NONLINEAR DISCRETE-TIME HAZARD MODELS FOR ENTRY INTO MARRIAGE

By Andy Batchelor, Heather L. Turner* and David Firth*

University of Warwick, Coventry, CV4 7AL, UK

When modeling the hazard of entry into marriage, the non-monotonic dependence on age needs to be taken into account. In this paper, nonlinear discrete-time hazard models based on a bell-shaped function are proposed, in which the support of the hazard function, the maximum hazard and the age of maximum hazard are estimated. Starting in the proportional hazards framework, the baseline hazard model proposed by Blossfeld and Huinink (1991) is extended to allow estimation of the support of the baseline hazard. A naive extension is shown to suffer from partial aliasing and thus an alternative parameterization is proposed, in which the partial aliasing is reduced. This parameterization includes the maximum hazard and the age of maximum hazard as parameters. A non-proportional hazards model is then obtained by allowing the age of maximum hazard, as well as the maximum hazard itself, to depend on covariates. The usefulness of the proposed models is demonstrated through application to data from the Living in Ireland Surveys conducted between 1994 and 2001.

1. Introduction. Changes in family formation over recent decades have provided an interesting field of research for social scientists. One aspect of interest is the propensity to marry, which can be studied by analyzing the timing of first marriage. In particular, the transition from being unmarried to entering marriage can be modeled using survival analysis techniques.

We define the survival time, T, to be the number of calendar years an individual remains unmarried from the year in which they reach the minimum legal age of marriage. If $t \in \{0, 1, 2, ...\}$ is the number of calendar years since reaching the minimum legal age, then the hazard of entry into marriage at time t is defined as

$$h(t) = P(T = t | T \geq t). \tag{1}$$

In this paper, we shall develop models for this hazard, starting within the proportional hazards framework. The discrete-time proportional hazards model for an individual i with covariates $\boldsymbol{x}_{it}$ may be formulated as

$$\text{logit}(h(t|\boldsymbol{x}_{it})) = h_0(t) + \boldsymbol{x}_{it}'\boldsymbol{\beta} \tag{2}$$

[*] Supported by ESRC Professorial Fellowship RES-051-27-0055 and by the ESRC National Centre for Research Methods, Lancaster-Warwick Node (ref RES-576-25-5020).

where $h_0(t)$ is the baseline hazard (Cox and Oakes, 1984). Use of the logit link provides a direct interpretation in terms of the conditional odds of marriage.

The baseline hazard of entry into marriage has a non-monotonic dependence on age, which needs to be represented in the model. Blossfeld and Huinink (1991) propose the parametric baseline hazard

$$h_0(t|age_{it}) = c + \beta_l \log(age_{it} - 15) + \beta_r \log(45 - age_{it}), \tag{3}$$

which forms a bell-shaped curve. This model makes the assumption that the hazard of entry into marriage only exists between the ages of 15 and 45. Blossfeld and Huinink (1991) do not justify this assumption and though the left endpoint would be governed by legal constraints, it would appear that the support of the baseline hazard was simply determined by the age range of women in their study.

It would be preferable to estimate the support of the baseline hazard as part of the model. An immediate extension of Equation 3 would give the nonlinear baseline hazard

$$h_0(t|age_{it}) = c + \beta_l \log(age_{it} - \alpha_l) + \beta_r \log(\alpha_r - age_{it}). \tag{4}$$

We shall demonstrate, using a novel graphical method, that this naive extension suffers from partial aliasing between the parameters. We therefore propose an alternative parameterization, in which the partial aliasing is reduced. With this model, it is possible to test whether assumptions made about the endpoints are validated by the data. Furthermore, the parameters of the proposed model have a more useful interpretation, allowing non-proportional hazard models to be considered, in which there are interactions between the covariates and parameters of the baseline hazard.

We present our approach through application to data from the Living in Ireland Surveys, which are described in the next section. In Section 3 we follow the approach of Blossfeld and Huinink (1991) to build a reference linear discrete-time hazard model. Then in Section 3.1, we demonstrate the partial aliasing that occurs when Equation 4 is used as a baseline hazard and the improvement offered by our proposed alternative. We repeat the analysis of Section 3 with the new baseline hazard and consider further improvements to the model. Our findings our summarized in Section 4.

