Stockholms universitet

Research Report

Department of Statistics

No. 2019:1

Maximum Likelihood Adjustment of Anticipatory Covariates in the Analysis of Retrospective Data

Gebrenegus Ghilagaber
Rolf Larsson

Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden

Maximum Likelihood Adjustment of Anticipatory Covariates in the Analysis of Retrospective Data

Gebrenegus Ghilagaber[a,*] and Rolf Larsson[b]

[a]Department of Statistics, Stockholm University, Stockholm, Sweden
[b]Department of Mathematics, Uppsala University, Uppsala, Sweden

Abstract

A multiplicative hazard model in the presence of anticipatory covariates is estimated by maximum likelihood. The case study concerns the effects of educational level on risks of divorce. For individuals with anticipatory educational levels, conditional probabilities of having attained the reported level before marriage are used as weights in the likelihood. The adjusted estimates of relative risks do not differ significantly from those from anticipatory analysis.

Keywords: Anticipatory analysis; Event-history analysis; Expected likelihood analysis; Maximum likelihood; Retrospective surveys

1 Introduction

Consider a retrospective survey where the interest is to investigate differentials in the risk of divorce across educational levels attained before marriage, but where available information is only respondents' highest educational level at the time of the survey. Common practice is to use the available information on educational level as a covariate in modelling the risk of divorce, an

*Contact: Gebrenegus Ghilagaber, Department of Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. Gebre@stat.su.se

event that took place before the survey. Educational progress is likely to occur between the time of entry into marriage and the date of the survey. To what extent can changes in patterns of divorce across educational levels be attributed to changes in the distribution of respondents across the various levels of education? To what extent do they reflect real differences in divorce due to differences in educational level? These questions can be answered by dealing with the fact that the covariate (education) is anticipatory and adjusting the corresponding parameters to correct the bias inherent in the time inconsistency of the anticipatory covariate.

Hoem (1996) warns that using anticipatory covariates is misleading, but concludes that adverse effects may be smaller in some situations. In their study of mortality clustering in India using past births and deaths, Arulampalam and Bhalotra (2003; 2006) discard anticipatory regressors: household asset, toilet facility, electricity or access to piped water at the date of the survey. However, much valuable information may be lost by ignoring such covariates. Hoem and Kreyenfeld (2006a; 2006b) argue that anticipatory covariates may provide useful information. They propose data imputation, but this procedure requires unrealistic assumptions. Faucett et al. (1998) dealt with missing data with Bayesian techniques. They give interval estimates with higher coverage probabilities compared to imputation. Todesco (2011) uses anticipatory covariates, area of residence, education, and religious commitment, to analyze marital dissolution in Italy, but argues that this should not jeopardize his results. Ghilagaber and Koskinen (2009) in a Bayesian model found that anticipatory analysis can lead to overestimate the relative risks associated with the anticipatory covariate.

We use a maximum likelihood to explore adverse effects of anticipatory covariates. We model divorce risks among 1312 Swedish men born between 1936 and 1964 in a piecewise constant hazard model. For individuals with anticipatory educational levels, we compute conditional probabilities that these levels were attained before marriage. These probabilities are then used as weights to the anticipatory contribution to the likelihood. The estimates of relative risks do not differ significantly from those obtained in the anticipatory analysis based on unweighted likelihood. The sign of the estimates is the same as in the Bayesian analysis of the same data set of Ghilagaber and Koskinen (2009), but the estimate was significant in the previous study.

2 The Multiplicative Two-factor Hazard Model

For a sample of individuals, consider J educational levels and the total number D_{ij} of divorces at marriage-duration i, $i = 1, ..., I$ in the j^{th} educational level, $j = 1, 2, ..., J$ for T_{ij} years of observed exposure to the risk of divorce. The covariate indexed by i is the grouped-time variable (duration of marriage) measured from the date of marriage until the date of divorce or until the interview date, whichever comes first.

Define

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

$$D_{++} = \sum_{i=1}^{I} D_{i+} = \sum_{j=1}^{J} D_{+j} = \sum_{i=1}^{I} \sum_{j=1}^{J} D_{ij}. \tag{2}$$

T_{i+}, T_{+j}, and T_{++} represent similar quantities for the exposure variable T. Divorce risks are assumed to be piecewise constant in each of the the I time intervals but may vary between intervals. The time to divorce then follows a piecewise exponential distribution for each educational level. The density function of the time to divorce in duration group i for a person k with educational level j is

$$f(t_{ijk}) = \lambda_{ij} \exp\left(-\lambda_{ij} t_{ijk}\right). \tag{3}$$

A multiplicative model for the hazard rate λ_{ij} (Breslow and Day, 1975; Hoem, 1987) corresponds to

$$\lambda_{ij} = \beta_i \alpha_j, \tag{4}$$

where β_i characterizes the group i and α_j the j^{th} level of education, α_j.

The model in Eq. (4) has $I+J$ parameters $\beta_1, \beta_2, ..., \beta_I$, and $\alpha_1, \alpha_3, ..., \alpha_J$. α_j measures the relative risk of divorce for individuals with educational level j, relative to those of the baseline level, say level 1 (where α_1 is set to 1); while β_i is the risk of divorce at duration-group i in the baseline educational group ($j = 1$).

To construct the likelihood function when Eq. (4) holds true we denote δ_{ijk} an indicator variable of whether the k^{th} sample member having the j^{th} level of education is divorced ($\delta_{ijk} = 1$) or is still married ($\delta_{ijk} = 0$) in

the i^{th} duration of marriage. From Eq. (3) and (4), the contribution to the likelihood of the sub-sample of individuals in the i^{th} duration-group and having the j^{th} level of education n_{ij} are obtained as

$$\Lambda_{ij} = \prod_{k=1}^{n_{ij}} (\beta_i \alpha_j)^{\delta_{ijk}} \exp(-\beta_i \alpha_j t_{ijk}) = (\beta_i \alpha_j)^{D_{ij}} \exp(-\beta_i \alpha_j T_{ij}), \quad (5)$$

where

$$D_{ij} = \sum_{k=1}^{n_{ij}} D_{ijk} \quad \text{and} \quad T_{ij} = \sum_{k=1}^{n_{ij}} t_{ijk}.$$

The likelihood for the entire sample is the product of the Λ_{ij} over all levels of i and j:

$$\Lambda = \prod_{i=1}^{I} \prod_{j=1}^{J} \Lambda_{ij} = \prod_{i=1}^{I} \prod_{j=1}^{J} (\beta_i \alpha_j)^{D_{ij}} \exp(-\beta_i \alpha_j T_{ij}), \quad (6)$$

so that

$$\begin{aligned} \ln \Lambda &= \sum_{i=1}^{I} \sum_{j=1}^{J} D_{ij} \ln(\beta_i \alpha_j) - \sum_{i=1}^{I} \sum_{j=1}^{J} (\beta_i \alpha_j) T_{ij} \quad (7) \\ &= \sum_{i=1}^{I} \sum_{j=1}^{J} D_{ij} \ln(\beta_i) + \sum_{i=1}^{I} \sum_{j=1}^{J} D_{ij} \ln(\alpha_j) - \sum_{i=1}^{I} \sum_{j=1}^{J} (\beta_i \alpha_j) T_{ij} \\ &= \sum_{i=1}^{I} D_{i+} \ln(\beta_i) + \sum_{j=1}^{J} D_{+j} \ln(\alpha_j) - \sum_{i=1}^{I} \sum_{j=1}^{J} (\beta_i \alpha_j) T_{ij}. \end{aligned}$$

Differentiating $\ln(\Lambda)$ in Eq. (7) with respect to β_i on the one hand and with respect to α_j on the other hand we get

$$\beta_i^* = \frac{D_{i+}}{\sum_{j=1}^{J} \alpha_j^* T_{ij}}, \ i = 1, 2, ..., I \quad (8)$$

and

$$\alpha_j^* = \frac{D_{+j}}{\sum_{i=1}^{I} \beta_i^* T_{ij}},\ j = 1, ..., J. \tag{9}$$

This system of $I + J$ equations has no analytical solution in general, but can be solved numerically.

3 Adjusting for Anticipatory Covariates

3.1 Expected Likelihood

From Eq. (8) and (9), the maximum likelihood estimates of the baseline hazards $\widehat{\beta}_i$ and relative hazards $\widehat{\alpha}_j$ are functions of the total number D_{ij} of events and exposure times T_{ij}. Misclassification of events or exposure times into wrong intervals or into wrong levels of the covariate as with anticipatory covariates may lead to incorrect estimates.

Consider those individuals who have completed their reported highest educational level after marriage. Inference is based on the individuals' education levels at the date of marriage, but only their highest level of education at the date of interview is observed. This is an incomplete data problem, which is handled by maximizing the expected likelihood conditional on available information (Orchard and Woodbury, 1972).

