Yearbook of Population Research in Finland 33 (1996), pp. 34–43

The Harmfulness or Harmlessness of Using an Anticipatory Regressor: How Dangerous Is It to Use Education Achieved as of 1990 in the Analysis of Divorce Risks in Earlier Years?

JAN M. HOEM

Professor
Demography Unit
Stockholm University
Stockholm, Sweden

Abstract

The case studies presented in this paper show that the chance of making a sensible analysis of the effect of education on divorce risks may be ruined for women who marry as teenagers if the educational variable is measured only at the end of the study period. By contrast, these adverse effects seem to be unimportant once the age at marriage is 20 or more.

Keywords: divorce, age, women, analysis, Sweden

Introduction

The nature of demographic analysis

Dynamic human behavior is often represented as streams of events occurring to individuals, and the corresponding data are analyzed by means of modern event-history analysis. In demographic investigations of family dynamics, for instance, the behavior studied may be marriage formation and dissolution, in which case the streams of events are marriages and divorces occurring to individuals in a study population. Demographic analysis of the determinants of the behavior in question typically consists of finding effects of factors that influence the rates of transition corresponding to the various events. The streams of events constitute the behavior and the rates of transition reflect the strength of the streams. In a manner of speaking, the rates are the "dependent" variables in the analysis, and the factors that influence them are the "independent" variables. The value of a determinant may be set from the start of the life segment analyzed, as is the case for the age at marriage and the number and ages of any premarital children in analyses of divorce among women in their first marriage. Alternatively, a determinant may be a dynamic covariate whose value is updated con-

tinuously during the life segment, such as current (childbearing) parity or the presence and age of a youngest child in any month of marriage.

Anticipatory covariates

It sometimes happens that the data available for demographic analysis contain the information required about the flows of demographic events but the value of some crucial dynamic determinant is known only at the end of the study period. A typical example is educational attainment in retrospective interviews in family and fertility surveys. Detailed life histories may be collected concerning births and union formation and dissolution and so on, but information about the respondent's education is sometimes obtained only for the interview date.

It is well known that attempts at studying the impact of educational attainment on demographic (or other) behavior from data of this nature may lead the analyst astray, for causal relations may be seriously disturbed by the lack of a firm time reference for the dynamics of the educational variable. If, on the basis of such data, one finds that more highly educated women have lower fertility rates in a certain age range, for instance, one does not know whether this reflects (i) lower fertility subsequent to the attainment of higher education, or (ii) an inability of women to continue their studies after early childbearing. In the first case, the causal direction would be from education to fertility, in that higher educational attainment would reduce fertility. If the second explanation were true, the causal direction would be the reverse, for childbearing would hamper further education. (Both of these possible causal relations, which are often regarded as obviously true by conventional wisdom, have been challenged in the literature, and they are used here only for illustrating our methodological point. For the sake of argument we are also ignoring any effects of selectivity; for instance, early and relatively extensive childbearing may reflect a low educational potential.) Presumably, individual demographic behavior at any stage in life is influenced by the education attained *at that time* (and by the nature of educational and other activities at the time) and not by any subsequent educational attainment. Educational status at the end of observation is an anticipatory variable that reflects the prior dynamics only in part, and potentially it may lead to a confusion of causal directions. The use of such a covariate may also produce biases in estimated effects. (For an account of the type of mathematics involved, see Hoem and Funck Jensen (1982, Section 5.3).)

Our present concern

The question we want to raise in this paper is whether the specter of causal confusion and estimation biases should scare us away from using anticipatory representations of determinants that are essentially dynamic "independent" covariates, or whether there are circumstances in which control over any adverse effects is sufficiently strong to make them harmless for the analytic purpose in hand. We provide a partial answer, as follows: In some investigations, an anticipatory educational covariate ruins analysis. In a particular investigation where we can compare anticipatory and dynamic analyses, the outcomes are so close that we can safely ignore the errors produced by the use of the anticipatory version of the educational variable. We feel confident that this must also hold for the analysis of the rest of the data set. Presumably, the lessons from our case studies can be further extended to similar situations in other data sets.