2. Data.

The Living in Ireland Surveys were conducted between 1994 and 2001 by the Economic and Social Research Institute. Full details of the surveys are given in (Watson, 2004). Data was collected by yearly household interviews, providing information on individuals' education, occupation and standard of living, as well as basic demographics.

We shall consider only a subset of the data here. In particular we restrict our attention to women who were members of the original sample of households and who were born between 1950 and 1975, giving five, five-year cohorts who by 2001 had passed the mean age at marriage for women in the full data set.

We represent the marital status simply by a binary variable, which is equal to one if the woman is married and zero otherwise. We focus our attention on a selection of other variables: the year and month of birth, the social class (a seven-level factor usually based on the father), the highest level of education attained (a seven-level factor) and the corresponding year of attainment.

We use the method of episode-splitting (see e.g. Powers and Xie, 2000) to generate yearly pseudo-observations data for each individual, from the year in which they became 16 up to the year in which they became 45 or were lost to follow up. The observations are assumed to be made at the start of the calendar year, so that the age at time t is taken to be

$$16 - (monb - 0.5)/12 + t$$

where $t \in \{0, 1, \ldots\}$ is the number of calendar years since that in which the woman became 16 and $monb \in 1, 2, \ldots, 12$ is the month of birth. Our final data set comprised 31009 records for 2902 women.

3. Linear Discrete-time Hazard Models.

We conduct an initial analysis of the data using the discrete-time proportional hazards model (Equation 2) with the linear baseline hazard of Equation 3, which assumes that the support of the baseline hazard is known. As far as possible, we follow the model-building strategy of Blossfeld and Huinink (1991): starting with the null model, adding the baseline hazard variables, then adding the social class, cohort and education variables in turn. The results are presented in Table 1.

As in Blossfeld and Huinink (1991), the effect of education is modeled using a time-varying binary variable which indicates whether the woman is in education or not at time t. Blossfeld and Huinink (1991) consider in addition a dynamic measure of the level of education, which we can not generate from our data. However Blossfeld and Huinink (1991) find this variable not to be significant; consistent with this result, we find that including the final level of education as a covariate does not significantly improve the model. Once

Table 1

Discrete-time proportional hazard models of entry into marriage for women born between 1950 and 1975, using the linear baseline hazard model defined in Equation 3. The body of the table shows parameter estimates, with standard errors in parentheses for Model 6.

	Model					
Variables	1	2	3	4	5	6
Intercept	−2.82	−18.11	−18.09	−19.74	−18.37	−18.33 (0.91)
Log(age - 15)		2.19	2.20	2.33	2.09	2.07 (0.10)
Log(45 - age)		3.66	3.70	4.25	3.96	3.93 (0.24)
Class s/skilled manual			−0.13	−0.11	−0.09	
Class skilled manual			−0.15	−0.07	−0.05	
Class non manual			−0.27	−0.23	−0.19	
Class low professional			−0.22	−0.20	−0.13	
Class high professional			−0.50	−0.44	−0.32	
Class missing			−0.07	−0.09	−0.04	
Cohort (54,59]				0.03	0.03	0.03 (0.07)
Cohort (59,64]				−0.09	−0.07	−0.08 (0.07)
Cohort (64,69]				−0.61	−0.59	−0.59 (0.08)
Cohort (69,74]				−1.46	−1.41	−1.42 (0.10)
In education					−2.17	−2.22 (0.31)
Deviance	13666	12573	12548	12181	12077	12089
Df	31008	31006	31000	30996	30995	31001

the educational status indicator is added to the model, the class factor can be dropped from the model without a significant increase in deviance.

Thus the final model includes the baseline hazard variables, the cohort factor and the educational status indicator. The coefficient of the first baseline variable is lower than that of the second (2.07 compared to 3.93), so the baseline hazard is right-skewed. Compared to the baseline cohort of women born in 1950-1954, the conditional odds of marriage for a woman of the same educational status are not significantly different in the 1955-1959 or 1960-1964 cohorts, but rapidly decrease through the later cohorts to 24% of the baseline conditional odds in the 1970-1974 cohort (95% confidence interval: 20 - 30%). The conditional odds of marriage for a woman who is in education are 11% of those for a woman of the same cohort who is not in education (95% confidence interval: 6 - 20%).