Eq. (3) denotes the density conditional on the level of education $x_k(T_k)$ for individual k at the age of marriage T_k. Because the reported educational level j is a function $j = j(k)$, the unconditional density is

$$g(t_{ijk}) = f(t_{ijk}) \left(P x_k(T_k) = j(k)\right). \tag{10}$$

We also impose the distribution, $P\{x_k(T_k) = j(k)\}$ which adds a term to the log likelihood:

$$\ln\left(\widetilde{\Lambda}\right) = \ln(\Lambda) + \sum_{k=1}^{n_{ij}} \ln\left(P\left(x_k(T_k) = j(k)\right)\right), \tag{11}$$

where $\ln(\Lambda)$ is as in Eq. (7), and the last term in Eq. (11) will yield the maximum likelihood estimates of the education time distribution.

3.2 Parameter Estimation in the Expected Likelihood

To calculate the components of the last term $\sum_{k=1}^{n_{ij}} \ln\left(P\left(x_k\left(T_k\right)=j\left(k\right)\right)\right)$ in Eq. (11) assume that $J=3$ and, as in Ghilagaber and Koskinen (2009), S_{jk} denotes the time of transition from educational level $j-1$ to educational level j for individual k; f_j denotes the density function of S_{jk}, and F_j its distribution function. We introduce Bernoulli variables Z_j of parameters ϕ_j as indicators of whether or not the level $x_k\left(T_k\right)=j$ is the highest educational level. Then, ruling out the possibility that $x_k\left(T_k\right)=0$, we get proposition 1:

Proposition 1 Writing $p_{jk} \equiv \left(Px_k\left(T_k\right)=j\right)$,

$$\begin{aligned} p_{1k} &= F_1\left(T_k\right) - \left(1-\phi_1\right)\int_0^{T_k} f_1\left(u\right) F_2\left(T_k-u\right) du, \qquad (12)\\ p_{2k} &= \int_0^{T_k} f_1\left(u\right) F_2\left(T_k-u\right) du \\ &\quad - \left(1-\phi_2\right)\int_{u=0}^{T_k} f_1\left(u\right)\int_{v=0}^{T_k-u} f_2\left(v\right) F_3\left(T_k-u-v\right) dvdu, \\ p_{3k} &= \int_{u=0}^{T_k} f_1\left(u\right)\int_{v=0}^{T_k-u} f_2\left(v\right) F_3\left(T_k-u-v\right) dvdu. \end{aligned}$$

Proof See appendix B.1.

To calculate all $P\left(x_k\left(T_k\right)=j\right)$ explicitly, we impose distributions on the times, S_{jk}, spent at the different educational levels. As in Ghilagaber and Koskinen (2009), we consider a piecewise gamma distribution with density function

$$f_j\left(s\right) = \frac{\eta_j^{\zeta_j} s^{\zeta_j-1}}{\Gamma\left(\zeta_j\right)} \exp\left(-\eta_j s\right), \qquad (13)$$

and distribution function

$$F_j\left(s\right) = \Gamma\left(\zeta_j, \eta_j s\right), \qquad (14)$$

where

$$\Gamma\left(a,s\right) = \int_0^s \frac{u^{a-1}}{\Gamma\left(a\right)} \exp\left(-u\right) du \qquad (15)$$

is the incomplete gamma function.

Writing $E_*(\cdot) = E(\cdot|y_1, ..., y_K)$, Eqs. (7) and (11) lead to

$$\begin{aligned} & E_*\left(\ln\left(\widetilde{\Lambda}\right)\right) \\ & = \sum_{i=1}^{I} E_*(D_{i+}) \ln(\beta_i) + \sum_{j=1}^{J} E_*(D_{+j}) \ln(\alpha_j) - \sum_{i=1}^{I}\sum_{j=1}^{J} E_*(T_{ij}) \beta_i \alpha_j \\ & + \sum_{k} E_*(\ln(P(x_k(T_k) = j(k)))). \end{aligned} \tag{16}$$

To make inference on the parameters β_i and α_j, we need to calculate the conditional expectations of the sufficient statistics D_{i+}, D_{+j} and T_{ij}. We re-write

$$\begin{aligned} D_{i+} &= \sum_{k=1}^{n_{ij}} D_{ijk} I_{(k\in\mathcal{A}_i)}, \\ D_{+j} &= \sum_{k=1}^{n_{ij}} D_{ijk} I_{(k\in\mathcal{B}_j)}, \\ T_{ij} &= \sum_{k=1}^{n_{ij}} t_{ijk} I_{(k\in\mathcal{A}_i\cap\mathcal{B}_j)} \\ &= \sum_{k=1}^{n_{ij}} \sum_{l\le i} \min(t_{ljk} - m(l), m(l+1) - m(l)) I_{(k\in\mathcal{A}_i\cap\mathcal{B}_j)}, \end{aligned} \tag{17}$$

where $\mathcal{A}_i$ is the set of individuals with common first index i, $\mathcal{B}_j$ is the set of individuals with common second index j, $m(i)$ is the lower duration limit in group i, and $I(A)$ is the indicator function of the event A. For each individual k, we observe only one marriage duration group $i = i(k)$ and one educational level $j = j(k)$. However, if that educational level is completed after marriage, the educational level at time of marriage is unknown. By introducing distributional assumptions on time to complete a certain educational level, the probabilities are calculated as

$$P(j(k) = j_0|y_k) = P(x_k(T_k) = j_0|x_k(t_k) = y_k), \tag{18}$$

where t_k is the age at completion of the reported level of education for individual k. This feature does not affect D_{i+}, since summation is over j.

However, it affects D_{+j} and T_{ij}, since different individuals may belong to different B_j depending on their unknown education level at time of marriage. Hence, we need to compute

$$E\left(D_{+j}|y_1,...,y_K\right)=\sum_k E\left(D_{ijk}I_{(k\in\mathcal{B}_j)}|y_k\right). \tag{19}$$

Re-writing the right-hand side as

$$E\left(D_{ijk}I_{\{k\in\mathcal{B}_j\}}|y_k\right)=D_{ijk}P\left(k\in\mathcal{B}_j|y_k\right)=D_{ijk}P\left(x_k\left(T_k\right)=j|x_k\left(t_k\right)=y_k\right), \tag{20}$$

gives

$$E\left(D_{+j}|y_1,...,y_K\right)=\sum_{k=1}^{n_{ij}}D_{ijk}P\left(x_k\left(T_k\right)=j|x_k\left(t_k\right)=y_k\right). \tag{21}$$

Similarly,

$$\begin{aligned}
&E\left(T_{ij}|y_1,...,y_K\right)\\
&=\sum_{k=1}^{n_{ij}}\sum_{l\leq i}E\left(\min\left(t_{ljk}-m\left(l\right),m\left(l+1\right)-m\left(l\right)\right)I_{(k\in\mathcal{A}_i\cap\mathcal{B}_j)}|y_k\right)\\
&=\sum_{k=1}^{n_{ij}}\sum_{l\leq i}\min\left(t_{ljk}-m\left(l\right),m\left(l+1\right)-m\left(l\right)\right)I_{(k\in\mathcal{A}_i)}P\left(x_k\left(T_k\right)=j|x_k\left(t_k\right)=y_k\right).
\end{aligned} \tag{22}$$

To maximize the adjusted log-likelihood in Eq. (11) over the parameters β_i and α_j, we first plug in the expression in Eq. (21) into Eq. (9) and the expression in Eq. (22) into Eqs. (8) and (9), and proceed in the usual manner.

To make inference on ζ_j, η_j and ϕ_j we first observe (see Eq. (11)) that

$$\begin{aligned}
&\sum_{k=1}^{n_{ij}}E\left(\ln P\left(x_k\left(T_k\right)=j\left(k\right)\right)|y_k\right)\\
&=\sum_{k=1}^{n_{ij}}\sum_{j=1}^{J}\ln P\left(x_k\left(T_k\right)=j\right)P\left(x_k\left(T_k\right)=j|x_k\left(t_k\right)=y_k\right).
\end{aligned} \tag{23}$$

Each set of ζ_j, η_j and ϕ_j values produces a numerical value of Eq. (23) as well as expected values of the sufficient statistics in Eq. (21) and Eq. (22),

which in turn are used to maximize $\ln(\Lambda)$ in Eq. (11). We calculate, this way, the expected log likelihood for any set of ζ_j, η_j and ϕ_j, maximized with respect to the α_j and β_i parameters. The next step is to maximize over ζ_j, η_j and ϕ_j, using the Newton-Raphson algorithm. According to Orchard and Woodbury (1972), such a procedure yields the ML estimates. We obtain the values of the $P(x_k(T_k) = j, x_k(t_k) = y)$ using proposition 2.