We encountered this problem in connection with a recent analysis of divorce risks among Swedish women in the 1970s and 1980s (Hoem 1995a,b). The analysis used data on streams of demographic events organized in Statistics Sweden's Fertility Reg-

ister and corresponding educational data from their Register of Educational Attainment. The Fertility Register is based on the Swedish population-register system and contains demographic life histories for all relevant birth cohorts of Swedish women; the data can be regarded as complete for our purposes. The Register of Educational Attainment contains detailed data about their education, but for large parts of the population we can only get educational attainment as of the census date in 1990. For younger cohorts, the register can also tell us when individuals completed the *gymnasium* and when they attained any subsequent highest education. The present paper is an account of how we have utilized this additional information to assess the harmfulness or of using no more than the educational level attained as of 1990 in our analyses of Swedish divorce risks. Our conclusion is that one should *not* use educational attainment by the end of the study period to estimate educational effects on divorce risks for women who married as teenagers. For women who married at higher ages, on the other hand, we can safely use our anticipatory educational variable in such an investigation. We explain how we reached this conclusion in what follows.

Data

To facilitate demographic analyses, Statistics Sweden maintains a special data bank called the Fertility Register. It contains all registered demographic events occurring in Sweden for all women born in Sweden in 1925 or later who were Swedish citizens and were domiciled in Sweden according to the census of 1960, as well as for all of their female descendants present in their households at that time and all women born in Sweden in 1961 or later. For these groups, their own dates of birth as well as all dates on which they bore children are registered, as are dates of marriage, of divorce, of out-migration from Sweden and of death. (For events other than births, dates are recorded for events that occurred after the 1960 census only. Immigrants are not included in the register. See Johansson and Finnäs (1983) and Qvist (1990) for more information.) The data are of varying quality for the first few years of the register's life and of excellent quality thereafter. According to Jan Qvist and Åke Nilsson (personal communications), we may trust all dates of birth and death in the register, marital status in 1960, all dates of marriage since 1960, most dates of divorce since the mid-1960s, and all dates of divorce since 1968. (Before 1968, the four last individual digits were missing in the identification number recorded when a divorce was granted. In the register, divorce events had to be assigned to individuals on the basis of their date of birth and other information contained in the records. The proportion that could be assigned to individuals in this manner, rose from 88 percent of all divorces granted in 1964 to 95 percent in 1967. Since 1968, there has been full coverage of the divorces granted.) We have records up to the end of 1991 in our data.

From a separate Register of Educational Attainment, Statistics Sweden has added information about each individual's education as registered at the end of 1990. For each individual included, we know the highest level of education ever recorded at that time as well as what kind of education it was (by means of a five-digit educational code (Statistics Sweden 1988)) and the year in which it was achieved. For women with a post-*gymnasium* education, we also know the year in which she completed the *gymnasium* and what kind of *gymnasium* education she received. (*Gymnasium* is the name of a school form normally completed at the age of 19, plus or minus one year. It covers all forms of upper-secondary education, both academic instruction leading to university-level studies and vocational educations leading to non-academic trades and skilled-manual occupations. A brief description of the Swedish educational system is given

in Appendix 1.) The educational attainment included in this register covers higher education completed since 1970 and all types of education completed since about 1977. This means that in our data set, we know the women's educational histories (as far as the completion of a *gymnasium* and of a post-*gymnasium* education is concerned) for the cohorts born since 1959, but we can only trust the information about the final educational level (and not her educational history otherwise) for women in our previous cohorts. For most women who died or migrated out of Sweden before 1990, we have no recorded educational level. For this reason, we have eliminated the records of all women who died or out-migrated before our period of observation was terminated at the end of 1991. This leaves us with records selected in a manner similar to that of a survey with interviews, except that we essentially have no nonresponse (beyond some limited further loss of records for which some essential educational information was missing). We have also eliminated a small number of records whose quality we are suspicious about because a divorce has been recorded before any marriage.

