Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2002/02

Grouped continuous and continuation ratio discrete time versions of the proportional hazard model with random effects: are they really equivalent?

Leonardo Grilli

Università degli Studi
di Firenze

Statistics

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Grouped continuous and continuation ratio discrete time versions of the proportional hazard model with random effects: are they really equivalent?

Leonardo Grilli
Università degli Studi di Firenze, Italy

Address for correspondence: Leonardo Grilli, Dipartimento di Statistica "G. Parenti", Università degli Studi di Firenze, Viale Morgagni 59, 50134 Firenze, Italy. E-mail: grilli@ds.unifi.it

Summary. When analyzing grouped time survival data having a hierarchical structure it is often appropriate to assume a random effects proportional hazards model for the latent continuous time and then derive the corresponding grouped time model. There are two formally equivalent grouped time versions of the proportional hazards model obtained from a different perspective, known as *continuation ratio* (Kalbfleisch and Prentice, 1973) and *grouped continuous* (McCullagh, 1980). However the two models require distinct estimation procedures and, more important, they differ substantially with respect to the extensibility to time-dependent covariates and/or non proportional effects. The paper discusses these issues in the context of random effects models, illustrating the main points with an application to a complex data set on job opportunities for a cohort of graduates.

Keywords: Complementary log-log link, Discrete time survival models, Proportional hazards model, Random effects.

1 Introduction

In many research areas the analysis of the time elapsed between two events of interest may require some special procedures because: a) the reported times are grouped into months, quarters etc.; b) the phenomenon under study is characterized by a hierarchical structure (for example, the graduates who seek for job may be grouped by school, or by neighborhood). Point *a* implies the use of discrete time survival models (continuous time models are inadequate due to the large number of ties), while point *b* calls for the inclusion of random effects which describe the correlation structure of the statistical units.

Discrete time survival models (Allison, 1982) and random effects (multilevel) models (Goldstein, 1995) each have a long history, but their conjoint use is quite recent: see Barber *et al.* (2000), Hedeker *et al.* (2000), Biggeri *et al.* (2001) and Reardon *et al.* (2001). Among the cited papers only the work of Hedeker *et al.* focus on the complementary log-log model, which is the grouped time version of the continuous time proportional hazards model (Cox, 1972; Kalbfleisch and Prentice, 1973). The present paper is intended to highlight, in the context of random effects models, the theoretical and practical differences between the two discrete time versions of the proportional hazards model, known as *grouped continuous* (McCullagh, 1980) and *continuation ratio* (Kalbfleisch and Prentice, 1973). The two models are formally equivalent, but they require distinct estimation procedures and, more important, they differ substantially with respect to the extensibility to time-dependent covariates and/or non proportional effects.

The structure of the paper is as follows. Section 2 shows the derivation of the two models under study as grouped time versions of the random effects proportional hazards model, while Section 3 discusses their differences with respect to estimation and extensibility to time-dependent covariates and/or non proportional effects. Section 4 presents an application of the models to an analysis of the time to obtain the first job for a cohort of graduates, showing the superiority, in that context, of the extended continuation ratio model. Section 5 concludes with some remarks.

2 From the random effects proportional hazards model to the corresponding grouped time models

When the time of interest is continuous and the subjects are clustered, a suitable model may be the classical proportional hazards model (Cox, 1972) with the addition of random effects (RPH model: e.g. Vaida and Xu, 2000):

$$S_Y(y \mid \mathbf{x}_{ij}, u_j) = \{S_{Y,00}(y)\}^{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}+u_j\right)}, \tag{1}$$

where S_Y is the conditional survivor function of the continuous r.v. Y, which represents the survival time; $\mathbf{x}_{ij}$ is the covariate vector of subject i in cluster j, possibly including

cluster-level variables; u_j is the random effect of cluster j; and $S_{Y,00}$ is the conditional baseline survivor function, i.e. $S_Y(y \,|\, \mathbf{x}_{ij} = \mathbf{0},\, u_j = 0)$. The model parameters are the regression slopes $\boldsymbol{\beta}$ and the random effects variance σ_u^2, while $S_{Y,00}$ is an arbitrary function which can be estimated non parametrically.

In this paper the words 'conditional' and 'marginal' refer to the random effects, the conditioning on the covariates being implicit. Usually the random effects are assumed to be iid Gaussian with zero mean and unknown variance σ_u^2, but other choices are possible as well. Note that (1) represents the simplest case of random effects model, with a single random effect on the intercept; to avoid unnecessary complications in the formulae, in the following only this case will be considered, the extension to multiple random effects being straightforward.