3.1. *Nonlinear Discrete-time Hazard Models.* We now turn to a nonlinear discrete-time proportional hazard model in which the endpoints of the support of the baseline hazard are to be estimated from the data. We first fit the baseline hazard model as defined in Equation 4 using the R package for generalized nonlinear models, gnm (Turner and Firth, 2007). The endpoint parameters α_l and α_r need to be constrained to ensure that the log terms remain finite. To enforce the constraints $\alpha_l < age_{[min]}$ and $\alpha_r > age_{[max]}$, where $age_{[min]}$ and $age_{[max]}$ are the minimum and maximum ages observed, we set

$$\alpha_l = age_{[min]} - exp(\alpha_l^*) \tag{5}$$

$$\alpha_r = age_{[max]} + exp(\alpha_r^*) \tag{6}$$

and estimate α_l^* and α_r^*. The gnm software detects numerically that the estimates of the parameters in the baseline hazard model are not identified, despite the design matrix being of full rank.

We can demonstrate the partial aliasing graphically using “recoil plots”, an example of which is given in Figure 1. We plot the fitted model on the probability scale, then plot the curve obtained by shifting one of the model parameters to a new value, and finally plot the model obtained when the parameters are re-estimated with the shifted parameter constrained to its new value. The partial aliasing is clearly apparent in Figure 1, since the re-fitted model coincides with the original model, i.e., the other parameters compensate for the arbitrary shift. A similar plot is obtained for the other parameters in the model.

Given the problem of partial aliasing, we propose a re-parameterization

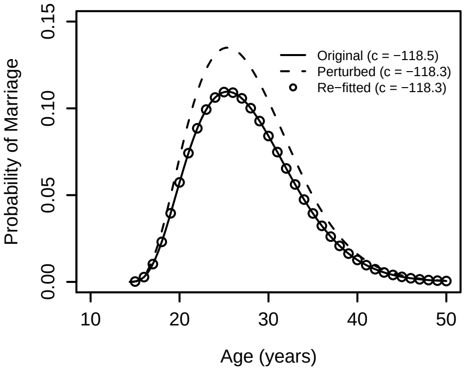

FIG 1. *A "recoil plot" for the intercept, c, demonstrating the aliasing in the baseline hazard model defined by Equation 4. Three hazard curves are shown: for the fitted model where* $c = -118.5$ *(Original); for the perturbed model with c shifted to* -118.3 *and the other parameters left at their fitted values (Perturbed), and for the re-fitted model with c constrained to* -118.3 *and the other parameters re-estimated (Re-fitted).*

of Equation 4 as follows:

$$h_0(t|age_{it}) = \gamma - \delta\left\{(\nu - \alpha_l)\log\left(\frac{\nu - \alpha_l}{age_{it} - \alpha_l}\right)\right\} + \delta\left\{(\alpha_r - \nu)\log\left(\frac{\alpha_r - \nu}{\alpha_r - age_{it}}\right)\right\}. \tag{7}$$

The corresponding set of recoil plots, Figure 2, show that the partial aliasing is greatly reduced.

An additional benefit of the new parameterization is that the parameters have a more useful interpretation, as illustrated in Figure 3. The left and right endpoints are given by the parameters α_l and α_r as before, while ν gives the location of the peak hazard and γ gives the maximum hazard on the logit scale. The fifth parameter, δ, does not have a direct interpretation, but relates to the sharpness of the peak and can be loosely interpreted as the 'fall off' from the peak.