To illuminate the issues of anticipatory analysis that we are concerned with in this paper, we have carried out two case studies, namely one for each of the cohorts born in 1959 and 1964. Our documentation is for the cohort of 1959 and is given in Section 4 after some theoretical considerations given in Section 3. The results for the cohort of 1964 are very similar and are not documented here. To study an additional concern with the possibility of reverse causation due to educational achievement after divorce, we give the cohorts born in 1962–1964 some further consideration in Section 5.

The cohort born in 1959 had almost 45,000 female members in our data. The investigation of which we present one aspect here concerns divorces to women in their first marriage, so we have also eliminated all women who never married before the end of our study period. In our study cohort from 1959, about 20,000 women were in this category, and almost 25,000 women were included in our analysis.

After some preliminary experiments, we have grouped women by educational level into three categories for the analysis reported here, namely into those who have attained (i) the pre-*gymnasium* level, (ii) the *gymnasium*, and (iii) some post-*gymnasium* level. Our reasoning will be based on this trichotomy.

What bias is produced by using an anticipatory instead of a dynamic definition of the educational level?

Presumably a married woman's divorce risk at any stage in life is influenced by the educational level she has reached at the time and not by any subsequent educational attainment. As we mentioned in our introductory remarks already, there is therefore some danger of estimation biases if we use the (fixed) anticipatory educational variable, which is all we have at our disposal for most of our respondents, instead of basing our analysis on a corresponding time-varying covariate. We will now argue that this danger is small and that under normal circumstances any real effect of the educational level should be stronger that what our analysis will bring out, at least if we leave out the youngest ages at marriage.

To see this, assume for instance that there is a negative gradient in the educational effect on divorce risks. Suppose that we select the lowest level of education as our basis of comparison (i.e., as our baseline level for computing relative risks). Women who have this level on our anticipatory variable have also had this level throughout, so their divorce risk should be correctly assessed by the practice that we follow. Women at a higher final level of education must have been subject to the higher divorce risks that pertain to lower levels of education during part of their life, however, so our as-

sessment of their divorce risk should be too high, as compared to what we would get if we had known their educational level at all times during their adult lives. Our upward bias could be progressively stronger the higher the woman's final educational level is, but at no stage should the overestimation be so strong that it would tip the estimated educational effect over from decrease to an increase as we proceed from one final educational level to the next. After all, a woman who ends up at a higher final educational level has been subject to the correspondingly lower divorce risks at least for some of her observed life. The total outcome should be that our procedure gives an educational gradient that is less negative than it should be, but we should not erroneously get a positive estimated gradient. In summary, our procedure should bias the educational gradient toward zero.

On closer scrutiny, this argument is actually seen to rest on a hidden premise to the effect that a woman's civil status history subsequent to first marriage formation does not much affect her ability to make educational progress, except for any effect of childbearing and other factors that may influence her behavior in the educational system. This would be the case if all education must be taken before first marriage, or if currently and previously married women have the same chances of getting more education, *ceteris paribus*. If educational progress is contingent on marital status, then things may change radically. For the sake of argument, assume, for instance, that women can get no further education when they are married but that divorced women *can* go back to school. Then a married woman must get a divorce if she wants to make educational progress. When divorce is a prerequisite for further education, the anticipatory procedure will tend to give a systematic upward bias in estimated educational effects on divorce risks. Even if divorce risks actually decline with an improving educational level, their biased estimates may come out with a positive gradient.

This has more relevance for the Swedish situation that one may suspect at first glance. A woman who marries at age 16 or 17, say, cannot have attained even a *gymnasium*-level education before marriage, simply because this level requires more schooling than she has had time for so far. Some of our respondents have married as young teenagers but have a *gymnasium* or post-*gymnasium* registered final level of education. We suspect that a divorce may have been beneficial for their educational progress, even in a country like Sweden. We are hesitant, therefore, to trust educational effects that we can estimate for women who first married at such young ages.