Suppose now that, due to coarse measurement, the continuous r.v. Y cannot be observed; rather the times are grouped into disjoint intervals

$$[y_0 = 0, y_1), [y_1, y_2), \ldots, [y_{t-1}, y_t), \ldots, [y_{t_{\max}-1}, y_{t_{\max}} = \infty).$$

The resulting discrete r.v. T, with values in $\{1, 2, \ldots, t_{\max}\}$, derives from Y accordingly to the following relationship:

$$\{T = t\} \Leftrightarrow \{y_{t-1} \leq Y < y_t\}\,.$$

Consequently the conditional survivor function of T is

$$S(t \,|\, \mathbf{x}_{ij}, u_j) \equiv \Pr(T > t \,|\, \mathbf{x}_{ij}, u_j) = \Pr(Y \geq y_t \,|\, \mathbf{x}_{ij}, u_j),$$

with corresponding hazard function

$$\lambda(t \,|\, \mathbf{x}_{ij}, u_j) \equiv \Pr\left(T = t \,|\, T \geq t\,,\, \mathbf{x}_{ij}, u_j\right) = \Pr\left(y_{t-1} \leq Y < y_t \,|\, Y \geq y_{t-1}\,,\, \mathbf{x}_{ij}, u_j\right).$$

Note that for the discrete r.v. T the hazard-survivor relationship is

$$\lambda(t \,|\, \mathbf{x}_{ij}, u_j) = 1 - \frac{S(t \,|\, \mathbf{x}_{ij}, u_j)}{S(t-1 \,|\, \mathbf{x}_{ij}, u_j)}, \tag{2}$$

while the distribution-survivor relationship is, as in the continuous time case,

$$F(t \mid \mathbf{x}_{ij}, u_j) = 1 - S(t \,|\, \mathbf{x}_{ij}, u_j).$$

Now, using the relationship $S_{Y,00}(y) = \exp\{-\Lambda_{Y,00}(y)\}$, where $\Lambda_{Y,00}$ is the conditional baseline integrated hazard function (see, for example, Kalbfleisch and Prentice, 1980), a little algebra yields

$$\log(-\log(1 - F(t \mid \mathbf{x}_{ij}, u_j))) = \alpha_t^{(RGC)} + \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j, \tag{3}$$

where $\alpha_t^{(RGC)} \equiv \log \Lambda_{Y,00}(y_t)$; and

$$\log(-\log(1 - \lambda(t \mid \mathbf{x}_{ij}, u_j))) = \alpha_t^{(RCR)} + \mathbf{x}'_{ij}\boldsymbol{\beta} + u_j\,, \tag{4}$$

where $\alpha_t^{(RCR)} \equiv \log\{\Lambda_{Y,00}(y_t) - \Lambda_{Y,00}(y_{t-1})\}$. Equation (3) defines the random effects grouped continuous (RGC) model (McCullagh, 1980) while equation (4) defines the random effects continuation ratio (RCR) model (Kalbfleisch and Prentice, 1973). Therefore the RGC and RCR models are two grouped time versions of the RPH model; although the link function is applied to different quantities (the distribution function and the hazard function, respectively), the two models are formally equivalent, the only difference being in the parametrization of the baseline hazard. The $\alpha_t^{(RGC)}$'s and $\alpha_t^{(RCR)}$'s are linked by the following formula:

$$\alpha_t^{(RGC)} = \log\left\{\sum_{s=1}^{t} \exp\left(\alpha_s^{(RCR)}\right)\right\}. \tag{5}$$

For both the RGC and RCR models the conditional survivor function is:

$$S(t \,|\, \mathbf{x}_{ij}, u_j) = S_{00}(t)^{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}+u_j\right)}, \tag{6}$$

where $S_{00}(t)$ is the conditional baseline survivor function

$$S(t \,|\, \mathbf{x}_{ij} = 0,\, u_j = 0) = \begin{cases} \exp\left\{-\exp\left(\alpha_t^{(RGC)}\right)\right\} & \text{for the RGC model} \\ \exp\left\{-\sum_{s=1}^{t}\exp\left(\alpha_s^{(RCR)}\right)\right\} & \text{for the RCR model .} \end{cases}$$

Note that the conditional survivor function (6) has the same form as in the RPH model (1). This is no longer true for the conditional hazard function, which is

$$\lambda(t \mid \mathbf{x}_{ij}, u_j) = 1 - \left\{\frac{S_{00}(t)}{S_{00}(t-1)}\right\}^{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}+u_j\right)}. \tag{7}$$

For two arbitrary subjects A and B (7) implies that

$$\frac{\log\left\{1-\lambda(t \mid \mathbf{x}_{ij}^{(A)}, u_j^{(A)})\right\}}{\log\left\{1-\lambda(t \mid \mathbf{x}_{ij}^{(B)}, u_j^{(B)})\right\}} = \exp\left\{\left(\mathbf{x}_{ij}^{(A)} - \mathbf{x}_{ij}^{(B)}\right)'\boldsymbol{\beta}\right\} \cdot \exp\left\{u_j^{(A)} - u_j^{(B)}\right\},$$

showing that in the grouped time versions of the proportional hazards model (with or without random effects) the proportionality does not apply directly to the hazard, but to a particular transformation of it.