Using the new baseline hazard, we repeat the analysis presented in Section 3, giving the results shown in Table 2. Compared to the linear baseline hazard model (Model 2, Table 1) the nonlinear baseline hazard model (Model 7, Table 2) reduces the residual deviance by 20 at the expense of two degrees of freedom. A similar reduction in deviance is seen across the models and the estimated effect of the covariates on the hazard is little changed. Therefore the qualitative interpretation remains the same, but the residual deviance is

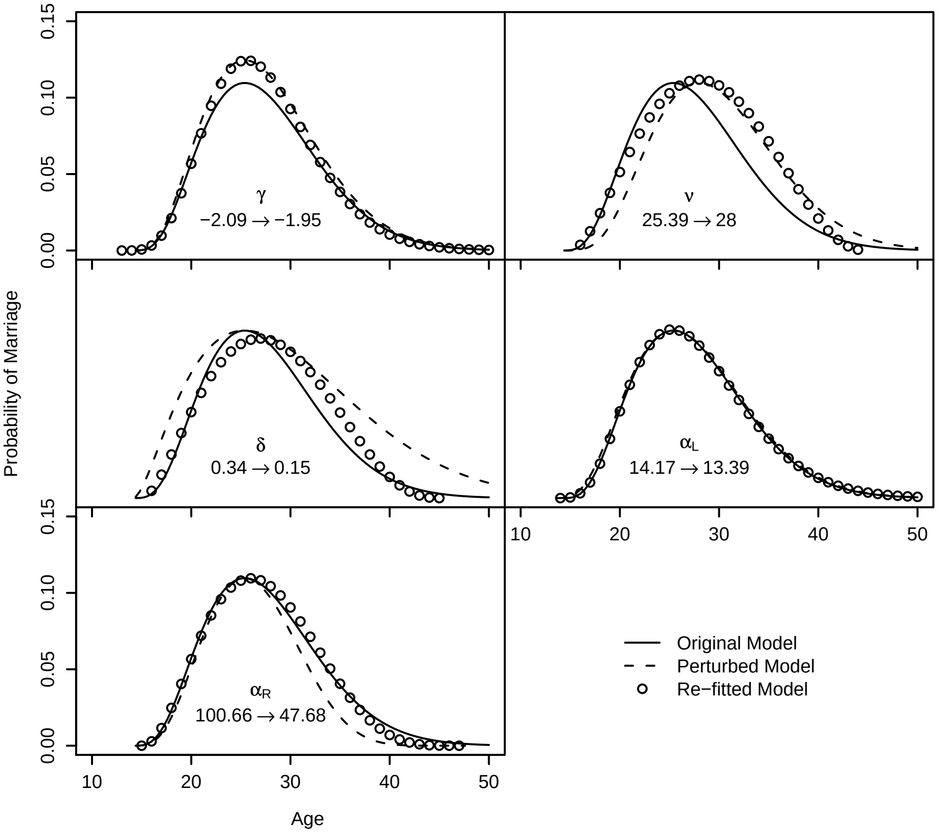

FIG 2. *Recoil plots for the parameters of the baseline hazard model defined in Equation 7. In each case three hazard curves are shown: for the fitted model (Original Model); for the perturbed model with the parameter of interest shifted to a new value and the other parameters left at their fitted values (Perturbed Model), and for the re-fitted model with the parameter of interest constrained at its new value and the other parameters re-estimated (Re-fitted Model).*

significantly reduced by estimating the support of the hazard function from the data.

As the theoretically important variables are added to the model, the estimated endpoints of the support of the baseline hazard diverge from their constraints of $15\frac{1}{24}$ and $45\frac{23}{24}$ for the left and right endpoints respectively. In particular, the right endpoint ends up at 400.15 with a standard error of 3342.66 (Model 11, Table 2). Given that this endpoint is so far from the data and indeed, in practical terms, may be regarded as representing the

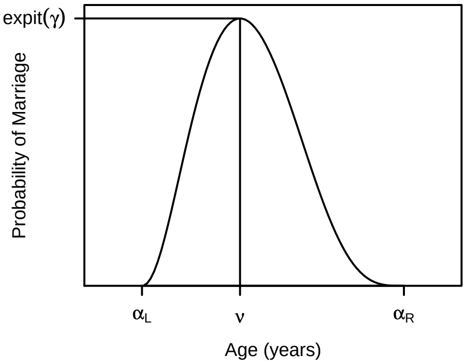

FIG 3. *An illustrative hazard curve, showing how the parameters of the baseline hazard model defined in Equation 7 relate to the features of the curve.*

TABLE 2

Discrete-time proportional hazard models of entry into marriage for women born between 1950 and 1975, using the nonlinear baseline hazard model defined in Equation 7. The body of the table shows parameter estimates, with standard errors in parentheses for Model 11.

