STATISTICAL ANALYSIS OF TENDENCY TO MARRY FOR COHABITANTS

Author: Xuan Lu

Supervisor: Rolf Larsson

Master Thesis in Statistics, June 2010
Uppsala University
Statistical Analysis of Tendency to Marry for Cohabitants

Author: Xuan Lu

Supervisor: Rolf Larsson

June, 2010

Abstract

The marriage rate is getting lower and lower and we are interested in what factors have an effect on the tendency to marry. In this paper we use the retrospective data set based on 1365 ever-cohabited men who were either married or still in cohabitation and focus on the two factors which are reported, highest educational level and the cohabitation durations which are classified into five groups by exposure time. The maximum likelihood procedure is proposed to estimate the effect parameters based on a multiplicative two-factor hazard model. After that, we use the bootstrap approach to estimate the standard errors and to obtain the confidence intervals of the estimates. The results show that the educational level and the cohabitation duration have an effect to the tendency of marriage. The aim is to find out the tendency to marry for the cohabitants. We also expect to analyze how the effects like educational level and cohabitation duration work on the marriage tendency. The problems need to be solved are discussed and we give some suggestions for further research.

Keywords: Retrospective data, Multiplicative model, Maximum likelihood, Bootstrap approach.

1 Email: xuaner0216@hotmail.com
2 Email: Rolf.Larsson@dis.uu.se
## Contents

1. Introduction ......................................................... 4
   1.1 Background .................................................. 4
   1.2 Aim ........................................................... 5
   1.3 Outline ....................................................... 5
2. Methodology ......................................................... 6
   2.1 The Standard Multiplicative Model ......................... 6
   2.2 Maximum Likelihood Estimators ......................... 7
   2.3 Bootstrap .................................................. 9
      2.3.1 Bootstrap Standard Errors ....................... 9
      2.3.2 Bootstrap Confidence Intervals ............... 10
3. Data .............................................................. 11
   3.1 Data Description ........................................... 11
   3.2 Data Selection .............................................. 13
4. Estimation Results ............................................... 15
5. Conclusion and Discussion ....................................... 17
1 Introduction

1.1 Background

“Marriage is slowly dying in Scandinavia and a majority of children in Sweden are born out of wedlock” (Kurtz, 2004). A lot of young couples select to live together without marriage. Then what factors affect the tendency to marry for the couples who already live together?

In this paper we have two effect factors: one is the number of years of cohabitation, the other is the educational level. We expect to find out to what extent they reflect the marriage tendency. In many cases, the events being recalled of marriage or still in cohabitation took place some years before the survey data. The quality of retrospective data is generally considered to cause a time-inconsistency problem because the data pertain to the time of event in the distant past (Haaga, 1988; Ghilagaber & Larsson, 2009). So that we encountered one problem for such retrospective survey data which is that the educational level at time for the cohabitation or marriage is unknown. We only know the time at which the highest educational level was attained for the individuals at the date of survey. And for some individuals, this may occur after the time of cohabitation. To solve this problem, we use two methods to process the data. One solution is simply not to take these individuals whose highest educational level time occurred after their cohabitation time into account, i.e. use what we call the reduced data. The other solution is to use all individuals in the analysis and assume that the individuals all attained their highest educational level prior to cohabitation.

Regarding the related problems, Hayford and Morgan (2008) urge caution in the use of retrospective data on cohabitation because of finding consistent discrepancies by analyzing the entry rate of cohabitation from the data collected in four major U.S. family surveys using a event-history analysis. They argue that the pattern of differences suggests that relative to cohabitation rates estimated closer to the time of survey, cohabitation retrospective data may underestimate rates in distant periods. In our view, this may not be comprehensive because the author does not take the individual characteristic in to account. While Haaga (1988) studies more than 1200 Malaysian women to retest the reliability of the retrospective survey data on breastfeeding duration and the type of supplementary food and then gets a
conclusion that the respondents’ characteristics play a more important role than the length of the period between the event time and the time of survey using a logistic regression analysis, and I agree with this view more. Ghilagaber and Larsson (2009) propose a maximum likelihood procedure to estimate the effects of educational level on divorce risks in a multiplicative piecewise-constant hazard model using anticipatory analysis and then the results show that the educational levels are more important determinants. However, regarding the same retrospective survey data, Ghilagaber and Koskinen (2009) propose a Bayesian approach to solve the relative problem and correct the biases in covariate effect estimators by reason of using anticipatory analysis. Ghilagaber and Larsson accomplish the same thing using maximum likelihood.

1.2 Aim

In this paper, the aim is to find out the tendency to marry for the cohabitants. We also expect to analyze how the effects like educational level and cohabitation duration work on the marriage tendency. After that, we will compare the results obtained by two different data processing methods to see whether the two outcomes are similar or completely contrary.

1.3 Outline

We first give a simple background of the article. In section 2, we introduce the methodology of the standard multiplicative model and maximum likelihood which will be used to estimate the effects in our analysis. We also give a simple introduction of the bootstrap standard errors. In section 3, the data is described in detail and we give a summary of the data structure. In section 4, we give the results of the data analysis and then compare the two groups of results obtained by reduced data and the entire data. In the final section, we have a conclusion on what we have done and discuss the problems for further research.
2 Methodology

2.1 The Standard Multiplicative Model

Related multiplicative models are considered in Breslow and Day (1975) and Hoem (1987) reviews the form of a multiplicative model and its statistical properties. After that, Ghilagaber & Larsson (2009) and Ghilagaber & Koskinen (2009) develop the multiplicative model applied to the data set similar to mine.