As for the relationship between conditional and marginal models, observe that the marginal survivor and hazard functions are, respectively,

$$\begin{aligned} S^{\text{marg}}(t \,|\, \mathbf{x}_{ij}) &\equiv \Pr(T > t \,|\, \mathbf{x}_{ij}) \\ &= E\left[S(t \,|\, \mathbf{x}_{ij}, u_j)\right] \end{aligned}$$

$$\begin{aligned} \lambda^{\text{marg}}(t \,|\, \mathbf{x}_{ij}) &\equiv \Pr(T = t \,|\, T \geq t\,,\, \mathbf{x}_{ij}) \\ &\neq E\left[\lambda(t \,|\, \mathbf{x}_{ij}, u_j)\right] \end{aligned}$$

Note that the structure which holds conditionally in general does not hold marginally; for example

$$S^{\text{marg}}(t \mid \mathbf{x}_{ij}) \neq \{S_0^{\text{marg}}(t)\}^{\exp\left(\mathbf{x}'_{ij}\boldsymbol{\beta}\right)},$$

where $S_0^{\text{marg}}(t)$ is the marginal baseline survivor function. The omission of the relevant random effects is thus a potentially serious misspecification error, which in the context of survival models causes, among other things, the underestimation of the baseline hazard (this phenomenon is known as duration bias; see Barber *et al.*, 2000).

3 A comparison between grouped continuous and continuation ratio models and their extensions

The fact that RGC and RCR models are formally equivalent may seem to close the question about which one to choose in a specific application. However there are important differences in two aspects: a) the possibility to extend the models in order to include time-dependent covariates and/or non proportional effects; b) the estimation method. It should be stressed that these differences, which will be discussed below, hold regardless of the presence of random effects.

As for the first point, note that the RGC model is based on the distribution function, so time-dependent covariates are meaningless. However it is possible to relax the proportionality assumption for a covariate x_{ij} by adding to the linear predictor, for each t, an interaction term $\gamma_t \cdot x_{ij}$ which modifies the baseline parameter α_t (Hedeker *et al.*, 2000). Thus the effect of x_{ij} freely changes with t, but at the price of adding $t_{\max}$-1 parameters (γ_1 is constrained to zero to guarantee identifiability). This price may be prohibitive when $t_{\max}$ is large or when there are many covariates with non proportional effect.

The RCR model, on the contrary, refers to the hazard function, which is a conditional probability, so it is not a problem to include a time-dependent covariate x_{ijt}. This opportunity is valuable also when in the data there are no variables of this kind, since it allows to:

- Adopt a parametric specification of the baseline hazard, replacing the α_t's with the parameters of a suitable function of time (time is a special case of time-dependent covariate). The typical choice is to use an R-grade polynomial $\sum_{r=0}^{R} \delta_r t^r$, whose adequacy can be judged through formal tests or graphical methods (Reardon *et al.*, 2001). The main advantage of a parametric specification is parsimony, especially when the number of time-intervals, $t_{\max}$, is high; moreover, a parametric specification allows a more reliable estimation for the time-intervals with few cases at risk (usually the last ones).

- Take into account the non proportional effect of a covariate x_{ij} in a parsimonious way by building interactions between the covariate and some suitable functions of time, for example $x_{ij} \cdot t$, $x_{ij} \cdot t^2$ etc.

As for estimation, maximum likelihood estimates for the RGC model and its extension to non proportional effects can be obtained using the standard methods for random effects ordinal models, with a simple modification to take into account right censoring. For example, Hedeker *et al.* (2000) describe an algorithm based on Gaussian quadrature which is implemented in the package MIXOR (Hedeker and Gibbons, 1996). Also the SAS NLMIXED procedure (SAS Institute, 1999) can be readily adapted to fit such a model.

On the other hand, there is currently no software for fitting the RCR model. In this case the usual strategy, described in detail in Barber *et al.* (2000), relies on the construction of a person-period data set in which every survival time and corresponding censoring indicator are replaced by a set of person-period indicators of event occurrence, which can be assumed as the response variable in a standard multilevel binary model. So for the RCR model and its extension to time-dependent covariates the estimation can be accomplished by any of the various algorithms for random effects binary models, but requires a preventive step of data transformation. This step causes an increase in the number of records which depends on the observed survival times; obviously, if the number of time intervals, $t_{\max}$, is high the new data set is likely to be considerably larger than the original one, with negative consequences on computing times.