Proposition 2 Writing $p_{jyk} \equiv P(x_k(T_k) = j, x_k(t_k) = y)$, we get

$$\begin{aligned}
p_{11k} &= F_1(T_k) - (1-\phi_1)\int_0^{T_k} f_1(u) F_2(t_k - u)\, du, \qquad (24)\\
p_{12k} &= (1-\phi_1)\left\{\int_0^{T_k} f_1(u)\left[F_2(t_k-u) - F_2(T_k-u)\right] du\right.\\
&\quad \left. - (1-\phi_2)\int_{u=0}^{T_k} f_1(u)\int_{v=T_k-u}^{t_k-u} f_2(v) F_3(t_k-u-v)\, dvdu\right\},\\
p_{22k} &= \int_0^{T_k} f_1(u) F_2(T_k-u)\, du\\
&\quad - (1-\phi_2)\int_{u=0}^{T_k} f_1(u)\int_{v=0}^{T_k-u} f_2(v) F_3(t_k-u-v)\, dvdu,\\
p_{13k} &= p_{1k} - p_{11k} - p_{12k},\\
p_{23k} &= p_{2k} - p_{22k},\\
p_{33k} &= p_{3k}.
\end{aligned}$$

Proof See appendix B.2.

Then, we express the $P(x_k(t_k) = y)$ in terms of the probabilities in Eq. (24) to get

$$P(x_k(t_k) = 1) = P(x_k(T_k) = 1, x_k(t_k) = 1), \qquad (25)$$

$$P(x_k(t_k) = 2) = P(x_k(T_k) = 1, x_k(t_k) = 2) + P(x_k(T_k) = 2, x_k(t_k) = 2), \qquad (26)$$

and

$$\begin{aligned}
&P(x_k(t_k) = 3) \qquad (27)\\
&= P(x_k(T_k) = 1, x_k(t_k) = 3) + P(x_k(T_k) = 2, x_k(t_k) = 3)\\
&+ P(x_k(T_k) = 3, x_k(t_k) = 3).
\end{aligned}$$

4 Effect of Education on Divorce

We use an extract from the 1985 Mail Survey of Swedish men as our data for illustration. The survey contained background variables as well as retrospective information on entry into and exit from marital and non-marital unions. Analyses is based on 1312 ever-married men who were either divorced or still married by the survey time. Their distribution across age at marriage and age at attainment of the reported educational level is shown in Figure 1. Those below the diagonal are the 245 (19%) observations whose reported educational level was completed after they married. Anticipatory analysis, common in the analysis of such type of data, amounts to moving the values below the diagonal in Figure 1 to the left, all the way to the diagonal reference line. A cross tabulation of the sample as displayed in Table 1 shows differentials in percentage divorced across the anticipatory status of education. The main goal of our work is investigating the role of misclassification on such differentials in educational gradients of divorce.

4.1 Models

We categorize the time variable, duration of marriage in years, into five intervals: $0\text{–}1^-$, $1\text{–}2^-$, $2\text{–}3^-$, $3\text{–}6^-$, and 6^+ years. We set primary level of education as baseline level and, hence, its corresponding relative hazard, α_1, is set to 1. To make the proposed adjustment comparable to previous work on the same data set (Ghilagaber and Koskinen, 2009), we also estimate the parameters from the common anticipatory approach and from a reduced model. Using anticipatory analysis amounts to "back-dating" the times of highest educational achievement, τ_k, for a number of individuals while in the reduced model, observations whose reported educational level was completed after marriage, $T_k < t_k$, are discarded.

Neither the anticipatory manipulations nor the proposed adjustment change the observed marginal occurrences D_{i+} and exposures T_{i+}. In the reduced model, these marginals are reduced due to the reduction of the total number of respondents. The time at which the highest educational level is achieved is irrelevant as long as it precedes the time of marriage, but it is relevant for calculating exposure times whenever it occurs after the time of marriage.

The conditions required for using the Fisher information to construct confidence intervals may not be fulfilled due to the nature of the problem at hand. Instead, we emply the bootstrap method. We bootstrap the indi-

viduals B times and maximize the expected likelihood for each such sample. We then calculate empirical 95% confidence intervals for each parameter by computing the 2.5% and 97.5% fractiles of the bootstrap distributions.[1] In the reduced and anticipatory cases, the number of bootstrap replications were $B = 10000$. In the adjusted case, $B = 1000$ because estimation of the additional parameters ζ and η required heavy computation.[2]

4.2 Numerical Considerations

We used numerical integration to calculate the integrals for given parameter values. When maximizing the expected likelihood over the ϕ parameters, one may arise. All the $P\{x(T) = j\}$ are linear in the ϕ_i and, hence, for individuals with $t_k \leq T_k$, the derivatives of $\ln P\{x_k(T_k) = j(k)\}$ may be nonzero for all possible ϕ_i. Hence, if all individuals would have $t_k \leq T_k$, and, thus, conditioning would be unnecessary, the likelihood may be maximized at the boundary of the parameter space, $\phi_i = 0$ or $\phi_i = 1$ for $i = 1$ and 2. These solutions might be considered unrealistic. For individuals with $T_k < t_k$, the situation is less clear-cut, but if majority of the individuals have $t_k \leq T_k$, there is a risk that their contribution dominates the likelihood, and this seems to be the case in the data set used for our illustration.

Hence, we fixed the ϕ_i parameters to the empirical proportions of men who did not continue to the next higher level (see Table 1). We assigned $\phi_1 = 0.34$, obtained from Table 1 as $\frac{442}{1312}$, and $\phi_2 = 0.66$, obtained from Table 1 as $\frac{488+94}{1312-442}$. These are also almost identical to the Bayesian estimates obtained in Ghilagaber and Koskinen (2009).

Instead of considering the gamma distribution parameters ζ_j, η_j explicitly in the maximization, it was more convenient to use the transformed parameters $\mu_j = \frac{\zeta_j}{\eta_j}$ and $\sigma_j = \frac{\zeta_j^{\frac{1}{2}}}{\eta_j}$. These correspond to the expectations and standard deviations, respectively. We found that the likelihood has an asymptote as $\sigma_3 \to 0$. To alleviate this problem, σ_3 was fixed and maximization was carried over the other parameters. This turned out to work out well.

[1] An alternative appraoch is to compute standard deviations and form normal approximation confidence intervals. But, this approach was considered problematic because of outliers in the bootstrap distributions.

[2] In the bootstrap replications, integrals and derivatives were calculated in the same manner as in the estimation. The estimated parameters were taken as starting values for each bootstrap replication.

Unlike in the ϕ parameter, there was no natural choice of ad hoc value of σ_3. However, the maximum with respect to all other parameters, including α_j and β_i, was very robust to different choices. We decided to put $\sigma_3 = 0.5$ which seemed to be a reasonable value.

To obtain the maximum of the expected likelihood, we combined several methods. Each likelihood calculation was done for a fixed set of the five gamma distribution parameters, μ_1, μ_2, μ_3, σ_1 and σ_2, including a maximization with respect to the α_j and β_i parameters. So in principle, the required task was to maximize over the gamma parameters. Initially, we performed a random choice of parameter values, and we selected the ones that gave the largest value of the likelihood. Then, using the Newton-Raphson maximization algorithm, we maximized the likelihood over the five gamma distribution parameters one at a time, keeping the others fixed. Finally, we maximized over μ_1, μ_2, μ_3, σ_1 and σ_2 simultaneously using the Newton-Raphson maximization algorithm in five dimensions.

To obtain confidence intervals for the parameters, we used a bootstrap procedure, sampling the individuals with replacement. We performed the maximization procedure described above each bootstrap replication. While bootstraping the μ_3 parameter drifted towards zero in about 6% of the bootstrap replications during the Newton-Raphson iterations. As negative value of μ_3 is not reasonable, we imposed an extra condition, $\mu_3 \geq 0.001$, to to overcome the problem. Further details on numerical issues, including Matlab codes, can be obtained from the authors upon request.

4.3 Results

4.3.1 Conditional Probabilities

Figures 2 displays the distribution of conditional probablities of various educational levels at marriage, given a corresponding reported educational level at time of survey, based on the 245 individuals who have compeleted their reported educational level after marriage. For comparison purposes, corresponding conditional probabilities computed by using estimates of the covariate-model parameters in a Bayesian analysis of the same data set (Ghilagaber and Koskinen, 2009) are also plotted.

The plots show that the probabilities of having had a lower educational level at marriage, given some higher level at interview, decrease with age at marriage. Thus, for someone who reports a post-secondary educational-level

at interview but has married early, say below age 20, the probability that he had primary-level education at marriage, $P_{1|3}$, is almost 1 (see Fig. 2). The probability that he had secondary level education at marriage, $P_{2|3}$, is also high but not as high as $P_{1|3}$. The combined probability that he had a lower level education, primary or secondary, is almost 1.

