Two sex proportional hazard model and marriage market analysis*

Toru Suzuki
National Institute of Population and Social Security Research
2-2-3 Uchisaiwaicho, Chiyoda-ku, Tokyo 100-0011 JAPAN
E-mail: suzuki-t@ipss.go.jp

Abstract:
Studies of marriage market should capture the simultaneous determination of male and female nuptiality by socio-economic characteristics of both sexes. In this line, an extension of Cox proportional hazard model into two sexes is proposed. The basic assumption is that the marital hazard of male and female characteristics (i,j) is proportional to that of reference combination (0,0). Two-sex partial likelihood function is defined by taking not an individual but a pair of male and female as the unit of analysis. The maximum likelihood estimator can be obtained by optimizing the function. Tied data can be treated in the same way as the one-sex model. Some methodological and interpretational problems are discussed. The author cannot provide either a complete computer program or a strict proof of asymptotic efficiency of the two-sex model but a small numerical example.

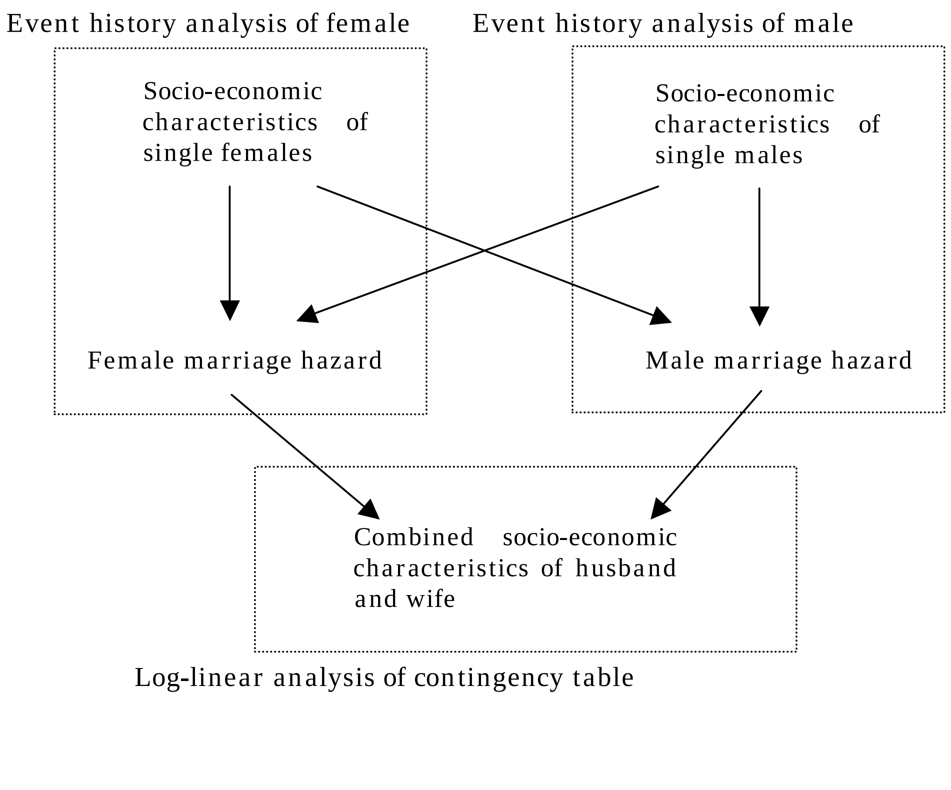
* This paper is for poster session, Statistical Demography (S36), The XXIV General Population Conference, Salvador, Bahia, Brazil, in August 2001.

1. Issue

Figure 1 illustrates the problem that conventional methods cannot examine the effects of marriage market structure on marital hazard. The event history analysis is inherently a one-sex model. Although it is a powerful tool to evaluate the effects of individuals' socio-economic characteristics on marriage, the availability or scarcity of desirable partner is out of the range. It would be essential to assume that a compositional change in characteristics of single males affect not only male nuptiality but also female nuptiality, and vice versa.

Marriage market studies have been concentrated on the analysis of contingency table of married couples. Many indices have been applied, including Blau's *OM* and *IM*, mobility ratios, Gini's *H*, Yasuda's *y*, Gray's *v*, Yule's *Q*, Goodman and Kruskal's *G*, Rockwell's hypergamy ratio, etc. Since 1980's, the log-linear analysis has been the standard method to analyze the contingency table. The problem here is that the contingency table refers to only the result of marriage hazard. Thus, it ignores persons who eventually do not marry, and the analysis of nuptiality change such as recent decline in marital hazard is out of the range.

Figure 1. Conventional methods for marriage market analysis



2. Extension of Proportional Hazard Model to Two-Sex

This paper proposes to extend the proportional hazard model (Cox regression) into two-sex. In the ordinary event history analysis, there is a simple relation between hazard and survivorship functions.

$$-h(x) = \frac{1}{S(x)}\frac{dS(x)}{dx} = \frac{d}{dx}\log S(x).$$

In a two-sex model, however, the survivorship function is not readily available. To allow a life table interpretation, the survivorship function needs be obtained with iteration as in the ETHNUP model by Schoen (1986, p. 215).