4 Application to the analysis of the time to obtain the first job

To illustrate the relative merits of the RGC and RCR models and their extensions we describe an analysis on the time to obtain the first job for a sample of Italian graduates, taking the data from a survey on the high-school graduates of the year 1995, carried out by the Italian National Statistical Institute three years later (Italian National Statistical Institute, 2000). For the present analysis we employed a subsample of 9404 graduates coming from 1448 schools, obtained by excluding from the whole sample: a) the individuals who already had a job before getting the certificate; b) the individuals who are not really interested in finding a job, mainly because they keep on studying; c) the cases with some missing value in the variables of interest (about 3% of the remainder). Though the survey is retrospective, the data allow to calculate the time in quarters (from 1 to 13) needed to obtain the first job; the survival time is censored on the 13^{th} quarter for the 3466 subjects who were not able to obtain a job by the date of the interview. Moreover, the data contain many individual-level variables (such as gender, final mark, military service, attendance of university courses, occupational status of the parents) and a few contextual variables (in essence, the type of the school and region in which it is located). In addition, we added to the data set the 1995's unemployment rate at regional level for young people, and its variation from 1995 to 1998. Table 1 reports the name, definition and sample average for the variables which are included in the final model.

The covariates relative to the military service and to the enrolment in and abandon of various types of courses are time-varying in nature, but the data do not allow to determine the corresponding time series, so they are included in the model as fixed-time covariates. From other items of the questionnaire and external sources we can infer that, in most cases, the beginning of the military service (which was compulsory and one year long) and the enrolment in courses is located in the first year after the achievement of the certificate, so we expect the effects to be stronger in the first half of the observation period, which is to say that such covariates have non proportional effects. Obviously, the use of such ill-defined covariates is subject to criticism; however we retained them because: a) they still carry some information about the pattern of the hazard; b) the interpretation of the other effects if safe, since in the modeling process we found no significant interactions with the other covariates.

Figure 1 shows the non parametric estimate of the hazard function: apart from the large value of the first quarter, which is, as expected, anomalous, the hazard fluctuates between 0.045 and 0.064 until the 10^{th} quarter, when it shifts to a higher level until the end of the observation period. Separate curves for specific values of the variables (not reported here) confirm the non proportional effect of the covariates relative to the military service and to the enrolment in courses.

The estimation methods for RGC and RCR models rely on the standard assumption of non-informative censoring. Since the data used in the analysis were gathered with a retrospective survey, it is natural to assume random right-censoring. However some problems could derive from the sampling non-response, which causes a special type of left-censoring (Reardon *et al.*, 2001). In the survey on the Italian graduates, realized with the CATI techique, the total non-response rate was quite high (39.2%), but the consequencies on the estimates should not be as serious as might seem at first sight, since over 90% of the non-responses is due to missing contact (wrong telephone number, no response to the calls etc.). In the present application a likely consequence of non-response is the underestimation of the hazard curves, since we expect the probability of missing contact to be higher for an individual who works.

The modelling process started with the basic RGC model, fitted by maximum likelihood with 10-point gaussian quadrature, as implemented in MIXOR (Hedeker and Gibbons, 1996; Hedeker *et al.*, 2000). The model selection was based on the Wald test at 95% level, except for the variance parameters, for which we used the more appropriate likelihood ratio test with p-value correction (Snijders and Bosker, 1999). The final model, reported in Table 2, has only one random effect on the intercept, whose standard deviation is estimated by the algorithm. To fit the formally equivalent RCR model it is necessary to construct the person-period data set, which results in 83302 records. Since MIXOR does not handle such a large data set, we performed the estimation through the PQL2 method implemented in MLwiN (Goldstein *et al.*, 1998). The results, shown in Table 2, are very close to the previous ones; note that the estimated α_t's of the two models are approximately linked by formula (5), while the estimated random effects variance is about the square of the value reported by MIXOR

for the standard deviation. Therefore, in the present case, maximum likelihood with gaussian quadrature and PQL2 can be considered as equivalent.

We then extended the RGC model to account for non proportional effects, trying, for each variable, the inclusion of a set of 12 interaction parameters whose joint significance was assessed through the likelihood ratio test. The final model, reported in Table 3, includes non proportional effects only for the variable which denotes the completion of the military service in the observation period (MS-Done). Note that the main effect of MS-Done is not significant, while all the corresponding interaction parameters have negative values, with a magnitude that increases until the 6^{th} quarter and then decreases monotonically to reach non significant values in the last two quarters. This pattern suggests that the forced withdrawal from the job market due to the military service has a negative effect on the chances to get job, but this effect is temporary and vanishes in a few quarters, as is confirmed by the sample percentages of graduates who have found their first job by the date of the interview: 70.0% for the males who have been exempted from the service and 69.6% for the males who have performed the service during the observation period. A comparison with the results for the basic RGC model shows that the inclusion of the interaction parameters for the covariate MS-Done has produced only minor changes in the estimates of the other parameters.

Finally we developed the RCR model using the person-period data set. The first target was to represent the baseline hazard in a more parsimonious way through a polynomial specification. The model selection procedure led to the choice a cubic function of the time with the addition of a specific parameter for the first quarter, for a total of 5 parameters. Subsequently we tried the inclusion of the time interaction terms, defined as interactions between a fixed-time covariate and the covariates which denote the quarters and the powers of the quarters until the third. In the final model, reported in Table 4, there are 5 variables with time interactions, for a total of 10 additional parameters; note that, apart from the gender indicator, these variables refer to events occurred in the observation period. The effects of the other covariates show only minor changes with respect to the basic RCR model.