In our case, for the sample of the investigated individuals, $D_{ij}$ denotes the number of occurrences in a specific time period. That is to say, in the $i^{th}$ ($i = 1, \ldots, I$) cohabitation duration group of the $j^{th}$ ($j = 1, \ldots, J$) educational level, $D_{ij}$ is the number of individuals who marry. Let $T_{ij}$ be the exposure years to marry that are grouped into 5 classes in cohabitation duration. The exposure time is different but constant in different time intervals $I$. We assume that the exposure time $T_{ij}$ follows the exponential distribution. Under this assumption, for the $k^{th}$ person at cohabitation duration $I$ of the $j^{th}$ educational level, we may obtain the density function of the time to marry as follows. (Ghilagaber and Larsson, 2009)

$$f(t_{ijk}) = \lambda_{ij} \exp(-\lambda_{ij}t_{ijk})$$

where $\lambda_{ij}$ is the rate parameter of marriage in the $i^{th}$ cohabitation duration group of the $j^{th}$ educational level.

Define

$$D_{i+} = \sum_{j=1}^{J} D_{ij}, \quad D_{+j} = \sum_{i=1}^{I} D_{ij}, \quad T_{i+} = \sum_{j=1}^{J} T_{ij}, \quad T_{+j} = \sum_{i=1}^{I} T_{ij},$$

and

$$D_{++} = \sum_{i=1}^{I} D_{i+} = \sum_{j=1}^{J} D_{+j} = \sum_{i=1}^{I} \sum_{j=1}^{J} D_{ij}, \quad T_{++} = \sum_{i=1}^{I} T_{i+} = \sum_{j=1}^{J} T_{+j} = \sum_{i=1}^{I} \sum_{j=1}^{J} T_{ij}$$

A particular multiplicative model for the rate parameter $\lambda_{ij}$ is expressed in the equation

$$\lambda_{ij} = \beta_i \alpha_j$$
whereby the duration-specific marriage rates are obtained from multiplicative contributions for the \(i^{th}\) cohabitation duration group \((\beta_i)\) and the \(j^{th}\) educational level \((\alpha_j)\). Meanwhile, \(\beta_i\) measures the marriage tendency at cohabitation duration group \(i\), while \(\alpha_j\) represents the tendency of marriage at the \(j^{th}\) educational level. Both \(\beta_i\) and \(\alpha_j\) must be estimated from the data set. Sometimes the issues become simplified if we select the standard comparison. In this paper, \(\alpha_1\) is limited by \(\alpha_1 = 1\). Then the model in (4) has \(I + J - 1\) parameters which are named by \(\beta_1, \beta_2, \ldots, \beta_I\) and \(\alpha_2, \alpha_3, \ldots, \alpha_J\).

### 2.2 Maximum Likelihood Estimators

The statistical method of maximum likelihood is the most popular method for deriving the estimators.

**Definition 1** For each sample point \(x\), let \(\hat{\theta}(x)\) be a parameter value at which \(L(\theta \mid x)\) attains its maximum as a function of \(\theta\), with \(x\) held fixed. A maximum likelihood estimator (MLE) of the parameter \(\theta\) based on a sample \(X\) is \(\hat{\theta}(X)\). (Casella and Berger, 2002)

To construct the likelihood function, first we need to let \(d_{ijk}\) be an indicator variable under the function (4). That is to say, for the \(k^{th}\) person at cohabitation duration \(I\) of the \(j^{th}\) educational level, \(d_{ijk} = 1\) represents the person is married while \(d_{ijk} = 0\) means the person is still in cohabitation. Then \(D_{ij} = \sum_k d_{ijk}\). Using equations (1) and (4), we may derive the likelihood function for the \(i^{th}\) cohabitation duration group and the \(j^{th}\) educational level as follows. (Ghilagaber and Larsson, 2009)

\[
L_{ij} = \prod_k [\lambda_{ij}^d d_{ijk} \exp(-\lambda_{ij} t_{ijk})] = \prod_k [(\beta_i \alpha_j)^d d_{ijk} \exp(-\beta_i \alpha_j t_{ijk})] = (\beta_i \alpha_j)^{D_{ij}} \exp(-\beta_i \alpha_j T_{ij})
\]

(5)

where \(T_{ij} = \sum_k t_{ijk}\).

Next for the entire sample, the likelihood function will be expressed at all levels of \(i\) and \(j\).

\[
L = \prod_i \prod_j L_{ij} = \prod_i \prod_j [(\beta_i \alpha_j)^{D_{ij}} \exp(-\beta_i \alpha_j T_{ij})]
\]

(6)
and
\[ \log L = \sum_i \sum_j D_{ij} \log(\beta_i \alpha_j) - \sum_i \sum_j \beta_i \alpha_j T_{ij} \] (7)

\[ = \sum_i \sum_j D_{ij} \log(\beta_i) + \sum_i \sum_j D_{ij} \log(\alpha_j) - \sum_i \sum_j \beta_i \alpha_j T_{ij} \]

\[ = \sum_i D_{i+} \log \beta_i + \sum_j D_{+j} \log(\alpha_j) - \sum_i \sum_j \beta_i \alpha_j T_{ij} \]

The partial derivatives, with respect to \( \beta_i \) and \( \alpha_j \) separately, are
\[ \frac{\partial}{\partial \beta_i} \log L = \frac{D_{i+}}{\bar{\alpha}_j T_{ij}} - \sum_j \alpha_j T_{ij} \] (8)

\[ \frac{\partial}{\partial \alpha_j} \log L = \frac{D_{+j}}{\bar{\beta}_i T_{ij}} - \sum_i \beta_i T_{ij} \] (9)