However, the estimation of coefficients in a proportional hazard model depends not on life table but on partial likelihood function. In fact, the effect of socio-economic characteristics on marriage hazard can be evaluated without referring to either hazard or survivorship function. The only necessary assumption is the proportionality between hazards.

$$h_k(x) = h_0(x)\exp(\boldsymbol{\beta}'\mathbf{z}_k), \tag{1}$$

where $h_k(x)$ is the hazard of individual k, $h_0(x)$ is the hazard of reference person whose covariate vector is [0, 0, ...], β' is the coefficient vector, and $\mathbf{z}_k$ is the covariate vector of individual k. Under the condition that there was one marriage at age x, the conditional probability that the marriage occurred to a particular person k is,

$$L(x) = \frac{h_k(x)}{\sum_k h_k(x)} = \frac{h_0(x)\exp(\boldsymbol{\beta}'\mathbf{z_k})}{h_0(x)\sum_k \exp(\boldsymbol{\beta}'\mathbf{z_k})} = \frac{\exp(\boldsymbol{\beta}'\mathbf{z_{(x)}})}{\sum_k \exp(\boldsymbol{\beta}'\mathbf{z_k})},$$

where $\mathbf{z_{(x)}}$ is the covariate vector of the individual k who actually married at age x. The denominator, risk set, is the sum of all hazards of the population at risk. Assume that all the covariates are indicator variables and let i be the identifier of pattern of covariates. If there are c covariates, the maximum number of i would be between c and 2^c. The former materializes when c covariates represent categories of one variable such as education or occupation, excluding a reference category. The latter materializes when c covariates represent different dichotomous factors. Let $N(x,i)$ be the number of survivors with covariate pattern i. Then,

$$L(x) = \frac{\exp(\boldsymbol{\beta}'\mathbf{z_{(x)}})}{\sum_i N(x,i)\exp(\boldsymbol{\beta}'\mathbf{z_{(i)}})},$$

where $\mathbf{z_{(i)}}$ is the 0-1 pattern identified with i. If there is no tied data, the partial likelihood is the product of these conditional probabilities. The coefficients are estimated so that they maximize the following function.

$$\log L = \sum_x \log L(x) = \sum_x \log \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{x})})}{\sum_i N(x,i)\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{i})})}. \tag{2}$$

To extend this model to two-sex, not an individual but a pair of male and female should be considered as the unit of analysis. Every possible pair of single male and single female is thought to have hazard of marriage. Assume that there were $N_f(x)$ females and $N_m(y)$ males at risk at joint ages *(x,y)*, and a pair identified with *k* actually married. The basic assumption of proportionality is,

$$h_k(x,y) = h_0(x,y)\exp(\boldsymbol{\beta}'\mathbf{z}_\mathbf{k}). \tag{3}$$

Let $\mathbf{z}_{(\mathbf{xy})}$ be the 0-1 pattern of the pair *k* that actually married at joint age *(x,y)*. If there are *c* indicator variables for an individual male and female, the length of the vector **z** should be 2*c*. The number of possible patterns should be between c^2 and 2^{2c}. Suppose that there were $N_f(x,i)$ females whose 0-1 pattern is identified with *i* and $N_m(y,j)$ males whose 0-1 pattern is identified with *j*. Let $\mathbf{z}_{(\mathbf{ij})}$ be the combination of female pattern *i* and male pattern *j*. The number of pairs that have pattern *(ij)* should be the product of $N_f(x,i)$ and $N_m(y,j)$. For each joint age at which a marriage took place, an element of partial likelihood function is,

$$L(x,y) = \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{xy})})}{\sum_{i,j} N_f(x,i)N_m(y,j)\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{ij})})}. \tag{4}$$

If there is no tied data, the partial likelihood function is simply the product of (4). Coefficient vector can be estimated by maximizing the following function. Ties can be treated by Breslow or Efron method in the same way as in the ordinary proportional hazard model.

$$\log L = \sum_x \sum_y \log L(x,y) = \sum_x \sum_y \log \frac{\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{xy})})}{\sum_{i,j} N_f(x,i)N_m(y,j)\exp(\boldsymbol{\beta}'\mathbf{z}_{(\mathbf{ij})})}. \tag{5}$$

3. Example

A small hypothetical example is presented here. Suppose that two single women, four single men and eight married couples were surveyed. Only one indicator variable is considered here. Table 1 shows that the third and fourth couple makes a tie at joint age (28, 29).