Therefore, in our application, the extended RGC and extended RCR models differ essentially in the representation of the hazard patterns: the extended RGC model uses 13 interval-specific parameters for the baseline hazard plus 12 interaction parameters to account for the non proportional effect of the covariate MS-Done; on the other hand, the extended RCR model employes 5 parameters to specify the baseline hazard plus 10 parameters to include the non proportional effects of five covariates. Figure 2 shows two estimated hazard curves for each model (the curves for the reference individual are obtained by assuming a null value for all the covariates and for the random effect). Obviously, the curves for the extended RCR model are smoother because of the parametric specification. Comparing, for the reference individual, the curves for the two models, we note that the curve for the extended RCR model tends to be higher in the first half of the observation period and lower in the second half: this pattern is due to the fact that the extended RCR model also adjusts for the

non proportional effects of the enrolement in and abandon of university and training courses, which exert their negative influence mainly in the first quarters.

Some further comments on the substantive results can be found in Grilli (1999).

5 Concluding remarks

While the RGC and RCR models are formally equivalent, the latter is more difficult to fit because it requires the contruction of a person-period data set which makes the estimation procedure less efficient in terms of computing time. However, in most practical applications it is necessary or convenient to extend the models to include time-varying covariates and/or non proportional effects: while the extended models can be fitted in the same manner as the basic ones, their properties are quite different. In particular, only the RCR model can be extended to include time-varying covariates, achieving a more parsimonious and accurate representation of the hazard patterns. In the analysis of the complex data set on the Italian graduates presented in Section 4 the superiority of the extended RCR model over the extended RGC model was manifest.

Replacing the complementary log-log link in the RCR model with the logit link produces the random effects version of the discrete time logit model proposed by Cox (1972), which has no direct relationship with the continuous time RPH model. It can be argued that the logistic model is the most natural choice when the time is truly discrete; however, nearly always the qualitative conclusions are unaffected by the link choice (Allison, 1982). A formal selection procedure could be devised using a generalized link function (e.g. Stukel, 1988).

A further interesting and straghtforward extention of the RCR model is obtained by including a random effect (often called frailty) at the individual level to represent unobserved eterogeneity among individuals. Using the person-period data set this extention is very simple, because it is sufficent to specify a three-level model in which the level-one units are the person-period observations, the level-two units are the individuals and the level-three units are the groups. In the analysis on the Italian graduates no significant unobserved eterogeneity among individuals was found.

References

[1] Allison P. D. (1982) Discrete-time methods for the analysis of event histories. In *Sociological Methodology* (ed. S. Leindhardt), pp. 61-98. San Francisco: Jossey-Bass.

[2] Barber J. S, Murphy S., Axinn W. G. and Maples J. (2000) Discrete-time multilevel hazard analysis. *Sociological Methodology*, **30**, 201-235.

[3] Biggeri L., Bini M., Grilli L. (2001) The transition from university to work: a multilevel approach to the analysis of the time to obtain the first job. *Journal of the Royal Statistical Society A*, **164**, 293-305.

[4] Cox, D. R. (1972) Regression models and life tables (with discussion). *Journal of the Royal Statistical Society B*, **34**, 187-220.

[5] Goldstein, H. (1995) *Multilevel Statistical Models*. London: Edward Arnold.

[6] Goldstein H., Rasbash J., Plewis I., Draper D., Browne W., Yang M., Woodhouse G. and Healy M. J. R. (1998) *A User's Guide to MLwin*. London: Institute of Education.

[7] Grilli L. (1999) *Sbocchi occupazionali e scelte formative dei diplomati: un'analisi multilivello*. PhD Thesis in Applied Statistics. Florence: Department of Statistics, University of Florence.

[8] Hedeker D. and Gibbons R. D. (1996) MIXOR: A computer program for mixed-effects ordinal regression analysis. *Computational Methods and Programs in Biomedicine*, **49**, 157-176.

[9] Hedeker D., Siddiqui O. and Hu F. B. (2000). Random-Effects Regression Analysis of Correlated Gruoped-Time Survival Data. *Statistical Methods in Medical Research*, **9**, 161-179.

[10] Italian National Statistical Institute (1999) *Percorsi di studio e di lavoro dei diplomati: Indagine 1998*. Informazioni n. 29, Rome: Italian National Statistical Institute.

[11] Kalbfleisch J. D. and Prentice R. L. (1973) Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60**, 267-278.

[12] Kalbfleisch J. D. and Prentice R. L. (1980) *The statistical analysis of failure time data*. New York: Wiley.

[13] McCullagh P. (1980) Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society B*, **42**, 109-142.

[14] Reardon S. F., Brennan R. and Buka S. L. (2001) *Estimating Multi-Level Discrete-Time Hazard Models Using Cross-Sectional Data: Neighborhood Effects on the Onset of Adolescent Cigarette Use*. Working Paper 01-07. University Park: Population Research Institute, The Pennsylvania State University.