An empirical experiment

The verbal arguments above can be supported and made more precise by mathematics of the kind contained in Section 5.3 in Hoem and Funck Jensen (1982). In the end, however, an operational assessment must rest on calculations based on empirical knowledge about risks of change in marital status and about chances of improving one's educational level. Such information is not available at the moment; in fact, our current investigations, of which the present paper forms a part, are an effort to improve our knowledge on this front. Fortunately, our data set contains certain information that can provide a basis for some assessment, namely the window into the educational histories of the cohorts born in 1959 and later that we described in Section 2. It enables us to carry out some limited experiments to test the suitability of the anticipatory definition of the educational level for our purposes.

For the cohorts of 1959 and 1964, we have run intensity regressions for married women at positive parities with the final level of education as a covariate in one set, and with the educational level as a time-varying covariate in a parallel set. In each case, we have included an interaction between the educational variable and age at mar-

riage, while we have let all other factors appear in a complete (five-way) interaction. A summary of the results for the cohort of 1959 is given in Table 1. The figures in Table 1 are computed by an improved version of indirect standardization, described in Appendix 2. The footnote to the table lists the covariates included, given in our normal short-hand notation. (Let us repeat that the results for the cohort of 1964 are quite similar and are not displayed here.) In the following discussion, we regard the educational effects estimated when the educational level is a time-varying covariate as an unbiased mirror of real effects, and we want to see to which extent the anticipatory educational effects reflect this reality.

Table 1. Educational effects on divorce risks to married mothers in the cohort of 1959 according to a dynamic and an anticipatory definition of the educational level[a]

	Educational level				
		as a time-varying covariate		as an anticipatory covariate	
Age at marriage	pre-gymnasium[b]	*gymnasium*	post-gymn.	*gymnasium*	post-gymn.	Biasing factor[c]
16–17	1	0.513		1.372		2.67
18–19	1	0.715	0.596	0.979	0.817	1.37
20–23	1	0.717	0.620	0.834	0.723	1.16
24–28	1	0.698	0.556	0.718	0.559	1.03
29–32	1	0.944	0.800	0.930	0.781	0.99

Notes

[a] The effects are estimated in a model of the following type:
Educational-level * Age-at-marriage (16–35) + Pregnant-at-marriage * Youngest-child-at-marriage * Parity (>0) * Period * Age-of-first-child.

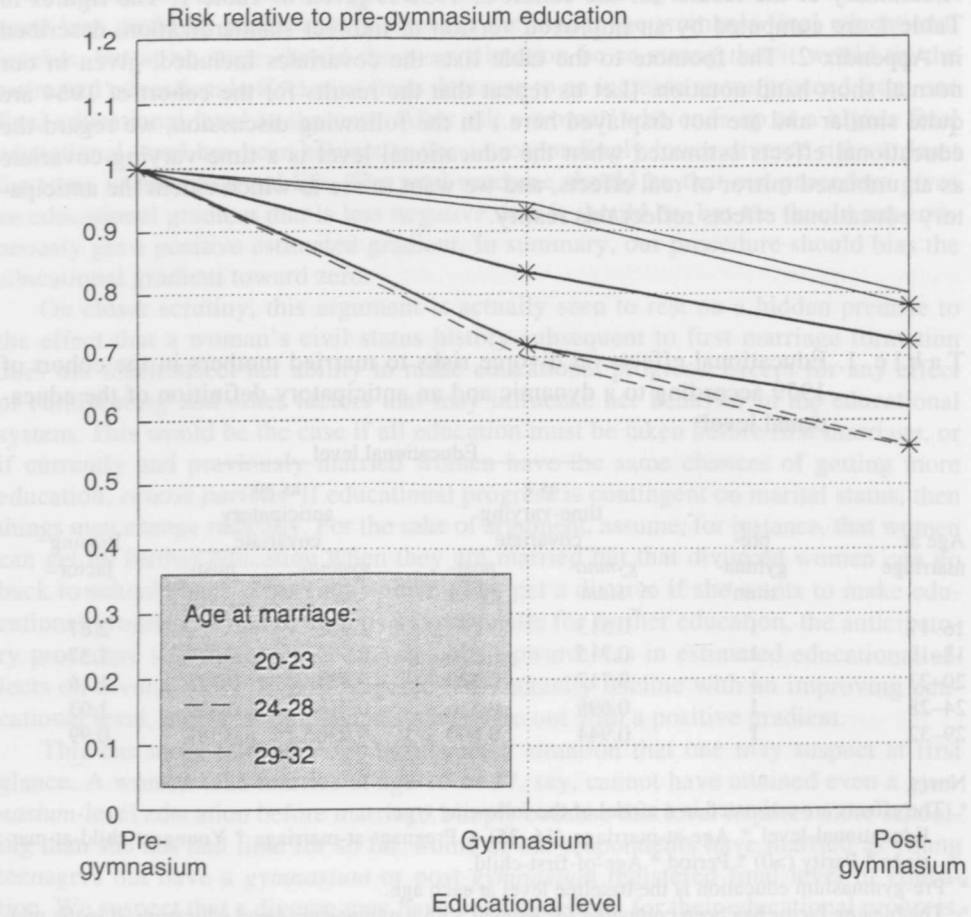
[b] Pre-gymnasium education is the baseline level at each age.