A comparison of the ML- and Bayes-estimates of the conditional probabilities indicates that the Bayes-estimates are, in general, higher than their corresponding ML-estimates. This is especially the case at older ages of marriage and for men who reported post-secondary level of education at time of survey.

4.3.2 Model parameters

Table 2 contains estimates of baseline risks, β_i expressed per 1000 exposure units, and relative risks of divorce, α_j, across the three models together with their corresponding 95% confidence intervals. Except for the effect of the second interval, β_2, which is much lower in reduced model, the estimates of the baseline risks, β_i, are close to each other across the three models.

The estimates of the relative risks α_j, on the other hand, vary appreciably across models. For instance, men with secondary level education have about the same risk of divorce as those with primary level education in the reduced model, 10% higher risk in the anticipatory model, and a negligible 5% higher risk in the adjusted model. Those with post-secondary education have much higher risks of divorce relative to the baseline men with primary education. The excess risk is 57% in the reduced model, 35% in the anticipatory model, and 34% in the adjusted model.

More interesting for the present purpose is the differences, or lack of it, in the estimates of relative risks across the models. That the anticipatory analyses lead to the same estimates of the relative risks α_j, 1.13 and 1.05 for α_2; 1.35 and 1.34 for α_3, indicates that anticipatory analysis is harmless in the sense that it does not lead to substantial bias in the estimates of the relative risks. These estimates of relative risks are somewhat higher than those obtained in the Bayesian adjustment by Ghilagaber and Koskinen (2009). This is especially true for the estimate of α_3 but it should be borne in mind that the 95% confidence intervals for α_2 and α_3 in both the ML and Bayes include 1 and, hence, are not significant at 5% significance level. Further, a re-estimation of α_3 using the Bayes estimated parameters of the gamma distribution yielded $\widehat{\alpha}_3 = 1.30$ which is almost identical to the $\widehat{\alpha}_3 =$

1.34 obtained through the Maximum likelihood approach.

The combined effects of the differences in the estimates of the model parameters, β_i and α_j, across the three models and for the three educational levels are depicted in Figures 3 and 4 which contain estimates, $\widehat{\lambda}_{ij} = \widehat{\beta}_i\widehat{\alpha}_j$ under various configurations. Figures 3(a) - 3(c) show educational profiles of divorce risks over time for each of the three models. In Figures 4(a) - 4(c), differences in the estimates of divorce risks across the three models are depicted for each educational level. Divorce risks increase over the first four time intervals, except in the reduced model which exhibits decrease in the risk between first and second interval, and decrease after about 6 years (Figure 3). Further, the levels and trends are alike across the anticipatory and adjusted models (Figure 4).

4.3.3 Covariate-model parameters

In the adjusted model, the parameters of the gamma-distribution for educational career, ζ_j and η_j, are also estimated in addition to the model parameters β_i and α_j. These are shown in the lower half of the column of the corresponding model in Table 2. These parameters are then used to compute estimates of the expected duration to complete the various educational levels as displayed in Table 3. Thus, it takes, on the average, 16.4 years after birth to complete primary-level education, 3.6 years to complete secondary level education after completing primary level, and 1.9 years to complete post-secondary level education after completing secondary level. The corresponding figures from the Bayesian analysis, extracted from the estimates in Ghilagaber and Koskinen (2009), were 14.8, 6.5 and 6.3. These figures differ much from those of the Maximum likelihood method. However, as can be seen in Table 5 where estimates of the model parameters α_j and β_i given the Bayesian estimates of the gamma-distribution parameters are presented, it is easy to observe, by comparing to Table 2, that the Maximum likelihood estimates of the model-parameters are very robust to this kind of changes in the gamma parameters.

In tables 3 and 4, in addition to estimates of means μ_j and standard deviations σ_j of education times, 80% bootstrap confidence intervals are also presented. The reason behind choosing 80% instead of 95% is due to problems in estimating μ_3 as already described in Section 4.3. No confidence interval is provided for σ_3 because it was fixed to 0.5. This might also have contributed towards the relatively narrow confidence intervals for ζ_3 and η_3 in Table 2.

5 Conclusion

Anticipatory analysis explaining current behavior by future outcomes in life-course research is problematic because it does not follow the temporal order of events. Event history data collected at enormous cost lack history on important explanatory variables such as education and often social status. It is the investigator's responsibility to explore the effects of anticipatory analysis and seek appropriate procedures to minimize, if not eliminate, the bias due to such errors in design before any attempt is made to interpret and use the parameters of interest. This issue was addressed and analytic procedure was proposed.

The specific problem has been that some individuals in the sample have achieved their reported highest educational level after marriage. At least some have had lower educational level at the time of marriage than at the time of the survey. There was little idea of how much lower it should be but there was information on the age of the individual and year at which he achieved the reported highest educational level. Using the education variable in its original form is likely to cause biases in the estimated relative hazards but the strength and direction of this bias is unclear. The main goal of the investigation was to come up with numerical estimates of the direction and strength of this bias.

We proposed maximum likelihood approach to use the available information optimally. We computed conditional probabilities that the reported educational levels were completed before marriage under certain distributional assumptions for educational career. We then used these probabilities as weights in the contributions of individuals to the likelihood function from which we derived the adjusted baseline and relative risks.

The results from our illustrative data show that anticipatory analysis is harmless because the adjusted estimates of relative risks are close to those from anticipatory analysis. Whether these results can be replicated on other data sets or on the same data set but with different events of interest, like family formation, through cohabitation or marriage, or childbearing, is an open question for future investigation.

References

[1] Arulampalam, W., and Bhalotra, S. (2003). Sibling death clustering in India: genuine scarring vs unobserved heterogeneity. *Discussion Paper* No. 03/552**,** Department of Economics, University of Bristol, UK.

[2] Arulampalam W, and Bhalotra, S. (2006). Sibling death clustering in India: state dependence versus unobserved heterogeneity *Journal of the Royal Statistical Society - Series A (Statistics in Society), 169*: 829-848.

[3] Breslow, N. E. and N. E. Day (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.

[4] Faucett, C. L., Schenker, N., and Elashoff, R. M. (1998). Analysis of Censored Survival Data with Intermittently Observed Time-Dependent Binary Covariates. *Journal of the American Statistical Association, 93*: 427 - 437.

[5] Ghilagaber, G., and Koskinen, J. H. (2009). Bayesian Adjustment of Anticipatory Covariates in Analyzing Retrospective Survey Data. *Mathematical Population Studies - An International Journal of Mathematical Demography, 16*: 105-130.

[6] 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.

[7] Hoem, J. M. (1996). The harmfulness and harmlessness of using anticipatory regressor. How dangerous is it to use education achieved as of 1990 in the analysis of divorce risks in earlier years. *Yearbook of Population Research in Finland, 33*: 34-43.

[8] Hoem, J. M., and Kreyenfeld, M. (2006a). Anticipatory Analysis and its Alternatives in Life-Course Research. Part 1: The Role of Education in the Study of First Childbearing. *Demographic Research*, Volume 15, Article 16, pp. 461 - 484 (29 November 2006). Available online at http://www.demographic-research.org/volumes/vol15/16/. (accessed 26 May 2007).

[9] Hoem, J. M., and Kreyenfeld, M. (2006b). Anticipatory Analysis and its Alternatives in Life-Course Research. Part 2: Two Interacting Processes. *Demographic Research*, Volume 15, Article 17, pp. 485 - 498 (29 November 2006). Available online at http://www.demographic-research.org/volumes/vol15/17/. (accessed 26 May 2007).

[10] Orchard, T. and Woodbury, M. A. (1972). A missing information principle: theory and applications. *Proceedings 6th Berkeley Symposium on Mathematical Statistics and Probability, 1*: 697-715.

[11] Todesco, L. (2011). A Matter of Number, Age or Marriage? Children and Marital Dissolution in Italy. *Population Research and Policy Review, 30*: 313-332.