[15] SAS Institute (1999) *SAS/STAT User's Guide Version 8*, Cary: SAS Institute Inc.

[16] Snijders T. A. B. and Bosker R. J. (1999) *An introduction to basic and advanced multilevel modeling*. London: Sage.

[17] Stukel T. A. (1988) Generalized Logistic Models. *Journal of the American Statistical Association*, **83**, 426-431.

[18] Vaida F. and Xu R. (2000) Proportional hazards model with random effects. *Statistics in Medicine*, **19**, 3309-3324.

List of Tables

1 Names, definitions and sample averages of the covariates 13
2 Estimates for the basic RGC and RCR models 14
3 Estimates for the extended RGC model 15
4 Estimates for the extended RCR model 16

List of Figures

1 Time to the first job: non parametric estimate of the hazard function 17
2 Time to the first job: estimated hazard functions 17

Table 1: Names, definitions and sample averages of the covariates

Name	Definition	Average
Female	1, female; 0, male	0.54
MS-Exempted*	1, exempted from military service; 0, otherwise	0.10
MS-Done	1, military service done by the date of the interview**; 0, otherwise	0.33
MS-ToBeDone	1, military service still to be done; 0, otherwise	0.03
FM36	1, final mark equal to 36; 0, otherwise	0.13
FM37-42	1, final mark from 37 to 42; 0, otherwise	0.40
FM43-49*	1, final mark from 43 to 49; 0, otherwise	0.29
FM50-59	1, final mark from 50 to 59; 0, otherwise	0.16
FM60	1, final mark equal to 60; 0, otherwise	0.02
OSP-Indep	number of parents that were independent workers (excluding businessmen and professionals)	0.31
OSP-Business	number of parents that were businessmen or professionals	0.08
EnrUniv	1, enrolled in a university course**; 0, otherwise	0.21
IntUniv-NoWork	1, abandoned university for reasons not linked to work**; 0, otherwise	0.10
IntUniv-Work	1, abandoned university because of work**; 0, otherwise	0.04
EnrTrC	1, enrolled in a regional training course**; 0, otherwise	0.19
EnrOtC	1, enrolled in a course different from a university or regional training course**; 0, otherwise	0.02
SchTP-Business*	1, received the high-school certificate in a technical/professional college - business type; 0, otherwise	0.31
SchTP-Industrial	1, received the high-school certificate in a technical/professional college - industrial type; 0, otherwise	0.20
SchTP-Other	1, received the high-school certificate in a technical/professional college different from business or industrial; 0, otherwise	0.27
SchGymnasium	1, received the high-school certificate in a gymnasium; 0, otherwise	0.07
SchOther	1, received the high-school certificate in a school different from technical/professional college or gymnasium; 0, otherwise	0.15
Unempl95	Unemployment rate (%) at regional level for people aged 14-25 in 1995 (variable centered at the value of Tuscany, 25.1)	11.14
Unempl95-98	Variation from 1995 to 1998 in the Unemployment rate (%) at regional level for people aged 14-25	-1.13
North	1, region of the north of Italy; 0, otherwise	0.40
Center*	1, region of the center of Italy; 0, otherwise	0.20
South	1, region of the south of Italy or islands; 0, otherwise	0.40

Sample size is 9404

The variables denoted by * are reference categories in model specification

The events denoted by ** have occurred at un unknown date between the end of the school and the interview