[c] The biasing factor has been computed for women with a gymnasium-level education. It is the relative risk estimated using the anticipatory educational covariate, divided by the corresponding estimate produced using the time-varying educational covariate. The corresponding estimates of the biasing factors for women at the post-gymnasium level deviate by at most 0.024.

We observe immediately that the "real" educational effects have a negative gradient for all ages at marriage in this cohort. We also see, however, that in the regression with the anticipatory educational variable, the educational effects erroneously have a positive gradient for ages 16–17 at marriage. (So few women in this group ever attain a post-*gymnasium* education that our comparison can only really be made concerning the transition from the pre-*gymnasium* to the *gymnasium* level.) Furthermore, there is a noticeable upward bias in the anticipatory estimates of educational effects for women who first marry at ages 18 or 19, but at least they only render the negative real gradient as a less negative biased gradient. For women first married at ages 20–23 or older, the anticipatory procedure provides reasonably accurate estimates of the real educational effects, and increasingly so as we pass to progressively higher age brackets. Figure 1 shows this in a graphic format.

On this basis, we have restricted our main investigation of educational effects (as reported in Hoem (1995a,b) to groups of women who first married at ages 20 or above. This only excludes about 1400 women who married as teenagers.

Figure 1. Educational effects on risks of divorce to married women born in 1959



Source: Table 1.
Unmarked curves: time-varying educational level.
Marked curves: anticipatory educational level.

Reverse causation

The above example of the teenage brides is a case of reverse causation: we look for effects of educational attainment on the divorce risk, but our analysis may pick up the effect of a divorce on a woman's educational progress. The same kind of effect may occur in our analysis if a particular line of study is mostly sought by divorced women looking for a way to make a living. Suppose hypothetically that many divorced women become recreational instructors. Then the relative divorce risk for women we have grouped in this category will be inflated in comparison to what we would have found if our educational variable were a time-varying covariate.

A special scrutiny of the data for the cohorts born in 1962–1964 suggests that this may be a problem for women who married as teenagers. About fourteen percent of such women in these three cohorts completed their education (so far as we can follow it) after a divorce, mostly by finishing the *gymnasium*. Only about two percent of our respondents in these three cohorts completed their final education after a divorce if

they married at ages 20–23, however, and there was almost no post-divorce school-completion among those who married later. We trust that reverse causation is a negligible problem for our analysis of women married at ages 20–35.

References

Dahllöf, Urban. 1990. Changes within the Swedish school system and their effect. In: The Comprehensive School Experiment Revisited: Evidence from Western Europe, edited by Achim Leschinsky and Karl Ulrich Mayer, pp. 174–209. Frankfurt am Main: Verlag Peter Lang.