APPENDIX

A Omitted proofs

A.1 Proof of proposition 1

We find (dropping the index k for ease of exposition)

$$\begin{aligned} p_1 &= P\{x(T)=1\} \\ &= (1-\phi_1)P\{x(T)=1|Z_1=0\} + \phi_1 P\{x(T)=1|Z_1=1\} \\ &= (1-\phi_1)P(S_1 \le T, S_1+S_2 > T) + \phi_1 P(S_1 \le T) \\ &= (1-\phi_1)\int_0^T f_1(u)\{1-F_2(T-u)\}du + \phi_1 F_1(T) \\ &= F_1(T) - (1-\phi_1)\int_0^T f_1(u)F_2(T-u)du, \end{aligned} \tag{28}$$

and similarly,

$$\begin{aligned} p_2 &= P\{x(T)=2\} \\ &= (1-\phi_2)P\{x(T)=2|Z_2=0\} + \phi_2 P\{x(T)=2|Z_2=1\} \\ &= (1-\phi_2)P(S_1+S_2 \le T, S_1+S_2+S_3 > T) + \phi_2 P(S_1+S_2 \le T) \\ &= (1-\phi_2)\int_{u=0}^T f_1(u)\int_{v=0}^{T-u} f_2(v)\{1-F_3(T-u-v)\}dvdu \\ &+ \phi_2\int_0^T f_1(u)F_2(T-u)du \\ &= \int_0^T f_1(u)F_2(T-u)du - (1-\phi_2)\int_{u=0}^T f_1(u)\int_{v=0}^{T-u} f_2(v)F_3(T-u-v)dvdu, \end{aligned} \tag{29}$$

while,

$$\begin{aligned} p_3 &= P\{x(T)=3\} \\ &= P(S_1+S_2+S_3 \le T) \\ &= \int_{u=0}^T f_1(u)\int_{v=0}^{T-u} f_2(v)F_3(T-u-v)dvdu. \end{aligned} \tag{30}$$

A.2 Proof of proposition 2

We get (dropping the index k for ease of exposition)

$$\begin{aligned} p_{11} &= P\left\{x\left(T\right)=1, x\left(t\right)=1\right\} \\ &= \left(1-\phi_1\right) P\left\{x\left(T\right)=1, x\left(t\right)=1 | Z_1=0\right\} \\ &+ \phi_1 P\left\{x\left(T\right)=1, x\left(t\right)=1 | Z_1=1\right\} \\ &= \left(1-\phi_1\right) P\left(S_1 \leq T, S_1+S_2>t\right)+\phi_1 P\left(S_1 \leq T\right) \\ &= \left(1-\phi_1\right) \int_0^T f_1\left(u\right)\left\{1-F_2\left(t-u\right)\right\} du+\phi_1 F_1\left(T\right) \\ &= F_1\left(T\right)-\left(1-\phi_1\right) \int_0^T f_1\left(u\right) F_2\left(t-u\right) du, \end{aligned} \tag{31}$$

and

$$\begin{aligned} p_{12} &= P\left\{x\left(T\right)=1, x\left(t\right)=2\right\} \\ &= \left(1-\phi_1\right)\left(1-\phi_2\right) P\left\{x\left(T\right)=1, x\left(t\right)=2 | Z_1=0, Z_2=0\right\} \\ &+ \left(1-\phi_1\right) \phi_2 P\left\{x\left(T\right)=1, x\left(t\right)=2 | Z_1=0, Z_2=1\right\} \\ &+ \phi_1 P\left\{x\left(T\right)=1, x\left(t\right)=2 | Z_1=1\right\} \\ &= \left(1-\phi_1\right)\left(1-\phi_2\right) P\left(S_1 \leq T,\ T<S_1+S_2 \leq t, S_1+S_2+S_3>t\right) \\ &+ \left(1-\phi_1\right) \phi_2 P\left(S_1 \leq T,\ T<S_1+S_2 \leq t\right)+0 \\ &= \left(1-\phi_1\right)\left(1-\phi_2\right) \int_{u=0}^T f_1\left(u\right) \int_{v=T-u}^{t-u} f_2\left(v\right)\left\{1-F_3\left(t-u-v\right)\right\} dvdu \\ &+ \left(1-\phi_1\right) \phi_2 \int_0^T f_1\left(u\right)\left[F_2\left(t-u\right)-F_2\left(T-u\right)\right] du \\ &= \left(1-\phi_1\right) \int_0^T f_1\left(u\right)\left[F_2\left(t-u\right)-F_2\left(T-u\right)\right] du \\ &- \left(1-\phi_1\right)\left(1-\phi_2\right) \int_{u=0}^T f_1\left(u\right) \int_{v=T-u}^{t-u} f_2\left(v\right) F_3\left(t-u-v\right) dvdu. \end{aligned} \tag{32}$$

Moreover,

$$
\begin{aligned}
p_{22} &= P\left\{x\left(T\right)=2,x\left(t\right)=2\right\} \\
&= \left(1-\phi_2\right)P\left\{x\left(T\right)=2,x\left(t\right)=2|Z_2=0\right\} \\
&+ \phi_2 P\left\{x\left(T\right)=2,x\left(t\right)=2|Z_2=1\right\} \\
&= \left(1-\phi_2\right)P\left(S_1+S_2\leq T,S_1+S_2+S_3>t\right) \\
&+ \phi_2 P\left(S_1+S_2\leq T\right) \\
&= \left(1-\phi_2\right)\int_{u=0}^{T} f_1\left(u\right)\int_{v=0}^{T-u} f_2\left(v\right)\left\{1-F_3\left(t-u-v\right)\right\}dvdu \\
&+ \phi_2\int_0^T f_1\left(u\right)F_2\left(T-u\right)du \\
&= \int_0^T f_1\left(u\right)F_2\left(T-u\right)du \\
&- \left(1-\phi_2\right)\int_{u=0}^{T} f_1\left(u\right)\int_{v=0}^{T-u} f_2\left(v\right)F_3\left(t-u-v\right)dvdu,
\end{aligned}
\tag{33}
$$

and the rest of the equalities are trivial.

B Tables and figures of empirical results

Table 1: Distribution of the sample of 1312 Swedish men across anticipatory status of education and status of marriage

		Status			
		Still married	Divorced	Total	% divorced
Non-Anticip	Primary	371	71	442	16
	Secondary	433	55	488	11
	Post Secon.	116	21	137	15
		920	147	1067	14
Anticipatory	Primary	-	-	-	-
	Secondary	66	28	94	30
	Post-Secon.	120	31	151	21
		186	59	245	24
	Total	1106	206	1312	16

Table 2: Estimated baseline and relative risks of divorce across the three models (95% confidence intervals in parentheses)

	Reduced	Anticip	Adjusted
β_1	7.2 (2.6, 13.3)	6.1 (2.5, 11.0)	6.4 (2.5, 11.1)
β_2	5.7 (1.7, 11.0)	10.1 (5.0, 15.9)	10.5 (5.5, 16.5)
β_3	13.1 (6.2, 21.7)	12.0 (6.2, 19.4)	12.5 (6.8, 20.2)
β_4	14.9 (10.0, 20.8)	14.9 (10.4, 20.3)	15.5 (10.8, 20.9)
β_5	11.9 (8.9, 15.2)	11.7 (8.8, 14.8)	12.1 (9.2, 15.3)
α_2	0.97 (0.67, 1.39)	1.13 (0.83, 1.57)	1.05 (0.78, 1.44)
α_3	1.57 (0.90, 2.51)	1.35 (0.93, 1.92)	1.34 (0.91, 1.92)
ζ_1	-	-	34900 (21600, 36900)
ζ_2	-	-	6.92 (3.97, 20.18)
ζ_3	-	-	14.50 (0, 66.31)
η_1	-	-	2130 (1260, 2360)
η_2	-	-	1.90 (0.97, 9.59)
η_3	-	-	7.62 (0, 16.29)

Table 3: Expected duration to complete various educational levels

		Estimate	80% confidence interval
Primary	$\frac{\zeta_1}{\eta_1}$	16.40	$(15.62, 16.99)$
Secondary	$\frac{\zeta_2}{\eta_2}$	3.65	$(2.89, 3.92)$
Post-secondary	$\frac{\zeta_3}{\eta_3}$	1.90	$(0.60, 3.09)$

Table 4: Standard deviation of the duration to complete various educational levels

		Estimate	80% confidence interval
Primary	$\frac{\zeta_1^{\frac{1}{2}}}{\eta_1}$	0.088	$(0.082, 0.095)$
Secondary	$\frac{\zeta_2^{\frac{1}{2}}}{\eta_2}$	1.40	$(0.95, 1.82)$

Table 5: ML estimates of model parameters given Bayesian estimated gamma parameters

β_1	6.4
β_2	10.5
β_3	12.6
β_4	15.6
β_5	12.2
α_2	1.07
α_3	1.30

Fig. 1: Distribution of the 245 anticipatory observations across reported educational levels

