1. Introduction

The analysis of the transition from school or university to work plays a crucial role in the evaluation of higher education institutions, since one of their aim is to endow the graduates with appropriate skills to be spent in the labour market. The analysis can focus on various aspects, e.g. the employment status at a given date, the time needed to get the first job, the degree of consistency between the current job and the curriculum, the satisfaction for the current job. The definition of the outcome to be studied depends on the purpose of the evaluation and ultimately on the policy objectives: for example, in the last few years there has been a tendency to narrow the definition of “employed graduate” by considering only jobs which are permanent or consistent with the curriculum. Moreover, a key point in the analysis of employment rates is the definition of the denominator, i.e. the criteria to identify the subset of graduates interested in finding a job (this is crucial to compare degree programmes whose graduates have markedly different propensities to keep on studying in contrast to search for job).

Whichever the outcome of interest, the phenomenon has a relevant hierarchical structure (e.g. graduates nested in degree programmes nested in universities and so on). Such a structure
calls for multilevel modelling, which allows to perform the analysis at all the hierarchical levels at the same time. Multilevel models are especially suited for comparing effectiveness (Goldstein & Spiegelhalter, 1996) as the relative effectiveness of an institution is explicitly represented in the model as a random effect (a latent variable at the institution level). Even if multilevel models represent a theoretically satisfactory tool for the assessment of educational institutions, their implementation must face serious problems such as misspecification due to omitted variables, measurement error bias, and low power in ranking the institutions.

In the analysis of the transition from school or university to work, survival models for the time to obtain the first job are potentially more informative than logit/probit models for the employment status at a given date, since they give insight also into the dynamics of the process, allowing explanatory variables with time-varying effects and even time-varying explanatory variables.

The time to obtain the first job is usually reported in months so it requires discrete-time survival models, as continuous-time models are inadequate owing to the large number of ties. Discrete-time survival models and multilevel models each have a long history, but their joint use is quite recent (Barber, Murphy, Axinn & Maples, 2000).

In this contribution I review my work on multilevel discrete-time survival models for the time to obtain the first job, making reference to Biggeri, Bini & Grilli (2001) and Grilli (2005). Other studies on graduates labour are in Grilli & Rampichini (2006a), a multilevel polytomous response model for studying where the skills needed for the current job have been acquired, and Grilli & Rampichini (2006b), a multilevel factor model for studying the satisfaction on various aspects of the current job.

This contribution is organized as follows. Section 2 outlines the main concepts on effectiveness, while Section 3 introduces multilevel models, stressing their role in the assessment of effectiveness. Section 4 introduces discrete-time survival models, while Section 5 deals with their multilevel extension. Finally, Section 6 reviews an application of the described methodology to the analysis of the time to obtain the first job for Italian graduates.
2. Effectiveness

The effectiveness of an organization is the degree of achievement of its institutional targets. In the case of education (schools, universities) some targets are internal, such as the attainment of an adequate level of knowledge, while other targets are external, such as a high proportion of employed graduates.

The degree of achievement of the targets can be measured in absolute terms (absolute effectiveness or impact analysis) or in relative terms (relative or comparative effectiveness). Absolute effectiveness is appropriate for the evaluation of interventions, e.g. a specific vocational training course, while relative effectiveness is suited for situations with many institutions offering the same service and thus interest focus on comparing the institutions. In a comparative setting, the effectiveness is usually operationalized as a measure of performance adjusted for the factors out of the control of the institution. In other words, the effectiveness is seen as an extra-performance entirely due to the behaviour of the institution itself.

In the following I will only consider comparative effectiveness on external targets concerning the labour market outcomes of the graduates.

In terms of economic theory, the issue of comparative effectiveness can be viewed through the Principal-Agent-User model (Fabbri, Fazioli & Filippini, 1996). In the context of education, the Principal is the Ministry of Education, the Agents are the educational institutions (schools or universities) and the Users are the students. The subjects are in a situation of asymmetric information and need some kind of assessment of the service offered by the Agents: in fact, each User must choose one Agent (the best for her), while the Principal must rank the Agents in terms of effectiveness in order to understand the good practices and to take actions to improve effectiveness (e.g. assigning incentives).