Erikson, Robert and Janne Jonsson. 1993. Ursprung och utbildning: social snedrekrytering till högre studier (Origin and education: social differentials in recruitment to higher studies). Statens offentliga utredningar 1993:85. Stockholm.

Hoem, Jan M. 1987. Statistical analysis of a multiplicative model and its application to the standardization of vital rates: a review. International Statistical Review 55 (2):119–152.

— 1995a. Educational gradients in divorce risks in Sweden in recent decades. Stockholm Research Reports in Demography 84. To appear in Population Studies.

— 1995b. Educational capital and divorce risks in Sweden in the 1970s and 1980s. Stockholm Research Reports in Demography 95.

Hoem, Jan M. and Ulla Funck Jensen. 1982. Multistate life table methodology: a probabilist critique. In: Multidimensional Mathematical Demography, edited by Kenneth C. Land and Andrei Rogers, pp. 155–264. New York: Academic Press.

Johansson, Leif and Fjalar Finnäs. 1983. Fertility of Swedish women born 1927–1960. Urval No. 14. Stockholm: Statistics Sweden.

Qvist, Jan. 1990. Kvalitets- och metodfrågor vid användning av registerdata: tre fallstudier inom befolkningsstatistiken (Issues of quality and method in the use of register data: three case studies in population statistics). Bakgrundsmaterial från demografiska funktionen 1990:2. Stockholm: Statistics Sweden.

Statistics Sweden. 1988. Swedish Standard Classification of Education. Part 1. Numerical Order. Report 1988:4 on Statistical Co-ordination.

Appendix 1: A brief account of the Swedish educational system

The Swedish educational system has gone through much reform since the early 1960s. This is not the place to describe all the various types of educational histories that our respondents could have. Here are only the essentials needed for our purposes. For more information, consult Dahllöf (1990) or Erikson and Jonsson (1993).

During most of the period when our respondents were educated, compulsory primary school started at age 7 and lasted for nine years. At age 16, all children could enter *gymnasium*, which lasted for two to four years, depending on the line of study. (Our early cohorts were not always so fortunate.) What we now call the *gymnasium* encompassed both academic and vocational upper-secondary education. Most vocational secondary education took two years. A secondary-level engineering education took four years. This "technical *gymnasium*" and all three-year *gymnasium* schooling opened for entry into subsequent university-level or other post-secondary education, including teacher's colleges, police academies, nursing schools and so on. For those who never entered the *gymnasium* or who dropped out before completion, programs of adult *gymnasium* education were available.

There have been two main reforms since this system was in use. (1) Vocational *gymnasium* schooling has been extended from two to three years and the technical *gymnasium* has been reduced from four to three years. (2) Children can now start school at age 6, if their parents so decide. None of these reforms influence the education of our respondents.

Appendix 2: Generalized indirect standardization

Suppose that the information about a given population segment is cross-classified by the levels on two factors G and H, indexed by g and h. Let D_{gh} be the number of divorces recorded in the subgroup (g,h) and let R_{gh} be the corresponding person-months of exposure. Then a standard measure of the risk of divorce in population group g on factor G, relative to some baseline group on the same factor, say group 1, net of the effect of factor H, is the *Standard Divortiality Ratio*

$$d_g = D_{g.}/\Sigma_h R_{gh} s_{1h}, \qquad (1)$$

where $D_{g.}$ is the sum $\Sigma_h D_{gh}$. In this analog of the Standard Mortality Ratio, the quantity s_{1h} should be chosen as a good estimator of the divorce risk in population group (1,h). A natural first choice is to let s_{1h} be the occurrence/exposure rate D_{1h}/R_{1h}. Note that this makes $d_1=1$. Formula (1) is a version of the classical indirect standardization method.