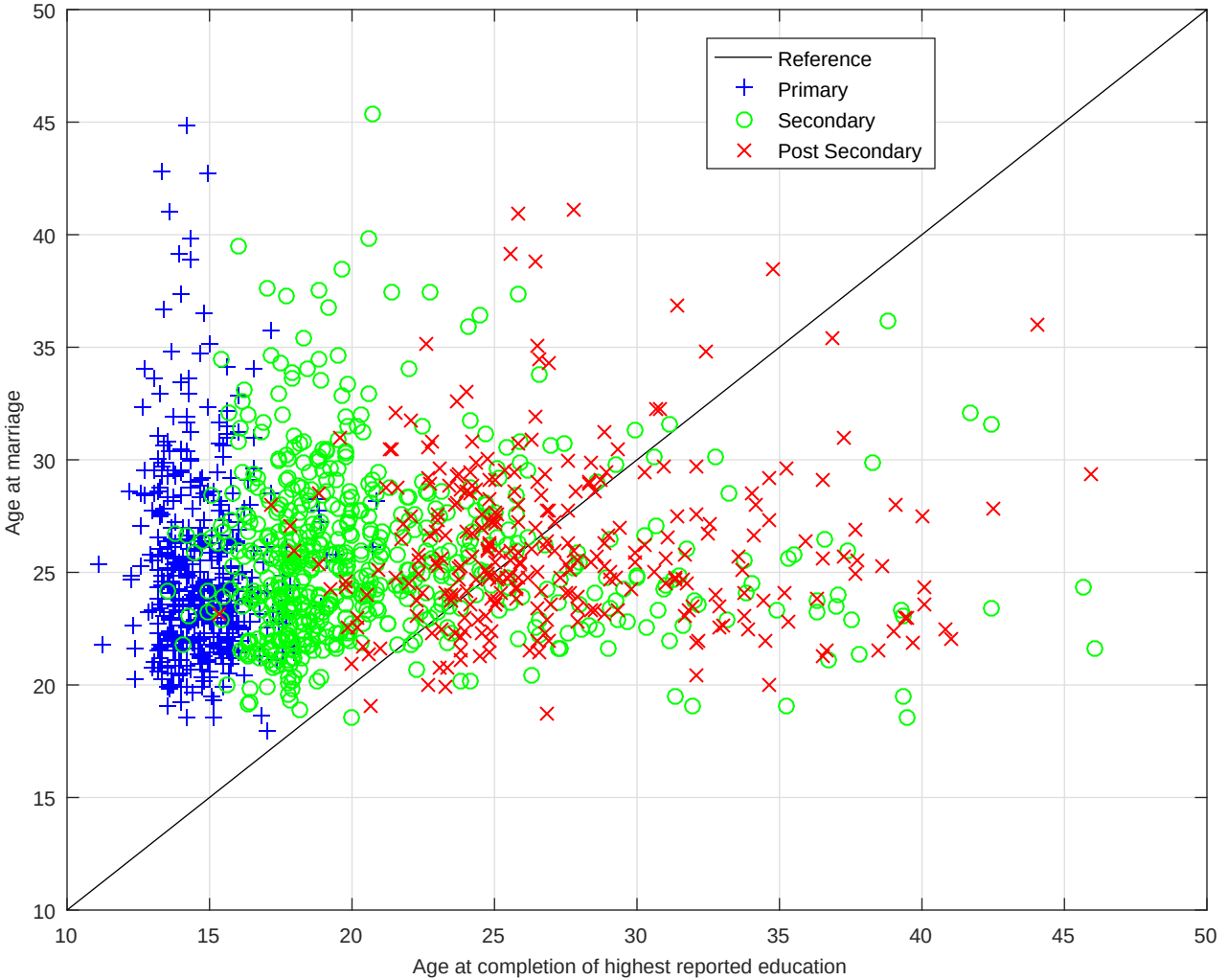

Fig. 2: Estimated conditional probabilities of educational levels at marriage given reported educational levels at interveiw