	Model					
Variables	7	8	9	10	11	
Intercept (γ)	−2.09	−1.92	−1.60	−1.67	−1.74	(0.06)
Peak age (ν)	25.39	25.37	24.85	25.41	24.71	(0.26)
Fall-off (δ)	0.33	0.34	0.44	0.21	0.47	(0.25)
Left endpoint (α_l)	14.18	14.14	13.57	14.90	12.59	(2.24)
Right endpoint (α_r)	100.92	102.82	195.81	44.96	400.15	(3342.66)
Class s/skilled manual		−0.13	−0.11	−0.09		
Class skilled manual		−0.15	−0.07	−0.05		
Class non manual		−0.27	−0.23	−0.19		
Class low professional		−0.23	−0.20	−0.13		
Class high professional		−0.50	−0.44	−0.32		
Class missing		−0.07	−0.09	−0.04		
Cohort (54,59]			0.03	0.03	0.03	(0.07)
Cohort (59,64]			−0.08	−0.08	−0.08	(0.07)
Cohort (64,69]			−0.62	−0.59	−0.60	(0.08)
Cohort (69,74]			−1.48	−1.41	−1.43	(0.1)
In education				−2.17	−2.21	(0.31)
Deviance	12553	12527	12154	12077	12064	
Df	31004	30998	30994	30994	30999	

Table 3

Discrete-time hazard models of entry into marriage for women born between 1950 and 1975, using the baseline hazard model with infinite right endpoint defined in Equation 8. The body of the table shows parameter estimates, with standard errors in parentheses for Model 15. Models 14 and 15 are non-proportional hazard models due to the dependence of the peak location parameter (ν) on the education level.

	Model			
Variables	12	13	14	15
Intercept (γ)	−1.73	−1.63	−1.59	−1.60 (0.06)
Peak age (ν)				
Intercept	24.70	24.66	13.74	14.42 (0.9)
Education level			0.97	0.88 (0.08)
Fall-off (δ)	0.50	0.50	0.50	0.46 (0.05)
Left endpoint (α_l)	12.33	12.38	13.98	13.86 (0.65)
Cohort (54,59]	12.33			
Cohort (59,64]	−0.08			
Cohort (64,69]	−0.60			
Cohort (69,74]	−1.43			
Birth year				
1950 effect		12.38	−0.02	−0.02 (0.01)
Exponential decay		0.19	0.20	0.20 (0.03)
In education	−2.21	−2.20	−1.13	−1.46 (0.33)
Early years post-educaton				−0.48 (0.11)
Deviance	12064	12046	11867	11847
Df	31000	31002	31001	31000

end of life, it seems sensible to consider an alternative baseline hazard in which the right endpoint tends to infinity. Letting $\alpha_r \to \infty$ in Equation 7 we obtain:

$$h_0(t|age_{it}) = \gamma - \delta \left\{ (\nu - \alpha_l) \log \left(\frac{\nu - \alpha_l}{age_{it} - \alpha_l} \right) - age_{it} - \nu \right\} \tag{8}$$

Re-fitting Model 11 (Table 2) with an infinite right endpoint, the deviance is increased by only 0.01 on one degree of freedom, so there is no significant difference (Model 12, Table 3). The estimates of the remaining parameters are largely unchanged, except that the estimated 'fall off' is slightly increased from 0.47 (s.e. 0.25) to 0.50 (s.e. 0.07) and the left endpoint moves down from 12.59 (s.e. 2.24) to 12.33 (s.e. 1.11).