Set these partial derivatives equal to 0 and then we can obtain the solution
\[ \hat{\beta}_i = \frac{D_{i+}}{\sum_j \hat{\alpha}_j T_{ij}}, \quad i = 1, 2, \ldots I \] (10)

\[ \hat{\alpha}_j = \frac{D_{+j}}{\sum_i \hat{\beta}_i T_{ij}}, \quad j = 1, 2, \ldots J \] (11)

Since the initial value of \( \alpha_j^{(1)} = 1 \) for \( \alpha_j \), we may insert this value to equation (10) to derive \( \hat{\beta}_i^{(1)} = \frac{D_{i+}}{T_{i+}} \) as the initial estimate of \( \beta_i \). This is not the explicit solution in the \( j^{th} \) educational level obtained by pooling the \( I \) cohabitation duration groups. The second cycle is
\[ \hat{\beta}_i^{(2)} = \frac{D_{i+}}{\sum_j \hat{\alpha}_j^{(2)} T_{ij}} \] (12)

and
\[ \hat{\alpha}_j^{(2)} = \frac{D_{+j}}{\sum_i \hat{\beta}_i^{(1)} T_{ij}} = \frac{D_{+j}}{\sum_i \hat{\beta}_i^{(1)} T_{ij}} \] (13)

After each cycle \( n = 1, 2, \ldots \) the estimates results \( \hat{\beta}_i^{(n)} \) and \( \hat{\alpha}_j^{(n)} \) are
serted in (7) and the procedure will draw to an end when \( \hat{\beta}_i^{(n)} \) and \( \hat{\alpha}_j^{(n)} \) converge to the maximum likelihood values of \( \hat{\beta}_i \) and \( \hat{\alpha}_j \).

This iteration process is suggested by Breslow and Day (1975). From these equations we may conclude that if we have known the number of occurrences \( (D_{ij}) \) and the exposure years \( (T_{ij}) \) at cohabitation duration \( I \) of the \( j^{th} \) educational level, we may easily derive the estimated values by the iteration procedure.

2.3 Bootstrap

2.3.1 Bootstrap Standard Errors

Bootstrap techniques, in some cases, expand the capacity of statistical inference which provides a method of calculating the standard errors. Bootstrap standard errors can sometimes replace the direct calculations of standard errors when the calculations are difficult. Standard errors estimated by bootstrap can reflect the small sample properties better. (Karlis and Kostaki, 2002)

Considering the same sample introduced in 2.1, we have derived the estimates of \( \beta_i \) and \( \alpha_j \). The steps of the bootstrap algorithm for estimating standard errors are (Johnson, 2001):

1 Select \( B (B = 1000) \) independent bootstrap samples from the given data with replacement.
2 Calculate the specific statistic, say, \( \hat{\beta}_i \) and \( \hat{\alpha}_j \) from each of the new sample.
3 Use the standard deviation of the \( B \) replications to estimate the standard error.

\[
\hat{se}(\beta_i) = \sqrt{\text{Var}_B^*(\hat{\beta}_i)} = \sqrt{\frac{1}{B-1} \sum_{n=1}^{B} (\hat{\beta}_i^{*n} - \overline{\hat{\beta}_i^*})^2} \quad (14)
\]
\[
\hat{se}(\alpha_j) = \sqrt{\text{Var}_B^*(\hat{\alpha}_j)} = \sqrt{\frac{1}{B-1} \sum_{n=1}^{B} (\hat{\alpha}_j^{*n} - \overline{\hat{\alpha}_j^*})^2} \quad (15)
\]

where \( \hat{\beta}_i^{*n}, \hat{\alpha}_j^{*n} \) are the estimates calculated from the \( n^{th} \) resample and \( \overline{\hat{\beta}_i^*} = \frac{1}{B} \sum_{n=1}^{B} \hat{\beta}_i^{*n}, \overline{\hat{\alpha}_j^*} = \frac{1}{B} \sum_{n=1}^{B} \hat{\alpha}_j^{*n} \) the mean of the resampled values.(Casella and Berger, 2002)
2.3.2 Bootstrap Confidence Intervals

In this section the bootstrap approach is used to produce the confidence intervals. Here we need to mention that the results of bootstrap standard errors in section 2.3.1 are not used to calculate confidence intervals. We use the bootstrap procedure to obtain the confidence intervals by using the empirical distribution. Hult and Lindskog (2009) introduce the procedure in their lecture notes. Let us take a simple example for explanation. Suppose we have observations \( x = (x_1, x_2, \ldots, x_n) \) from the random variables \( X_1, \ldots, X_n \) and \( F_n \) is the empirical distribution which puts point masses \( 1/n \) at the points \( X_1, \ldots, X_n \). We may resample from \( F_n \) simply by drawing with replacement among \( X_1, \ldots, X_n \) and let the resample be \( X_1^*, \ldots, X_n^* \). Then we may resample many times, say, \( N \) times to produce new samples \( X_1^{*(i)}, \ldots, X_n^{*(i)} \), \( i = 1, \ldots, N \). To obtain the empirical distribution, we use these samples to calculate the estimate, for instance \( \theta \), and \( \theta_i^* = \hat{\theta}(X_1^{*(i)}, \ldots, X_n^{*(i)}) \). Then the empirical distribution \( F_{\theta} \) is expressed by

\[
F_{\theta}(x) = \frac{1}{N} \sum_{i=1}^{N} 1_{[\theta_i^*, \theta_i^*]}(x)
\]