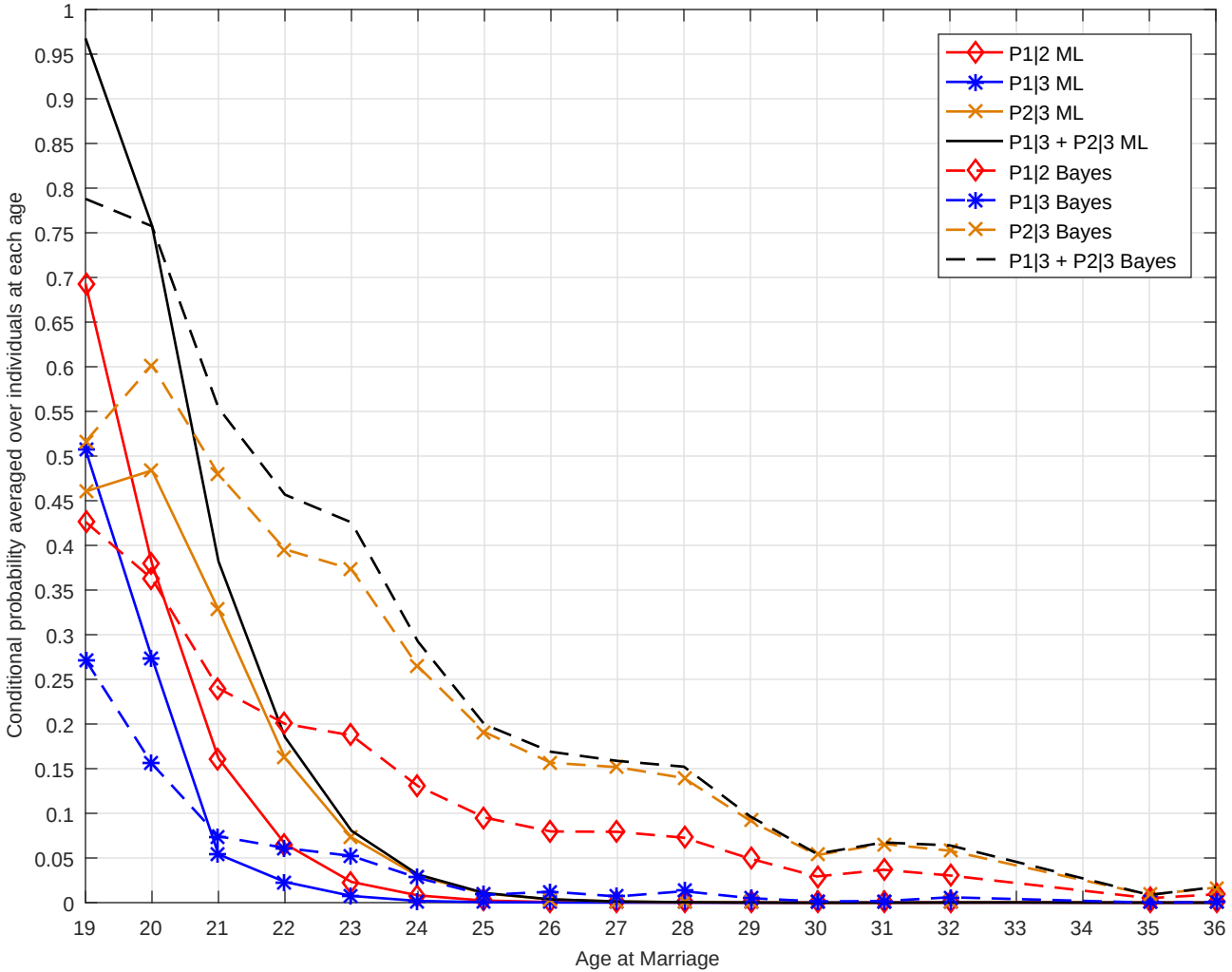

Fig. 3(a): Educational profiles of divorce risks over time in the Reduced Model

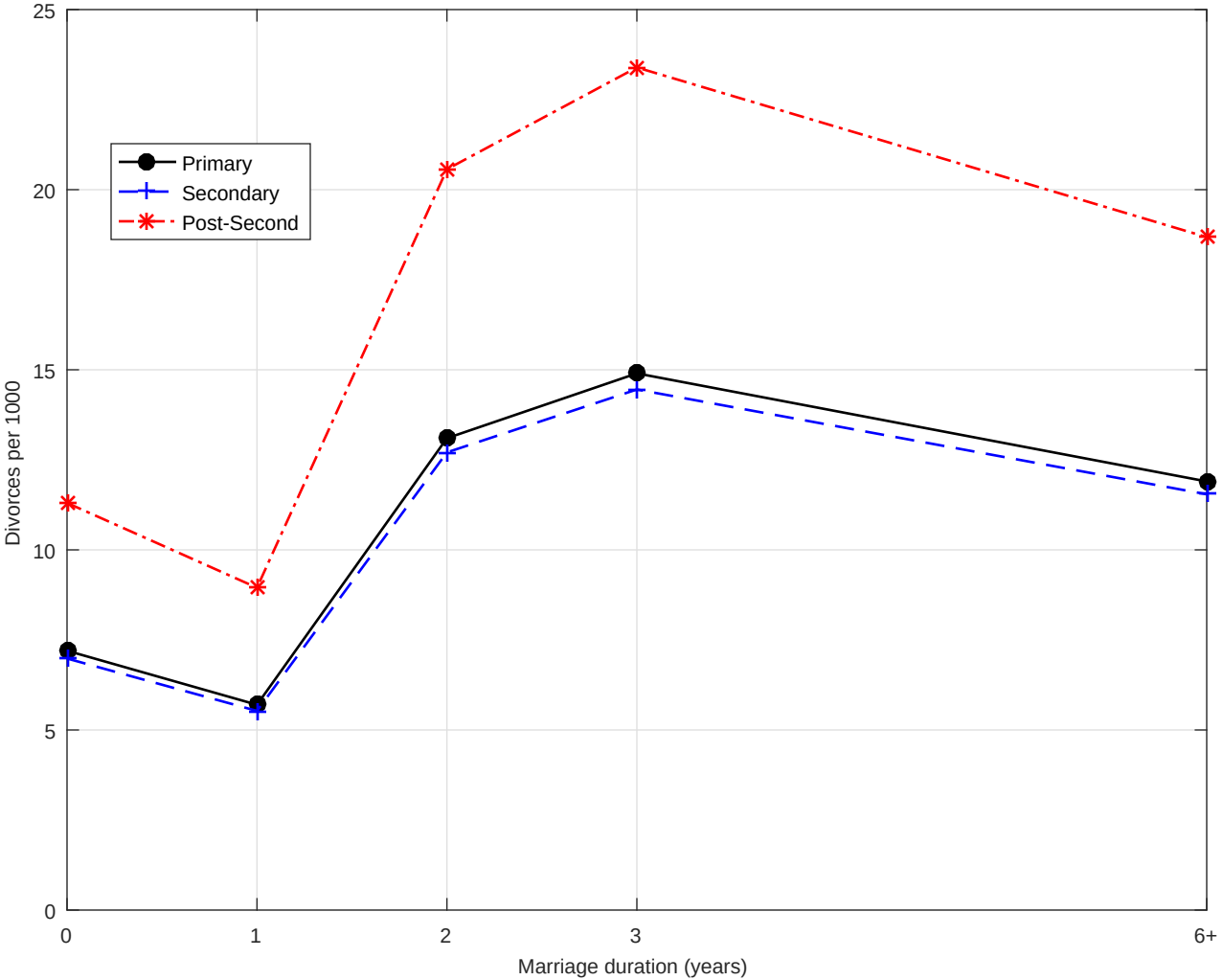

Fig. 3(b): Educational profiles of divorce risks over time in the Anticipatory Model

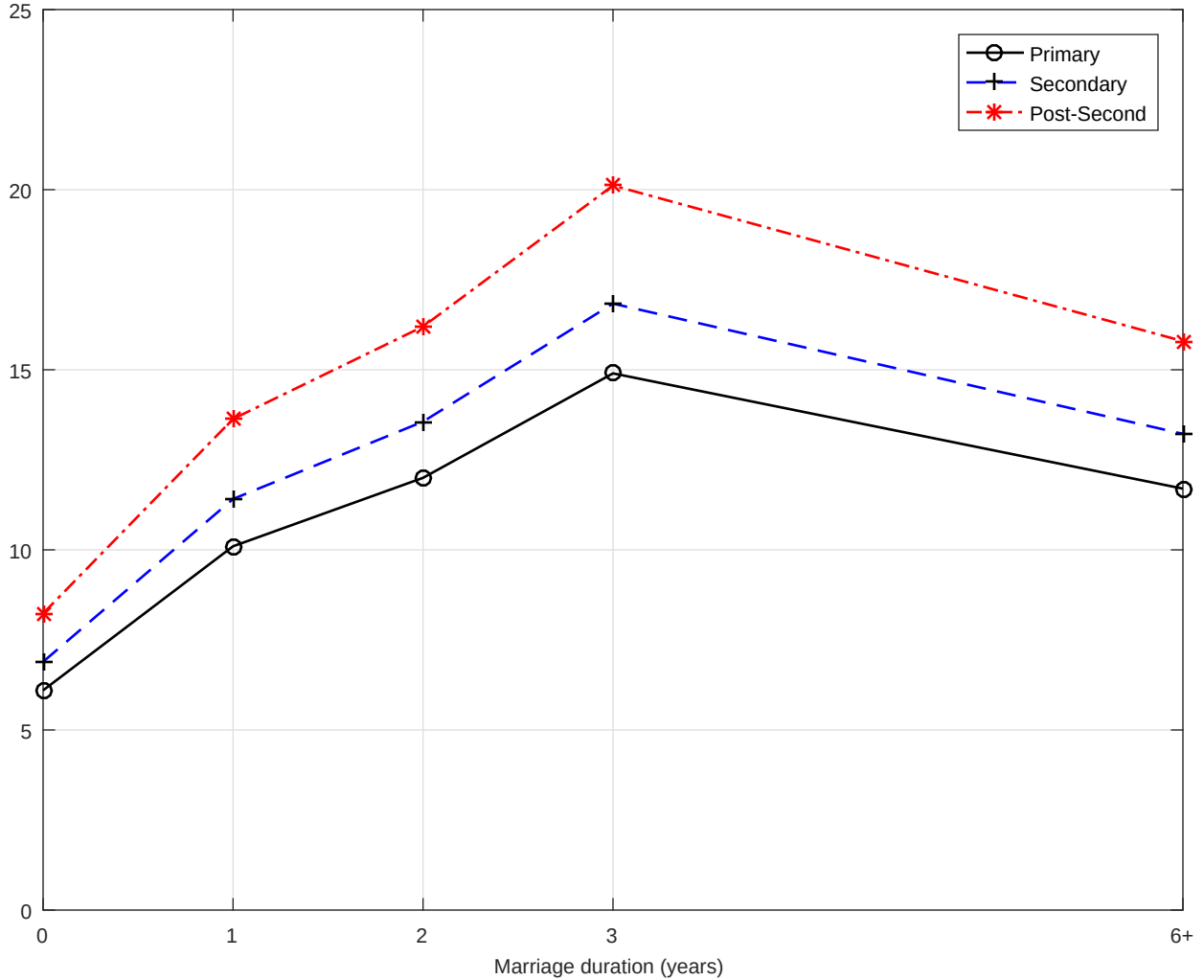

Fig. 3(c): Educational profiles of divorce risks over time in the Adjusted Model

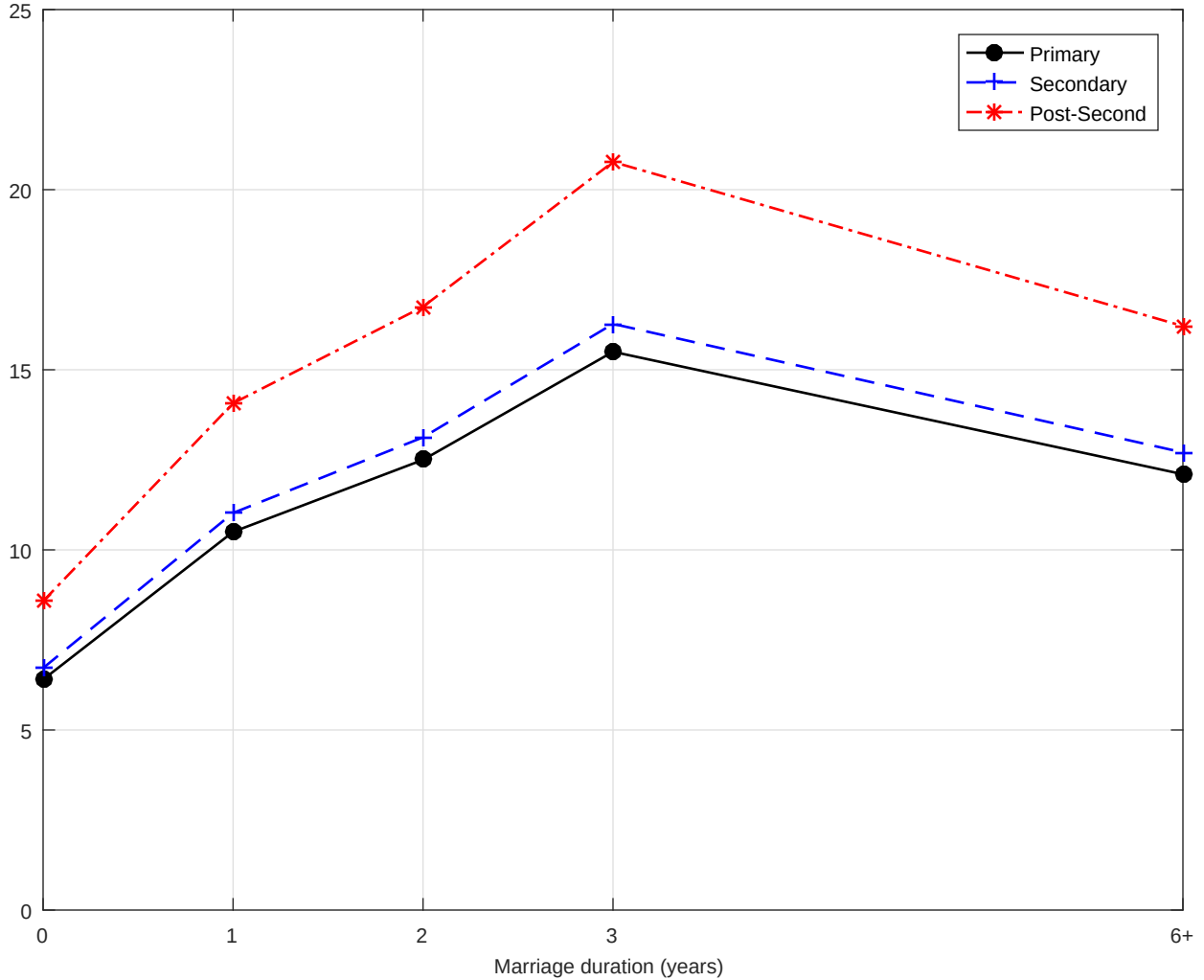

Fig. 4(a): Estimates of divorce risks for men with primary-level education across the three models