Table 1. Hypothetical Data

Singles	Female#	Current Age	Level	Male#	Current Age	Level
	1	27	1	1	30	0
	2	25	0	2	24	0
				3	38	1
				4	33	1
Couples	**Wife#**	**Age at Marriage**	**Level**	**Husband#**	**Age at Marriage**	**Level**
	1	23	0	1	26	0
	2	22	0	2	30	1
	3	28	0	3	29	0
	4	28	1	4	29	1
	5	18	0	5	22	0
	6	29	1	6	33	1
	7	24	1	7	24	0
	8	30	1	8	30	1

Since there are ten women and twelve men, there should be 120 possible pairs in the beginning. The number of surviving pairs decreases with marriage and censoring. The following table indicates the combined level (i, j) of the couple that actually married at joint age (x, y) and the number of surviving pairs by combined level. Tied data was processed by Efron's approximation.

Table 2. Events and Survivors by Combined Level

	Events				Surviving Pairs			
Joint Age	(0,0)	(0,1)	(1,0)	(1,1)	(0,0)	(0,1)	(1,0)	(1,1)
18,22	1	0	0	0	30	30	30	30
22,30	0	1	0	0	4	20	5	25
23,26	1	0	0	0	9	18	15	30
24,24	0	0	1	0	10	12	25	30
28,29	1	0	0	0	2	6	6	18
28,29	0	0	0	1	1.5	6	6	17.5
29,33	0	0	0	1	0	0	0	6
30,30	0	0	0	1	0	0	1	5

A Japanese version of S-Plus for Windows (4.0J) was used to estimate the coefficients. First, matrix of events and that of pairs at risk were produced as S-object named coef.test$cnum and coef.test$cden, respectively.

```
> coef.test
$cnum:
     0;0 0;1 1;0 1;1
18,22   1   0   0   0
22,30   0   1   0   0
23,26   1   0   0   0
24,24   0   0   1   0
28,29   1   0   0   0
28,29   0   0   0   1
29,33   0   0   0   1
30,30   0   0   0   1

$cden:
      0;0 0;1 1;0  1;1
18,22 30.0  30  30 30.0
22,30  4.0  20   5 25.0
23,26  9.0  18  15 30.0
24,24 10.0  12  25 30.0
28,29  2.0   6   6 18.0
28,29  1.5   6   6 17.5
29,33  0.0   0   0  6.0
30,30  0.0   0   1  5.0
```

The two-sex partial likelihood function was defined as a user function named partlik().

```
#Partial Likelyhood Function (Sign Reversed)
prtlik <- function(bval,nummat,denmat)
{
        - sum(nummat %*% log(c(1,bval))) + sum(log(denmat %*% c(1,bval)))
}
```

Coefficients were estimated using S function nlminb().

```
#Applying nlminb()
soltn <- nlminb(start=rep(0.01,3),objective=prtlik,nummat=coef.test$cnum,denmat=coef.test$cden)

> soltn$parameters
[1] 0.14108725 0.12699803 0.08642293

#Partial likelihood of null hypothesis and alternative hypothesis
loglik.nul <- -prtlik(bval=rep(1,times=3),nummat=coef.test$cnum,denmat=coef.test$cden)
loglik.alt <- -prtlik(bval=soltn$parameters,nummat=coef.test$cnum,denmat=coef.test$cden)

#Likelihood ratio test
likratio <- 2 * (loglik.alt - loglik.nul)

> 1 - pchisq(likratio,3)
[1] 0.114021
```

Although the result of likelihood ratio test is not significant, the indicator variable seems to imply smaller hazard of marriage. The table below suggests that it is unlikely for a pair that both are indicated "YES" to marry.

Table 3. Coefficients Estimated

		Male	
		j=0	j=1
Female	i=0	$\exp(\beta_{00}) = 1$	$\exp(\beta_{01}) = 0.14108725$
	i=1	$\exp(\beta_{10}) = 0.12699803$	$\exp(\beta_{11}) = 0.08642293$

4. Discussion

This paper is just a presentation of an immature idea and the author cannot provide a strict proof of asymptotic centrality (Fleming and Harrington, 1991) for the two-sex proportional hazard model. This model contains some difficulties in addition to theoretical validity.

The most serious problem would be the range of marriage market. In ordinary Cox regression of marriage, the availability of the opposite sex population is seen as environment and out of the model. The two-sex model, however, processes all possible pairs of male and female as the unit of analysis and assumes that all those pairs have risk of marriage. Using a national sample survey would imply that there is risk of marriage between male and female living hundred or thousand kilometers apart.

Use of retrospective data causes another problem of marriage market identification.

Even though two couples married at the same joint age (x, y), the year of marriage would be different if their joint ages at the survey date are different. One may want to identify participants in marriage market at one point of time. In such a case, a tie appears only when two couples married in the same year at the same joint age (x, y). Complicated data processing would be required for such a two-sex marriage market analysis.

References

Fleming, Thomas R., and David P. Harrington, 1991

Counting Processes and Survival Analysis, New York, A Wiley-Interscience Publication.

Schoen, Robert, 1988

Modeling Multigroup Populations, New York, Plenum Press.

Statistical Sciences, 1995

S-PLUS Guide to Statistical and Mathematical Analysis, Version 3.3, Seattle, StatSci, a division of MathSoft, Inc.