Then the confidence intervals are constructed as the 2.5% and 97.5% fractiles of the empirical distribution, i.e. \( q_{(1-p)/2} F_{\theta} \), \( q_{(1+p)/2} F_{\theta} \). If the sample in our case ordered as \( \theta_1^* \leq \cdots \leq \theta_N^* \), then we may get the confidence interval as \( (\theta_{(N(1-p)/2)+1}, \theta_{(N(1+p)/2)+1}) \). In our case, the steps of bootstrap confidence intervals are (Johnson, 2001):

1. Resample \( B \) times among the observations to obtain new samples.
2. Calculate the estimates, say, \( \hat{\beta}_i \) and \( \hat{\alpha}_j \) for each of the new sample.
3. Construct a confidence interval \( I_p \) with the confidence level \( p = 95\% \) as

\[
I_p(\beta_i) = (\beta_i^{*[((B(1-p)/2)+1), B]}, \beta_i^{*[((B(1+p)/2)+1), B]})
\]

\[
I_p(\alpha_j) = (\alpha_j^{*[((B(1-p)/2)+1), B]}, \alpha_j^{*[((B(1+p)/2)+1), B]})
\]

where \( \beta_i^{*[1,B]} \leq \cdots \leq \beta_i^{*[B,B]} \) is the ordered sample of \( \beta_i^*, \ldots, \beta_i^* \), and this explanation is also applied to \( \alpha_j^* \).
3 Data

3.1 Data Description

The data set used in this paper is an extract from the 1985 Mail Survey of Swedish men and its information was collected on background variables as well as an entry into and exit form marital and non-marital unions which is detailed retrospective history (Ghilagaber and Larsson, 2009). There are 1365 observations who were either married or still in cohabitation which are organized as follows:

Tab. 1: Summary of the data structure for the sample of 1365 individuals

<table>
<thead>
<tr>
<th>Cohabitation-duration</th>
<th>Educational level</th>
<th>Status</th>
<th>Cohabitation age</th>
<th>Highest educational level age</th>
<th>Exposure years</th>
<th>Anticipatory</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>Min.: 16.08</td>
<td>Min.: 10.08</td>
<td>Min.: 0.08</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>Mean: 23.15</td>
<td>Mean: 20.06</td>
<td>Mean: 4.718</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Max.: 40.42</td>
<td>Max.: 46.83</td>
<td>Max.: 27.83</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Row 1: Name of the variables.

Column 1: Cohabitation duration. The durations of cohabitation are grouped in 5 classes by exposure years. The detailed classification is shown as follows.

CohabDur = 1 corresponds to Exposures less than 2 years.
CohabDur = 2 corresponds to Exposures of between 2 years (inclusive) and 4 years (exclusive).
CohabDur = 3 corresponds to Exposures of between 4 years (inclusive) and 6 years (exclusive).
CohabDur = 4 corresponds to Exposures of between 6 years (inclusive) and 10 years (exclusive).
CohabDur = 5 corresponds to Exposures of 10 years (inclusive) or above.

Column 2: Educational level. There are grouped in three levels: 1, primary; 2, secondary; 3, Tertiary.

Column 3: Status. There are two status: status=0 represents that the individual is still in cohabitation by the survey time while status = 1 means that the respondent gets married.
Column 4: Cohabitation age - age at Cohabitation.

Column 5: Highest educational level age - age at Completion of the reported highest educational level.

Column 6: Exposure years. It is these exposure years that are grouped into 5 classes in cohabitation duration.

Column 7: Anticipatory variable. Indicator of whether the Education-variable is anticipatory.

We may see the relationship between the age at time of attainment of the reported highest educational level and the age at cohabitation clearly in Figure 1.

![Fig. 1: The relationship between the age at time of attainment of the highest educational level and the age at cohabitation.](image)

Figure 1 shows us that for a large proportion of individuals, the time at which they have obtained their reported highest educational level is prior to the time they cohabit. What is more, although the individuals belong to different educational levels and the ages when they obtain the highest educational level display a significant difference, their cohabitation ages concentrate in the interval between twenty and thirty years old.
3.2 Data Selection

In retrospective surveys, the anticipatory variables whose values refer to the contents and results of the investigation by the data of interview are common in such surveys and the values are used to explain the behavior occurred prior to the survey (Ghilagaber and Koskinen, 2009). We may draw a table to illustrate the summary of the data structure for the sample across the anticipatory status of education.

Table 2: Summary of data structure for the sample of 1365 Swedish men across the anticipatory status of education

<table>
<thead>
<tr>
<th>Anticipatory Status</th>
<th>Highest educational level</th>
<th>Married</th>
<th>Still in cohabitation</th>
<th>Total</th>
<th>% married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Anticipatory</td>
<td>First-level</td>
<td>302</td>
<td>110</td>
<td>412</td>
<td>73.30</td>
</tr>
<tr>
<td></td>
<td>Second-level</td>
<td>366</td>
<td>178</td>
<td>544</td>
<td>67.28</td>
</tr>
<tr>
<td></td>
<td>Third-level</td>
<td>72</td>
<td>21</td>
<td>93</td>
<td>77.42</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>740</td>
<td>309</td>
<td>1049</td>
<td>70.54</td>
</tr>
<tr>
<td>Anticipatory</td>
<td>First-level</td>
<td>10</td>
<td>1</td>
<td>11</td>
<td>90.91</td>
</tr>
<tr>
<td></td>
<td>Second-level</td>
<td>84</td>
<td>43</td>
<td>127</td>
<td>66.14</td>
</tr>
<tr>
<td></td>
<td>Third-level</td>
<td>139</td>
<td>39</td>
<td>178</td>
<td>78.09</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>233</td>
<td>83</td>
<td>316</td>
<td>73.73</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>973</td>
<td>392</td>
<td>1365</td>
<td>71.28</td>
</tr>
</tbody>
</table>