Table 2: Estimates for the basic RGC and RCR models

	RGC model		**RCR model**	
Parameter	*Estimate*	*Std. Err.*	*Estimate*	*Std. Err.*
α_1	-1.583	0.073	-1.580	0.073
$\alpha_2 - \alpha_1$	0.453	0.019	-0.559	0.055
$\alpha_3 - \alpha_1$	0.793	0.024	-0.451	0.055
$\alpha_4 - \alpha_1$	0.982	0.025	-0.776	0.064
$\alpha_5 - \alpha_1$	1.171	0.027	-0.589	0.062
$\alpha_6 - \alpha_1$	1.323	0.028	-0.640	0.065
$\alpha_7 - \alpha_1$	1.500	0.029	-0.318	0.060
$\alpha_8 - \alpha_1$	1.626	0.029	-0.506	0.066
$\alpha_9 - \alpha_1$	1.767	0.029	-0.261	0.063
$\alpha_{10} - \alpha_1$	1.963	0.031	0.237	0.055
$\alpha_{11} - \alpha_1$	2.122	0.031	0.209	0.059
$\alpha_{12} - \alpha_1$	2.277	0.032	0.341	0.060
$\alpha_{13} - \alpha_1$	2.422	0.033	0.422	0.062
Female	-0.345	0.045	-0.346	0.048
MS-Done	-0.257	0.044	-0.258	0.047
MS-ToBeDone	-0.458	0.116	-0.459	0.116
FM36	-0.144	0.047	-0.144	0.047
FM37-42	-0.114	0.033	-0.114	0.033
FM50-59	0.091	0.040	0.092	0.042
FM60	0.257	0.099	0.258	0.096
OSP-Indep	0.087	0.025	0.088	0.025
OSP-Business	0.257	0.043	0.258	0.047
EnrUniv	-1.666	0.105	-1.669	0.105
IntUniv-NoWork	0.956	0.119	0.957	0.114
IntUniv-Work	1.720	0.118	1.722	0.119
EnrTrC	-0.440	0.042	-0.441	0.037
EnrOtC	-0.666	0.112	-0.669	0.114
SchTP-Industrial	0.135	0.048	0.135	0.048
SchTP-Other	-0.211	0.044	-0.211	0.042
SchGymnasium	-0.256	0.072	-0.255	0.071
SchOther	-0.443	0.053	-0.444	0.053
Unempl95	-0.026	0.003	-0.026	0.002
Unempl95^2	0.00019	0.00008	0.00019	0.00008
Unempl95-98	-0.022	0.007	-0.022	0.006
North	0.130	0.051	0.131	0.047
South	-0.246	0.069	-0.246	0.066
σ_u (RGC) or σ_u^2 (RCR)	0.257	0.023	0.068	0.011
Number of records	9404		83302	
Estimation method	ML (Gaussian quad.)		PQL2	

Table 3: Estimates for the extended RGC model

Parameter	*Estimate*	*Std. Err.*
α_1	-1.654	0.077
$\alpha_2 - \alpha_1$	0.538	0.027
$\alpha_3 - \alpha_1$	0.918	0.033
$\alpha_4 - \alpha_1$	1.128	0.035
$\alpha_5 - \alpha_1$	1.329	0.037
$\alpha_6 - \alpha_1$	1.489	0.038
$\alpha_7 - \alpha_1$	1.654	0.038
$\alpha_8 - \alpha_1$	1.774	0.038
$\alpha_9 - \alpha_1$	1.887	0.039
$\alpha_{10} - \alpha_1$	2.055	0.040
$\alpha_{11} - \alpha_1$	2.180	0.041
$\alpha_{12} - \alpha_1$	2.322	0.041
$\alpha_{13} - \alpha_1$	2.456	0.042
MS-Done	-0.074	0.072
γ_2*MS-Done	-0.227	0.039
γ_3*MS-Done	-0.342	0.049
γ_4*MS-Done	-0.405	0.052
γ_5*MS-Done	-0.437	0.055
γ_6*MS-Done	-0.462	0.057
γ_7*MS-Done	-0.424	0.059
γ_8*MS-Done	-0.403	0.060
γ_9*MS-Done	-0.323	0.060
γ_{10}*MS-Done	-0.245	0.062
γ_{11}*MS-Done	-0.147	0.063
γ_{12}*MS-Done	-0.108	0.065
γ_{13}*MS-Done	-0.074	0.065
Female	-0.336	0.046
MS-ToBeDone	-0.451	0.119
FM36	-0.149	0.047
FM37-42	-0.113	0.034
FM50-59	0.091	0.041
FM60	0.248	0.101
OSP-Indep	0.088	0.025
OSP-Business	0.259	0.043
EnrUniv	-1.648	0.106
IntUniv-NoWork	0.928	0.120
IntUniv-Work	1.690	0.120
EnrTrC	-0.442	0.043
EnrOtC	-0.656	0.113
SchTP-Industrial	0.146	0.048
SchTP-Other	-0.208	0.044
SchGymnasium	-0.250	0.073
SchOther	-0.435	0.054
Unempl95	-0.026	0.003
Unempl95^2	0.00018	0.00008
Unempl95-98	-0.022	0.007
North	0.125	0.051
South	-0.254	0.069
σ_u	0.255	0.023
Number of records	9404	
Estimation method	ML (gauss. quad.)	

Table 4: Estimates for the extended RCR model

Parameter	*Estimate*	*Std. Err.*
Intercept	-1.627	0.108
Quarter (from 0 to 12)	-0.160	0.056
Quarter^2	0.016	0.010
Quarter^3	-0.00019	0.00055
1^{st} quarter indicator	0.242	0.083
Female	-0.479	0.068
Female*Quarter	0.030	0.011
MS-Done	-0.265	0.079
MS-Done*Quarter	-0.497	0.058
MS-Done*Quarter^2	0.114	0.012
MS-Done*Quarter^3	-0.006	0.001
IntUniv-NoWork	-0.569	0.246
IntUniv-NoWork*Quarter	0.707	0.143
IntUniv-NoWork*Quarter^2	-0.096	0.026
IntUniv-NoWork*Quarter^3	0.004	0.001
IntUniv-Work	1.335	0.148
IntUniv-Work*Quarter	0.061	0.015
EnrTrC	-1.097	0.095
EnrTrC*Quarter	0.162	0.036
EnrTrC*Quarter^2	-0.006	0.003
MS-ToBeDone	-0.419	0.115
FM36	-0.149	0.047
FM37-42	-0.114	0.033
FM50-59	0.085	0.042
FM60	0.245	0.095
OSP-Indep	0.089	0.025
OSP-Business	0.269	0.047
EnrUniv	-1.622	0.105
EnrOtC	-0.640	0.114
SchTP-Industrial	0.144	0.047
SchTP-Other	-0.204	0.041
SchGymnasium	-0.251	0.071
SchOther	-0.432	0.052
Unempl95	-0.026	0.002
Unempl95^2	0.00019	0.00008
Unempl95-98	-0.022	0.006
North	0.124	0.046
South	-0.252	0.065
σ_u^2	0.059	0.011
Number of records	83302	
Estimation method	PQL2	