The cohort effects have the pattern noted in Section 3, that is, there is no significant cohort effect until after 1964, when the cohort effects become increasingly negative. This suggests that a cohort factor may not be the best way to model the effect of the year of birth. To investigate this further, we

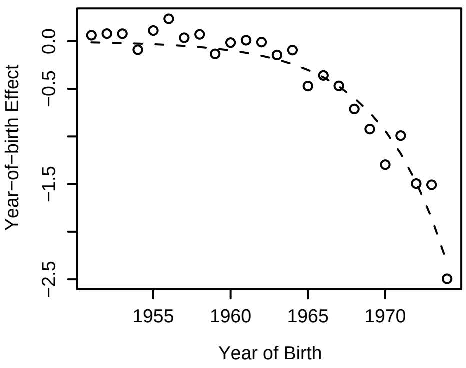

FIG 4. *Estimated year-of-birth effects when the cohort factor in Model 12 is replaced by a year-of-birth factor. The effect for year-of-birth equal to 1950 is set to zero.*

fit a model in which the cohort factor is replaced by a year-of-birth factor and plot the fitted effects (Figure 4). The pattern of these effects suggests that a more appropriate model might be

$$\theta \exp(\lambda(yrb_i - 1950)). \tag{9}$$

where yrb_i is the year of birth for individual i. Fitting this curve directly to the year-of-birth effects seems to give a reasonable fit (Figure 4), so we include this nonlinear term in our model and drop the cohort factor. This reduces the deviance by 19 while increasing the residual degrees of freedom by 2 (Model 13, Table 3).

In order to check the fit of Model 13, we compare the observed and fitted proportions in different sub-groups of the data. Figure 5, shows the observed and fitted proportions for each year of age, by highest level of education attained. The seven levels of education have been reduced to five, since the "sub-primary" group and the "post-leaving certificate" group were small in size and could be merged with the "primary" and "institute of technology" groups respectively, since the pattern of observed proportions were not dissimilar. From Figure 5 we can see that the fitted model fits quite well for the "upper secondary" group, but peaks too late for lower education levels and too early for higher education levels.

This suggests that the peak location depends monotonically on education level. We quantify the education level, *ed*, as the equivalent average years spent in education (calculated from the data) and allow for a linear trend

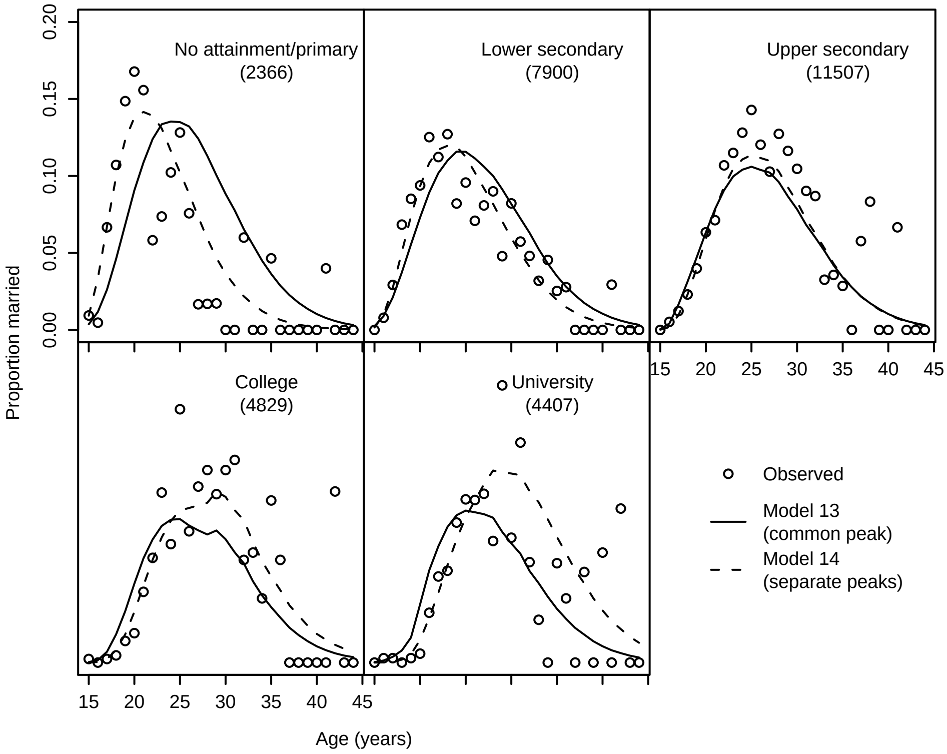

FIG 5. *Fitted hazard curves for Model 13, with a common peak location for all levels of education (common ν in the baseline hazard) and Model 14, with a separate peak location for each level of education (ν dependent on equivalent average years spent in education). The curves are laid over the observed proportion married for each year of age.*