Table 2 illustrates that the sub-total average percentages of marriage in non-anticipatory and anticipatory status are both close to the total average marriage percentage. How to deal with the anticipatory variables, we have mentioned in the introduction part. That is, we use two methods to process the data set. Since there are only 316 individuals whose highest educational level time occurred after their cohabitation time, one possible solution is simply not to take these individuals into account, i.e. use what we call the reduced data. In the reduced data model, we ignore the observations whose age at highest educational level achievement is larger than that of cohabitation. The other solution is to use all individuals in the analysis, i.e. use what we call the entire data and assume that the individuals all attained their highest educational level prior to cohabitation. To be more specific, consider a person $k$, and $t_k$ denotes the age at acquiring the reported highest educational level while $t_k^*$ denotes the age at cohabitation. In the reduced data model, the observations of the individuals whose $t_k > t_k^*$ (anticipatory
variable=1) are ignored. In the entire data model, we assume that all the individuals follow $t_k < t_k^*$ and then we take all the observations into account. The direct distributions of the two groups for the sample are shown in Table 3 and 4.

**Tab. 3: Distribution of the data regarding the educational effect processed by two different methods**

<table>
<thead>
<tr>
<th>Data</th>
<th>Highest educational level</th>
<th>Married</th>
<th>Still in cohabitation</th>
<th>Total</th>
<th>% married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced data</td>
<td>First-level</td>
<td>302</td>
<td>110</td>
<td>412</td>
<td>73.30</td>
</tr>
<tr>
<td></td>
<td>Second-level</td>
<td>366</td>
<td>178</td>
<td>544</td>
<td>67.28</td>
</tr>
<tr>
<td></td>
<td>Third-level</td>
<td>72</td>
<td>21</td>
<td>93</td>
<td>77.42</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>740</td>
<td>309</td>
<td>1049</td>
<td>70.54</td>
</tr>
<tr>
<td>Entire data</td>
<td>First-level</td>
<td>312</td>
<td>111</td>
<td>423</td>
<td>73.76</td>
</tr>
<tr>
<td></td>
<td>Second-level</td>
<td>450</td>
<td>221</td>
<td>671</td>
<td>67.06</td>
</tr>
<tr>
<td></td>
<td>Third-level</td>
<td>211</td>
<td>60</td>
<td>271</td>
<td>77.86</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>973</td>
<td>392</td>
<td>1365</td>
<td>71.28</td>
</tr>
</tbody>
</table>

**Tab. 4: Distribution of the data regarding the cohabitation duration processed by two different methods**

<table>
<thead>
<tr>
<th>Data</th>
<th>Cohabitation duration</th>
<th>Married</th>
<th>Still in cohabitation</th>
<th>Total</th>
<th>% married</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduced data</td>
<td>Cohabitation duration 1</td>
<td>381</td>
<td>18</td>
<td>399</td>
<td>95.50</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 2</td>
<td>180</td>
<td>41</td>
<td>221</td>
<td>81.45</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 3</td>
<td>97</td>
<td>54</td>
<td>151</td>
<td>64.24</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 4</td>
<td>69</td>
<td>78</td>
<td>147</td>
<td>49.94</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 5</td>
<td>13</td>
<td>118</td>
<td>131</td>
<td>9.92</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>740</td>
<td>309</td>
<td>1049</td>
<td>70.54</td>
</tr>
<tr>
<td>Entire data</td>
<td>Cohabitation duration 1</td>
<td>508</td>
<td>18</td>
<td>526</td>
<td>96.58</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 2</td>
<td>228</td>
<td>46</td>
<td>274</td>
<td>83.21</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 3</td>
<td>130</td>
<td>61</td>
<td>191</td>
<td>68.06</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 4</td>
<td>85</td>
<td>97</td>
<td>182</td>
<td>46.70</td>
</tr>
<tr>
<td></td>
<td>Cohabitation duration 5</td>
<td>22</td>
<td>170</td>
<td>192</td>
<td>11.46</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>973</td>
<td>392</td>
<td>1365</td>
<td>71.28</td>
</tr>
</tbody>
</table>

From Table 3 we may see that among different reported highest educational levels, there may not exist a great impact to the marriage percentage differences. In terms of different highest educational levels, not considering other possible effects, the marriage percentage differs from each other
4 Estimation Results

Using the techniques in the methodology part, first we derive the number of occurrences ($D_{ij}$) and the exposure years ($T_{ij}$) at cohabitation duration $I$ of the $j^{th}$ educational level and the results of $D_{ij}$ and $T_{ij}$ are shown in Appendix A. Then, we use the maximum likelihood procedure to calculate the estimates of contributions for the $i^{th}$ cohabitation duration group $\beta_i$ and the $j^{th}$ educational level ($\alpha_j$). Meanwhile, we set the lowest educational level (the first-level) as the baseline level and the initial value of $\alpha$ is limited to $\alpha_1 = 1$. Finally, the bootstrap approach is used to estimate the standard errors.
Tab. 5: Estimated results of $\beta_i$, $\alpha_j$, bootstrap standard errors and 95% confidence intervals from the reduced data and the entire data