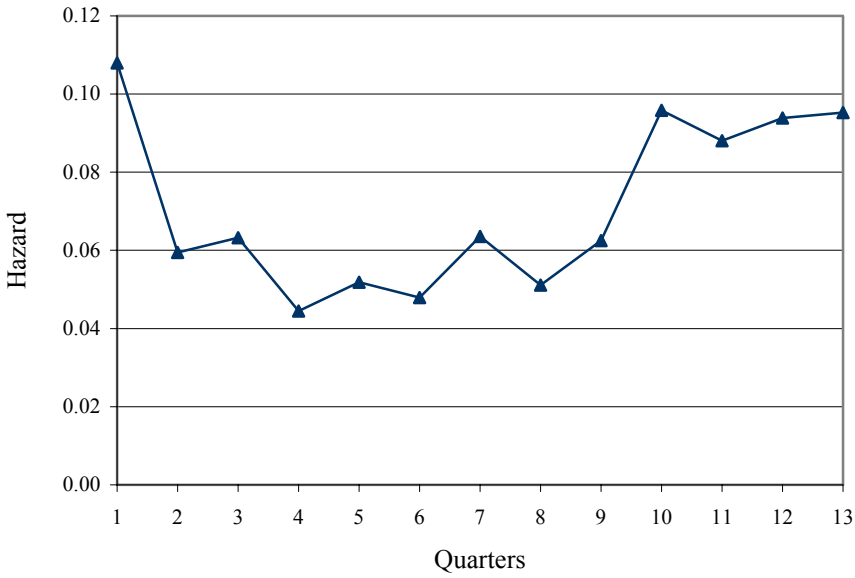

Figure 1: Time to the first job: non parametric estimate of the hazard function

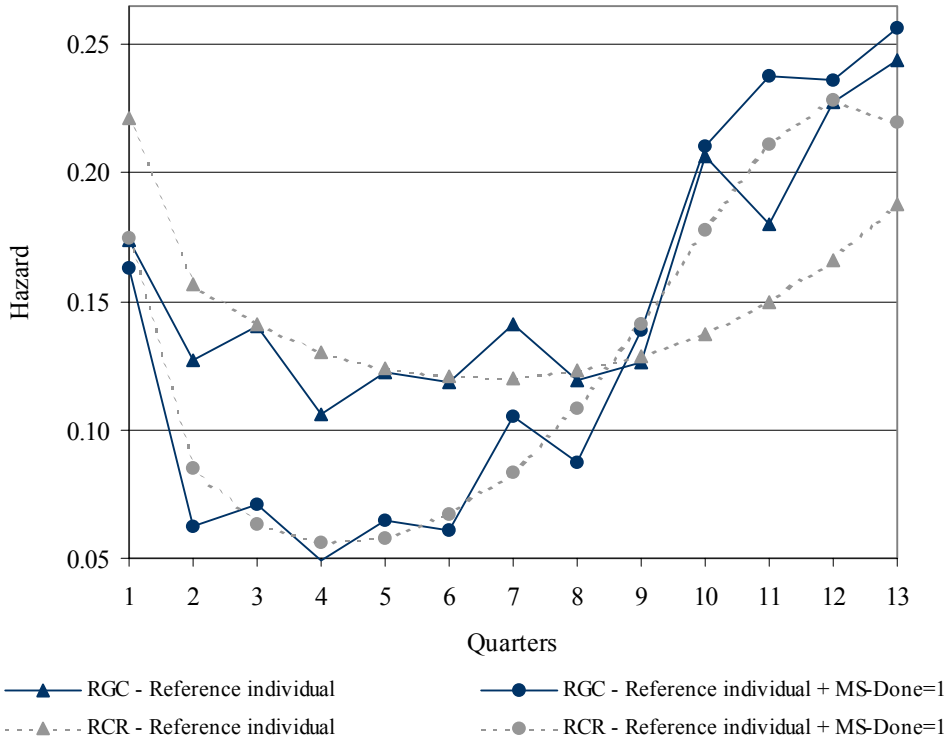

Figure 2: Time to the first job: estimated hazard functions