Table 1 is based on a multidimensional extension of formula (1). In this table, the two anonymous factors G and H are replaced by the following seven factors, defined for each woman after entry into both marriage and motherhood:

- educational level, indexed by i,
- age at marriage, indexed by j, and grouped into the seven age groups given in the first column of the table,
- pregnancy status at marriage, indexed by k, with k=1 for women who were pregnant at marriage and k=0 for women who were not,
- age of any youngest child at marriage, indexed by l, with l=0 if the mother had no premarital children, and with l=1, 2, or 3 for women who had premarital children and whose youngest child at marriage was 0–2 years, 3–6 years, or 7–15 years old, respectively,
- current childbearing parity, indexed by m, for parities 0, 1, 2, and 3+, (a time-varying covariate),
- current calendar period, indexed by n, for periods 1971–1973, 1974 (to cover the year of the Swedish divorce reform), 1975–1977, 1978–1980, 1981–1984, 1985–1988, and 1989–1991 (another time-varying covariate) and
- current age of the woman's oldest child, indexed by t and suitably grouped (the basic time variable).

The index g in the simple two-factor theory is replaced by a pair (i,j) and the index h is replaced by a quintuple (k,l,m,n,t). Formula (1) is transformed as follows:

Let $D_{ijklmnt}$ be the number of divorces registered for women at educational level i who are in age group j at marriage, pregnancy status k at marriage, age group l of any youngest child at marriage, current parity m, for calendar period n and at a time when the woman's oldest child was in age group t. Let $R_{ijklmnt}$ be the corresponding months of exposure and let $D_{ij\ldots} = \Sigma_k\Sigma_l\Sigma_m\Sigma_n\Sigma_t D_{ijklmnt}$. Then the Standard Divortiality Ratio of group (i,j) is

$$d_{ij} = D_{ij}...../\Sigma_k\Sigma_l\Sigma_m\Sigma_n\Sigma_t R_{ijklmnt} s_{13klmnt}, \quad (2)$$

where $s_{13klmnt}$ is a suitable estimator of the divorce risk in group (1,3,k,l,m,n,t). This corresponds to a choice of the group with a pre-*gymnasium* education (group i=1) and age 20–23 years at marriage (group j=3) as a baseline group for the pair (i,j). If $s_{13klmnt}$ is taken to be $D_{13klmnt}/R_{13klmnt}$, then $d_{13}=1$. We interpret d_{ij} as a measure of the divorce risk for women who have educational level *i* and age group *j* at marriage, relative to the corresponding risk for women who have educational level 1 and age group 3 at marriage, when we control for the effects of the five other factors in the data.

From the d_{ij}, we can define

$$f_{ij} = d_{ij}/d_{ij} \quad (3)$$

as a measure of the divorce risk for women who have educational level *i* and age group *j* at marriage, relative to the corresponding risk for women who have educational level 1 and *the same age group* at marriage (i.e., also age group *j*, not age group 3 as for d_{ij}). This allows us to keep age at marriage constant and to concentrate on comparisons between commensurate divorce risks for women at different educational levels, net also of age at marriage.

In statistical terms, $s_{ijklmnt}$ is an estimator the divorce intensity $\sigma_{ijklmnt}$ in group (i,j,k,l,m,n,t). If this underlying model parameter satisfies a multiplicative decomposition of the form $\sigma_{ijklmnt}=\delta_{ij}\,\beta_{klmnt}$, then we get an improved (indeed optimal) version of the Standard Divortiality Ratio if we replace $s_{13klmnt}$ in (2) by a Maximum Likelihood Estimator of β_{klmnt}, say b_{klmnt}. This makes d_{ij} similarly the MLE for the model parameter δ_{ij}. (In reality, both sets of MLEs are computed simultaneously by an iterative numerical algorithm.)

The figures tabulated in Table 1 are the values of f_{ij} computed for the current data set, based on the improved method of indirect standardization just described. The method is a special case of intensity regression (hazard regression) with a piecewise constant baseline intensity.

In footnote a of Table 1, * indicates interaction and + indicates a multiplicative decomposition. The + notation is explained by the additive relation

$$\ln(\sigma_{ijklmnt}) = \ln(\delta_{ij})+\ln(\beta_{klmnt})$$

that we get by taking logarithms in the multiplicative formula for $\sigma_{ijklmnt}$.