<table>
<thead>
<tr>
<th></th>
<th>Reduced data</th>
<th></th>
<th></th>
<th>Entire data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Standard error</td>
<td>Confidence Intervals</td>
<td>Estimate</td>
<td>Standard error</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.2410</td>
<td>0.0037</td>
<td>(0.2097, 0.2768)</td>
<td>0.2504</td>
<td>0.0099</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.1815</td>
<td>0.0037</td>
<td>(0.1510, 0.2157)</td>
<td>0.1779</td>
<td>0.0095</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>0.1506</td>
<td>0.0038</td>
<td>(0.1187, 0.1870)</td>
<td>0.1518</td>
<td>0.0093</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>0.0922</td>
<td>0.0027</td>
<td>(0.0717, 0.1200)</td>
<td>0.0825</td>
<td>0.0061</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>0.0215</td>
<td>0.0014</td>
<td>(0.0098, 0.0366)</td>
<td>0.0242</td>
<td>0.0034</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>0.8826</td>
<td>0.0152</td>
<td>(0.7487, 1.0302)</td>
<td>0.8605</td>
<td>0.0400</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>1.2340</td>
<td>0.0365</td>
<td>(0.9407, 1.6135)</td>
<td>1.0979</td>
<td>0.0605</td>
</tr>
</tbody>
</table>

Table 5 contains the values of the estimates of contributions for the $i^{th}$ cohabitation duration group $\beta_i$ and the $j^{th}$ educational level ($\alpha_j$) together with their bootstrap standard errors from the reduced and entire data models.

Generally there exist little differences between the estimated values of $\beta_i$ between the two different models. However, the estimate of $\beta_i$ decreases with the cohabitation duration increasing. It is interesting that when the cohabitation duration changes from 2 to 3, the estimate of $\beta_i$ decreases slightly. However, when the other cohabitation durations change, the $\beta_i$ estimate decreases more quickly but steadily. Since we have classified the cohabitation duration by exposure time into 5 intervals: $0^{-2^{-}}, 2^{-4^{-}}, 4^{-6^{-}}, 6^{-10^{-}}$, and 10 years and over, it is not surprising that the results are like this. With the growth in the number of years for the individuals who have already lived together, they may feel that it makes no sense whether they marry or not. If they did not marry in the first years, they may be accustomed to their current life and may not have a passion to change their status. Thus the contribution of the cohabitation duration to the tendency of marriage becomes gradually smaller and smaller when the cohabitation time is larger and larger.

The estimate of $\alpha_j$, however, varies a lot between the two different models, especially in the third educational level. Since the contribution of the first-level $\alpha_1 = 1$, then the contribution of the tendency to marry in the second-level is less than in the first-level and third-level. The estimate of $\alpha_j$ in the third-level is the largest among the three $\alpha_j$ estimates. It is in accord
with the marriage percentage variation that we obtained in Table 3. Comparing the two data models, the corresponding estimated values of \( \alpha_j \) using the entire data are smaller than that using the reduced data. The possible reason is due to the assumption that all the individuals attained their highest educational level prior to cohabitation. Then we may underestimate the parameters since there exist some individuals who attained their highest educational level later than the cohabitation time.

From the bootstrap standard errors, we may see that all the standard errors are not big so our estimates are reasonable. However, the standard errors in the entire data model are all larger than that in the reduced model. Since the assumption is used in the entire data model, there must exist bigger deviation than in the reduced data model. All the estimates are included in the 95% confidence intervals. The length of the confidence interval in the entire data model is larger than that in the reduced model. From Figure a-1 and a-2 in Appendix A, we may see the histograms of these estimates after the bootstrap procedure and the estimated values are distributed mainly around the point estimate after resampling many times.

To understand the estimates clearly, we may select a specific group to interpret the meaning of the estimate. For instance, we select the individuals belonged to the cohabitation duration 1 of the second educational level in the reduced data model. From Table 5 we can see that the value of \( \beta_1 \) is equal to 0.2410, and the value of \( \alpha_2 \) is equal to 0.8826. Since the rate parameter \( \lambda_{ij} = \beta_i \alpha_j \), we may calculate the rate parameter based on this group. That is, \( \lambda_{12} = 0.2410 \times 0.8826 = 0.2127 \). Then we may say that for the individuals in the cohabitation duration 1 (exposure years from 0 to 2) of the second educational level, the rate of marriage is 0.2127 in the group.

5 Conclusion and Discussion

In this paper, the maximum likelihood procedure is proposed to estimate the effect parameters based on a multiplicative two-factor hazard model. And then the bootstrap approach is applied to estimate the standard errors of the parameter. Finally, we get a conclusion that both the educational level and the cohabitation duration have effects on the tendency of marriage. There exist little differences between the estimated values of the cohabitation duration contribution according to the two different models and the
estimated values decrease when the cohabitation time increases. Speaking of the contribution of the different educational level, the third educational level contributes to the tendency of marriage the most while the second-level the least. The two factors are both contributions to the tendency to marry. When we use the entire data, there exists bigger deviation than in the reduced data model.

As Ghilagaber & Koskinen (2009) and Ghilagaber & Larsson (2009) state in their paper, there still exist some problems needed to be solved and discussed. One specific problem is that we do not know the educational level at time for the cohabitation or marriage. We only know the time at which the highest educational level was attained for the individuals at the date of survey. And for some individuals, this may occur before or after the time of cohabitation. The highest educational level is reported by the date of the survey, for the individuals who acquire the highest educational level after the cohabitation, they may have lower educational level at the time of cohabitation. We have no idea about the exact educational level at the time of cohabitation. It will cause biases to the estimated parameters and we do not know the extent of deviation. The solution of trying to take into account that some of the individuals had a lower level of education than the highest one reported at the time of cohabitation need to be studied in further research.

