36 months, 37–60 months, and 61–120 months.
4. The initial study has as its starting point a cohort of men and women, born around 1945 in the parish of Fryksände. This parish comprises all of Torsby, the municipal centre of a large forest-land municipality in northern Värmland, a province in western Sweden. At the time of the study, some members of this cohort were still living in *Torsby*. Others have moved, before the time of the study, to *other parts of Värmland*; while still others have moved to *other parts of Sweden, outside Värmland*. The *Residence* variable with these regions as levels was thus chosen to capture the effects of (internal) migration on the rate of transition to parenthood.
5. Two-pairs of baseline levels arise in situations where one has two first-order interactions (interaction involving only two factors). One three-way baseline level arises when the model contains only one second-order interaction term (interaction involving three factors)
6. The exponents in (12) and (13) have the term $(c_k - c_{k_0})$ in common. Thus, one can partition this contribution of $(c_k - c_{k_0})$ among the two expressions in an infinite number of ways. In other words, the partitioning is not unique, or that the partitioning of $(c_k - c_{k_0})$ is not identified. The natural formulation of f_{kr} and g_{ks} can, therefore, *not* be used to estimate ρ_{kr} and ψ_{ks}. This is reflected in the transformation from the log-linear to the multiplicative parametrization. The problem arises because the factor indexed by k interacts with the two other factors indexed by r and s, respectively, in the same model. When Bernhardt and Bjerén (1990) used Equation (12) to compute the corresponding relative intensities, they have ignored the fact that the factor indexed by k is also interacting with another factor as shown in (13). A similar error was committed while computing the relative intensities in (13).
7. Graphical models are those models for multivariate random observations whose independence structure is characterized by a graph. Modeling data through graphs helps in understanding and interpreting the inter-relationship between various variables. Details of theoretical and applied works in graphical models may be found in Darroch, Lauritzen, and Speed, 1980; Lauritzen, 1989, 1996; Whittaker, 1990; McKee and Khamis, 1996; and Khamis, 1996.
8. To avoid cells with zero entries analysts usually restrict attention to two-way interaction models by collapsing (marginalizing) over a third factor. This issue is of some importance to practical data analysts because in many studies it is impossible to measure all potential covariates. However, the procedure of collapsing over a covariate might lead to considerable scope for paradox and error, especially if one marginalizes over potential conditioning variables. The dangers of amalgamating tables are well known and are documented in Simpson (1951) and later works that followed him.

References

Bernhardt, E.M. & Bjeren, G. (1990). Quantitative life history analysis with small data sets: A study of gender differences in the transition to parenthood in Sweden. Paper presented at the World Congress of Sociology, Madrid–Spain, July 9–13, 1990.

Breslow, N.E. & Day, N.E. (1975). Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronic Diseases* 28: 289–303.

Darroch, J.N., Lauritzen, S.L., & Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics* 8: 522–539.

Ghilagaber, Gebrenegus (1995). Similarities among some hazard-rate and duration models for grouped and continuous life-time data with covariates: an exploration and application to modeling correlates of marital dissolution. Research Report 95-3, Department of Statistics, Uppsala University.

Ghilagaber, Gebrenegus (1998). Analysis of survival data with multiple causes of failure: a comparison of hazard- and logistic-regression models with application in demography. *Quality & Quantity* 32(3): 297–324.

Hoem, J.M. (1987). Statistical analysis of a multiplicative model and its application to the standardization of vital rates: A review. *International Statistical Review* 55: 119–152.

Khamis, H.J. (1996). Application of multigraph representations of hierarchical log-linear models. In A. Von Eye and C. C. Clogg (eds), *Categorical Variables in Developmental Research: Methods of Analysis*. San Diego: Academic Press, pp. 215–229.

Laird, N., & Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association* 76: 231–240.

Lauritzen, S.L. (1989). Mixed graphical association models (with discussion). *Scandinavian Journal of Statistics* 16: 273–306.

Lauritzen, S.L. (1996). *Graphical Models*. Oxford Statistical Science Series # 17. Oxford: Clarendon Press.

McKee, T.A. & Khamis, H.J. (1996). Multigraph representations of hierarchical loglinear models. *Journal of Statistical Planning and Inference* 53: 63–74.

Olivier, D. & Neff, R. (1976). *LOGLIN 1.0: User's Guide*. Harvard: Harvard School of Public Health.

Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society – Series B* 13: 238–241.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.