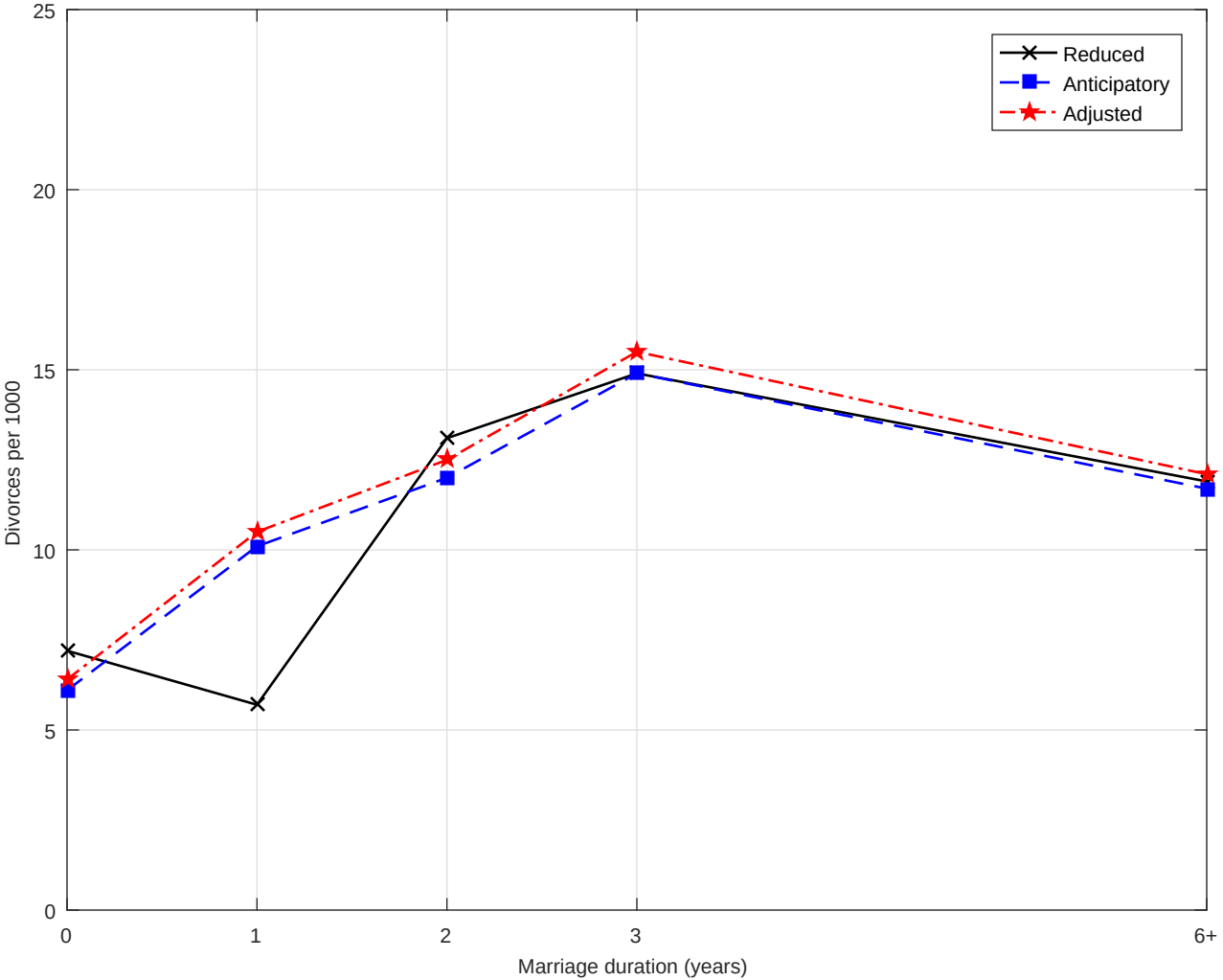

Fig. 4(b): Estimates of divorce risks for men with secondary-level education across the three models

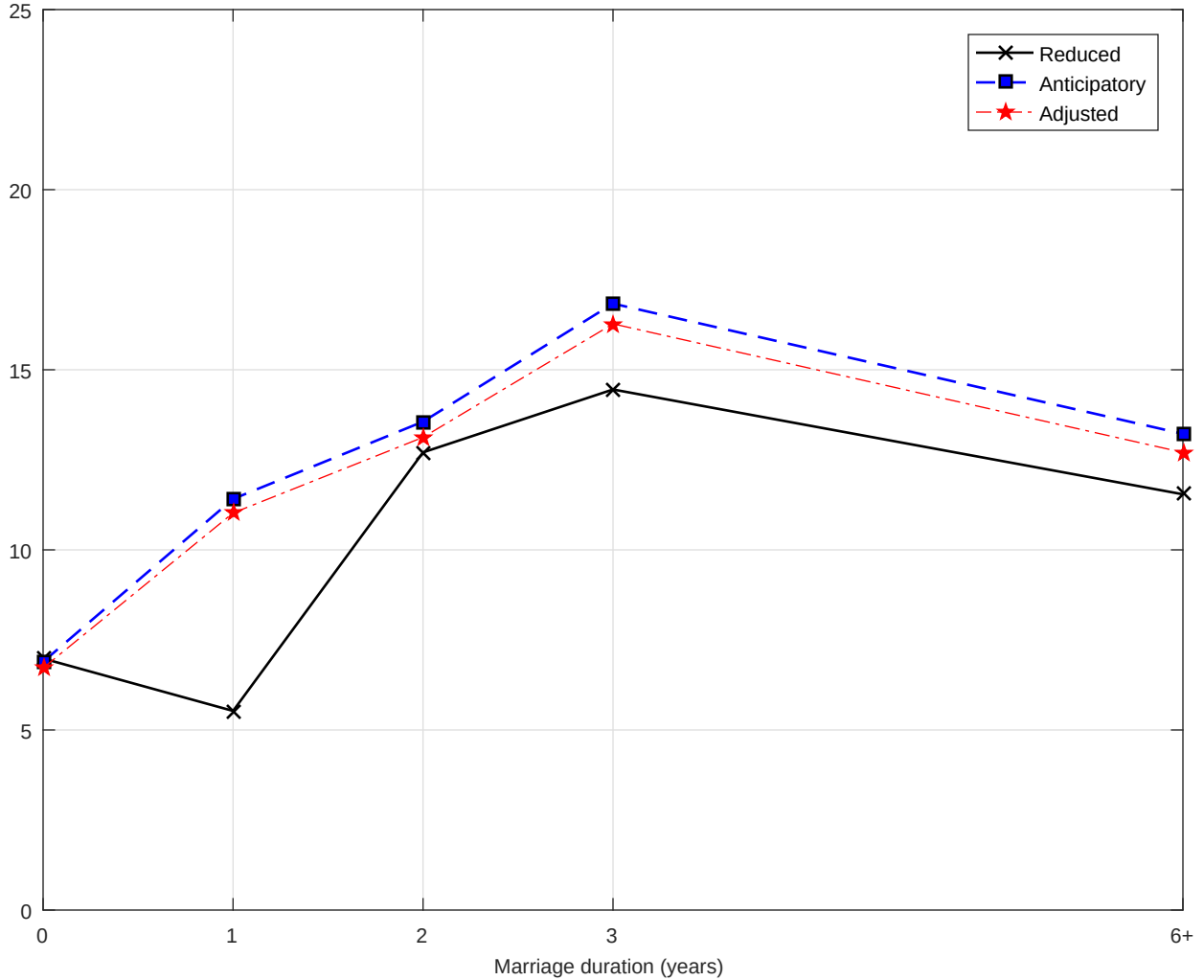

Fig. 4(c): Estimates of divorce risks for men with post-secondary level education across the three models