References


[3] Ghilagaber, G. and Larsson, R. (2009), Maximum Likelihood Adjustment of Anticipatory Covariates in Analyzing Retrospective Survey Data. Stockholm University, Uppsala University, Gebre@stat.su.se, Rolf.Larsson@dis.uu.se.


### APPENDIX

#### A Omitted empirical results

Table a-1 The number of occurrences $D_{ij}$ in different cohabitation duration and different educational level

<table>
<thead>
<tr>
<th>Cohabitation duration</th>
<th>Highest educational level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First-level</td>
</tr>
<tr>
<td>Reduced data</td>
<td></td>
</tr>
<tr>
<td>Cohabitation duration 1</td>
<td>165</td>
</tr>
<tr>
<td>Cohabitation duration 2</td>
<td>70</td>
</tr>
<tr>
<td>Cohabitation duration 3</td>
<td>41</td>
</tr>
<tr>
<td>Cohabitation duration 4</td>
<td>20</td>
</tr>
<tr>
<td>Cohabitation duration 5</td>
<td>6</td>
</tr>
<tr>
<td>Entire data</td>
<td></td>
</tr>
<tr>
<td>Cohabitation duration 1</td>
<td>170</td>
</tr>
<tr>
<td>Cohabitation duration 2</td>
<td>73</td>
</tr>
<tr>
<td>Cohabitation duration 3</td>
<td>42</td>
</tr>
<tr>
<td>Cohabitation duration 4</td>
<td>20</td>
</tr>
<tr>
<td>Cohabitation duration 5</td>
<td>7</td>
</tr>
</tbody>
</table>

Table a-2 The exposure years $T_{ij}$ in different cohabitation duration and different educational level

<table>
<thead>
<tr>
<th>Cohabitation duration</th>
<th>Highest educational level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First-level</td>
</tr>
<tr>
<td>Reduced data</td>
<td></td>
</tr>
<tr>
<td>Cohabitation duration 1</td>
<td>629.93</td>
</tr>
<tr>
<td>Cohabitation duration 2</td>
<td>394.43</td>
</tr>
<tr>
<td>Cohabitation duration 3</td>
<td>263.90</td>
</tr>
<tr>
<td>Cohabitation duration 4</td>
<td>335.49</td>
</tr>
<tr>
<td>Cohabitation duration 5</td>
<td>365.51</td>
</tr>
<tr>
<td>Entire data</td>
<td></td>
</tr>
<tr>
<td>Cohabitation duration 1</td>
<td>647.60</td>
</tr>
<tr>
<td>Cohabitation duration 2</td>
<td>402.77</td>
</tr>
<tr>
<td>Cohabitation duration 3</td>
<td>269.32</td>
</tr>
<tr>
<td>Cohabitation duration 4</td>
<td>343.49</td>
</tr>
<tr>
<td>Cohabitation duration 5</td>
<td>370.51</td>
</tr>
</tbody>
</table>
Figure a-1 The histogram of the estimates using bootstrap approach from the reduced data.
**Figure a-2** The histogram of the estimates using bootstrap approach from the entire data.

**B R code**

```r
#The columns represent the following variables:
#1. CohabDur: Duration of Cohabitation (Grouped in 5 classes)
#2. Educ: Educational level (three levels)
#3. Status: Status indicator (event = 1, censored = 0)
#4. AgeatCohab: Age at Cohabitation
#5. AgeatEduc: Age at Completion of the reported educational level
#6. Exposure: Exposure months (it is these exposure months that are grouped into 5 classes in CohabDur)
#7. Anticip: Indicator of whether the Education-variable is anticipatory
# read data
data<-read.table('CohabReduced.txt',head=TRUE)
# empirical plot
plot(data$AgeatCohab,data$AgeatEduc,type='n',xlab='Age at time of attainment of highest educational level',
ylab='Age at Cohabitation',xlim=(10,50),ylim=(10,50))
points(data$AgeatEduc[data$Educ==1],data$AgeatCohab[data$Educ==1],pch=21,col='2')
points(data$AgeatEduc[data$Educ==2],data$AgeatCohab[data$Educ==2],pch=20,col='3')
points(data$AgeatEduc[data$Educ==3],data$AgeatCohab[data$Educ==3],pch=22,col='4')
abline(coef=c(0,1),col='6')
legend(30,50,c('first-level','second-level','third-level','reference'),text.col = "green4",
pch=c(21,20,22,NA),lty=c(0,0,0,1),col=c(2,3,4,6))
##Summary of data structure for the sample of 1365 Swedish men across the anticipatory status of education
J<-3
Non_A<matrix(c(rep(0,6)),nrow=J,ncol=2)
```
for (j in 1:J)
  {Non_A[j,1]<-sum(data$Status[data$Anticip==0&data$Educ==j])
   Non_A[j,2]<-length(data$Status[data$Anticip==0&data$Educ==j])-Non_A[j,1]
  }
Non_A

A<-matrix(c(rep(0,6)),nrow=J,ncol=2)
for (j in 1:J)
  {A[j,1]<-sum(data$Status[data$Anticip==1&data$Educ==j])
  }
A