in the model as follows:

$$\text{(10)}\quad \operatorname{logit}(h(t|\boldsymbol{x}_{it})) = \gamma - \delta\left\{(\nu_0 + \nu_1 ed_i - \alpha_l)\log\left(\frac{\nu_0 + \nu_1 ed_i - \alpha_l}{age_{it} - \alpha_l}\right)\right\} + \delta\left\{age_{it} + \nu_0 + \nu_1 ed_i\right\} + \theta\exp(\lambda(yrb_i - 1950)) + \beta_1 edstat_{it}.$$

where $edstat_{it}$ is the binary variable indicating whether a woman is in education or not at time t. This is no longer a proportional hazards model, since both the scale and location of the hazard curve can vary between individuals. Allowing separate peak locations for each education level visually improves the fit (Figure 5) and significantly reduces the deviance (Model 14, Table 3).

Continuing to check the fit of the model as before, we find no particular lack of fit over age, class or year of birth. Grouping the data by the

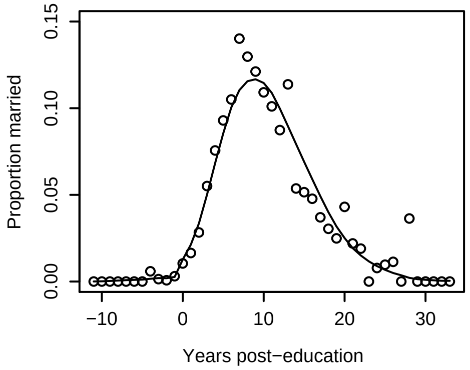

Fig 6. *The fitted hazard curve for Model 14, laid over the observed proportion married for each year post-education.*

educational status indicator would not be meaningful, so instead we group by years post-education as shown in Figure 6. There are several points to note about this plot. First, there is some evidence that the rate of increase in the proportion married is lower in the first three years post-education than it is three to six years post-education. Second, there appears to be a sharp change in the pattern of marriages at seven years post-education and the model does not capture this change. Finally, there are unusually high proportions of women marrying at 13, 20 and 28 years post-education.

With regard to the outlying points, we note that the sample size decreases with increasing years post-education, so the observed proportions would be more influenced by small fluctuations in the number of marriages. From a total sample size of 2902, the outlying points at 13, 20 and 28 years post-education are based on 545, 209 and 55 women respectively. Examining the data at 13 years post-education more closely, we find no unusual characteristics of the 62 women who marry that might explain the high proportion of marriages. There is a high proportion of 29 and 30 year olds (36/62), but this is in keeping with the observed trend over years post-education.

Similarly, we find no simple explanation for the sharp peak at seven years post-education. The model does not capture this peak, even if we re-fit the model with the data for 13 years post-education removed. We could improve the fit by allowing the peak location to change after six years post-education, however we are concerned that the sharp peak may be a feature specific to this data set and we do not wish to over-fit.

On the other hand, allowing a separate effect for the early years post-education seems quite natural, as one might expect a lower rate of marriage while women establish themselves in their careers. Furthermore the sample sizes are at their highest here: over 2400 women contribute to the observed proportions. We supplement the “in education” indicator by a second binary variable that indicates when a woman is in her first three years post-education. This reduces the deviance by 19.5 for 1 degree of freedom (Model 15, Table 3).