## MLE function
MLE<-function(data)
{
  J<-3 #education level j=1,2,3
  I<-5 #Duration of Cohabitation
  k<-length(data[,1]) #number of individuals
  D=matrix(ncol=J,nrow=I)
  T=matrix(ncol=J,nrow=I)
  Educ<-data[,2]
  CohabDur<-data[,1]
  i<-1
  while(i<=I)
    {
      j<-1 while(j<=J)
      {D[i,j]<-sum(data[,3][data[,1]==i&data[,2]==j])
       T[i,j]<-0 # set T as a all zero matrix
       j<-j+1}
      i<-i+1}
  t<-data[,6]
  Educ<-data[,2] #E<-numeric()
  for (s in 1:k)
    {
      j<-Educ[s]
      T[1,j]<-T[1,j]+min(t[s],2)
      if (t[s]>=2) {T[2,j]<-T[2,j]+min(t[s]-2,2)}
      if (t[s]>=4) {T[3,j]<-T[3,j]+min(t[s]-4,2)}
      if (t[s]>=6) {T[4,j]<-T[4,j]+min(t[s]-6,4)}
      if (t[s]>=10) {T[5,j]<-T[5,j]+t[s]-10}
    }
  #construct D_r and T_r, for i=1,...,5, j=2,3
  D_r<-as.matrix(numeric(I))
  D_c<-as.matrix(numeric(J))
  for (i in 1:I) {D_r[i]<-sum(D[,i])
  for (i in 1:J) {D_c[i]<-sum(D[,i])
  T_r<-as.matrix(numeric(I))
Conclusions and Discussion

T_c<-as.matrix(numeric(J))
for (i in 1:I){ T_r[i]<-sum(T[,i])}
for (i in 1:J){ T_c[i]<-sum(T[,i])}
# Iterative method to estimate
n=100
b1-c-b2-c-b3-c-b4-c-b5=c(numeric(n+1))
a2-c-a3-c-numeric(n+1)

beta=rbind(b1,b2,b3,b4,b5)
alpha=rbind(a2,a3)

alpha[,1]<-c(1,1)
alpha[,1]<-as.matrix(alpha[,1])
beta[,1]<-D_r/T_r
a<0.000000001
s<-1
temp<-numeric()
while (s<n)
  for (i in 1:J)
    for (j in 1:I)
      temp[i]<-T[i,j]*beta[i,s]
    alpha[j,s+1]<-D_c[j]/sum(temp)
  temp<-numeric()
for (i in 1:J)
  for (j in 1:I)
    temp[j]<-T[i,j]*alpha[j,s+1]
  beta[i,s+1]<-D_r[i]/(sum(temp)+T[i,1])
if (c(alpha[1,s+1]-alpha[1,s])<=a && c(alpha[2,s+1]-alpha[2,s])<=a && c((beta[1,s+1]-beta[1,s]))<=a && c((beta[2,s+1]-beta[2,s]))<=a && c((beta[3,s+1]-beta[3,s]))<=a && c((beta[4,s+1]-beta[4,s]))<=a && c((beta[5,s+1]-beta[5,s]))<=a)
  {cat('Iterative number','
  cat(s+1,'
  cat('Alpha','
  t<-alpha[,s+1]
  #cat('Beta','
  t2<-beta[,s+1]
  s<-s+1 }
MLE<-c(t,t2)
MLE

## Bootstrap confidence interval function
boot.ci<-function(data,p)
  {n<-length(data)
data<-sort(data)
a<-data[round(n*(1-p)/2)+1]
b<-data[round(n*(1+p)/2)+1]
c(a,b)
  }

#### Reduced data

data<-read.table('CohabReduced.txt',head=TRUE)
CohabDur<-data$CohabDur[data$Anticip==0]
Educ<-data$Educ[data$Anticip==0]
Status<-data$Status[data$Anticip==0]
Exposure<-data$Exposure[data$Anticip==0]
Conclusion and Discussion

AgeatCohab <- data$AgeatCohab[data$Anticip == 0]
AgeatEduc <- data$AgeatEduc[data$Anticip == 0]
data <- data.frame(cbind(CohabDur, Educ, Status, AgeatCohab, AgeatEduc, Exposure))

# Bootstrap approach
source('MLE.R')
source('boot.ci.R')
B = 1000 # Bootstrap number
b <- matrix(nrow = B, ncol = 7)
k <- length(data[,1])
for (s in 1:B)
{
c <- sort(sample(1:k, 1000, replace = T)) # sample number = 1000
data <- read.table('CohabReduced.txt', head = TRUE)
### data <- data.frame(cbind(CohabDur, Educ, Status, AgeatCohab, AgeatEduc, Exposure)) ###
data <- data[c,]
b[s,] <- MLE(data)
}

# Estimate error
error <- c(sqrt(var(b[,1])), sqrt(var(b[,2])), sqrt(var(b[,3])), sqrt(var(b[,4])), sqrt(var(b[,5])), sqrt(var(b[,6])), sqrt(var(b[,7])))
error

# Confidence intervals
p <- 0.95
.ci <- c(boot.ci(b[,1], p), boot.ci(b[,2], p), boot.ci(b[,3], p), boot.ci(b[,4], p), boot.ci(b[,5], p), boot.ci(b[,6], p), boot.ci(b[,7], p))
.ci

# Histogram
par(mfrow = c(3, 3))
hist(b[,1], breaks = 30, xlab = "Alfa2", main = "Histogram of Alfa2")
hist(b[,2], breaks = 30, xlab = "Alfa3", main = "Histogram of Alfa3")
hist(b[,3], breaks = 30, xlab = "Beta1", main = "Histogram of Beta1")
hist(b[,4], breaks = 30, xlab = "Beta2", main = "Histogram of Beta2")
hist(b[,5], breaks = 30, xlab = "Beta3", main = "Histogram of Beta3")
hist(b[,6], breaks = 30, xlab = "Beta4", main = "Histogram of Beta4")
hist(b[,7], breaks = 30, xlab = "Beta5", main = "Histogram of Beta5")