The hazard and survival curves for our final model are shown in Figure 7. The left endpoint of the support of the hazard function is estimated as 13.86 years (s.e. 0.65) and the deviance is significantly increased if this endpoint is constrained to 15 years as in the linear model. The location of the peak hazard varies from 21.32 years (s.e. 0.10) for the group with no formal education to 27.60 years (s.e. 0.14) for university graduates. For a woman born in 1950, the peak hazard of entry into marriage is 0.17 (s.e. 0.002). This probability drops slightly to 0.15 (s.e. 0.001) for a woman born in 1960, but drops to 0.07 (s.e. 0.003) for a woman born in 1970. Clearly this model is inappropriate for predicting the hazard of marriage for women born after 1974 (the last year included in the analysis) as the peak hazard would soon be near-zero. A logistic term of the form

$$\frac{\theta}{1+\exp(\lambda(\mu - yrb_i))} \tag{11}$$

may be more appropriate in this case, but it did not significantly improve the fit for our data.

4. Summary.

We have shown that restricting the support of the hazard function to the age range represented in the sample may not be justified and an improved model may be obtained by estimating the support of the hazard function from the data. We have illustrated the problem of partial aliasing that can arise from an naive extension of a linear model to a nonlinear model, but have demonstrated how such problems can be overcome through re-parameterization.

In addition to estimating the support of the hazard function from the data, our proposed model has the benefit of more interpretable parameters, allowing investigation of the effect of covariates on both the location and scale of the maximum hazard. These features of the hazard curve are both the most interesting from a substantive point of view (Raymo, 2003) and the ones on which there is the most information in the data. Therefore, although it would theoretically be possible to allow dependence of the endpoints or the fall-off parameter on covariates, we do not recommend this.

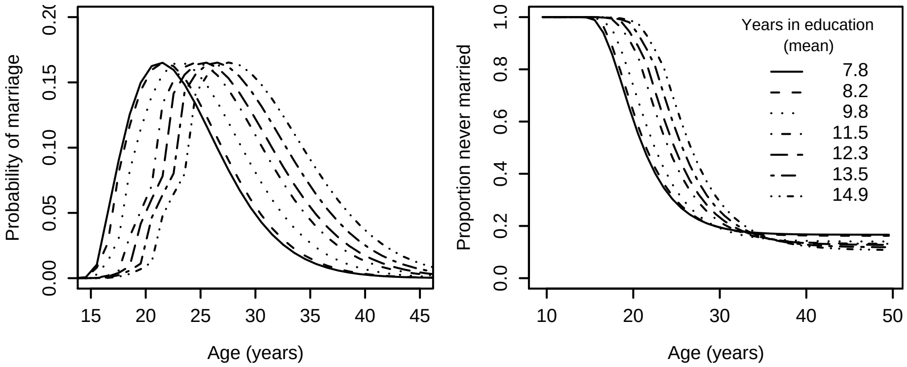

Fig 7. *The fitted hazard curves (left) and survival curves (right) for women born in 1950, for each level of education: Sub-primary (average 7.8 years in education), Primary (8.2 years), Lower Secondary (9.8 years), Upper Secondary (11.5 years), Post-Leaving Certificate (12.3 years), Institute of Technology (13.5 years), University (14.9 years).*

The parametric form of our model does impose some restrictions on the shape of the hazard curve. While these restrictions reduce the influence of outliers on the model, it also means that the model does not always capture fully the pattern of the data. Nevertheless, we believe that our proposed model strikes a useful balance between flexibility and interpretability.

Acknowledgements. The data are from The Economic and Social Research Institute Living in Ireland Survey Microdata File (©Economic and Social Research Institute). We gratefully acknowledge Carmel Hannan for introducing us to this application and providing background on the data.

References.

Blossfeld, H.-P. and J. Huinink (1991). Human Capital Investments or Norms of Role Transition? How Women's Schooling and Career Affect the Process of Family Formation. *American Journal of Sociology 97*, 143–168.

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. London: Chapman & Hall.

Powers, D. A. and Y. Xie (2000). *Statistical Methods for Categorical Data Analysis*. London: Academic Press.

Raymo, J. (2003). Educational Attainment and the Transition to First Marriage among Japanese Women. *Demography 40*, 83–103.

Turner, H. and D. Firth (2007). Generalized nonlinear models in R: An overview of the gnm package. Documentation in the *gnm* package, `http://cran.r-project.org`.

Watson, D. (2004). Living in Ireland Survey - Technical Overview. http://issda.ucd.ie/documentation/esri/lii-overview.pdf.