The key point is that the quality of the output of the educational process cannot be defined in absolute terms, but only with respect to the effects on the students. Therefore, education is a field where the evaluation of the Agents must be adjusted for the features of the Users. This need is widely recognized in the literature on value-added student achievement, where the main methodological point is how to properly adjust the final raw achievement for the initial conditions (initial level of knowledge, motivation, social and economical status…). The issue of adjustment is crucial also in external effectiveness evaluations (employment chances, consistency between job and curriculum), but in such cases the adjustment for a ceteris paribus comparison is
even more difficult: in fact, there is no initial measure of the outcome under study and the external nature of the result requires adjusting also for external conditions (e.g. the unemployment rate). In essence, to achieve a fair evaluation the main difficulty is to make a proper adjustment.

A further point is that the kind of adjustment required for assessing effectiveness is not the same for the various subjects interested in the results. In this regard, it is useful to distinguish between two types of effectiveness. In fact, a potential student (User) and the Ministry of Education (Principal) are interested in different types of effectiveness of the educational institutions (Agents):

- **Type A - Potential student**: interested in comparing the results she can obtain by enrolling in different institutions, irrespective of the way such results are yielded;
- **Type B - Ministry of Education**: interested in assessing the “production process” in order to evaluate the ability of the institutions to exploit the available resources.

The two types of effectiveness are called A and B after Raudenbush & Willms (1995), who focused on internal measures, but the concept naturally extends to external measures. In a comparative setting the effectiveness is usually assessed through a measure of performance adjusted for the factors out of the control of the institution, so the difference between Type A and Type B effectiveness simply lies in the kind of adjustment:

- **Type A effectiveness**: performance of the Agent adjusted for the features of its Users;
- **Type B effectiveness**: performance of the Agent adjusted for the features of its Users, the features of the Agent itself (out of its control) and the context in which it operates.

In the evaluation of schools or universities the features of the students to adjust for are the initial knowledge, ability, motivation etc., or proxies easier to measure, such as the socio-economic status. Examples of features of the institutions to adjust for are the public or private status, the student/teacher ratio and the amount of funding. The features of the context requiring adjustment depend on the kind of evaluation, for example to assess the effectiveness in terms of chances of employment an adjustment should be made for the conditions of the local labour market.

As pointed out by Raudenbush & Willms (1995), in practice the adjustment required for the assessment of Type B effectiveness is not easy, as it involves many variables whose measurement is problematic.
Regardless of the kind of adjustment, any assessment of effectiveness must face the issue of the accidental variability in the outcome due to sources other than the effectiveness. Common sources of accidental variability are: fluctuations in the unobserved features of the Users, sampling error, measurement error. As usual, the amount of accidental variability should be carefully estimated and taken into account when comparing the Agents, in order to avoid results that do not reflect actual differences in effectiveness.

A broad review of the methodological and statistical issues connected with performance indicators is Bird, Cox, Farewell, Goldstein, Holt & Smith (2005).

3. Multilevel models as a tool for measuring effectiveness

The statistical models for assessing the relative effectiveness of educational institutions must face two main problems:

- Adjustment: the measures must be adjusted at least for the features of the students, since this is necessary for a fair (ceteris paribus) comparison;
- Quantification of uncertainty (accidentally variability): this is necessary in order to formulate judgements strongly supported by the empirical evidence, avoiding judgements that may be originated by the sampling variability or other sources of error.

The raw rankings, sometimes called ‘League Tables’, ignore both issues (Goldstein & Spiegelhalter, 1996).

The main statistical tool for making a proper adjustment, while quantifying uncertainty, is regression. However, in a comparative evaluation of educational institutions, standard regression models (such as the Generalized Linear Models) are not adequate as they do not take into account a crucial feature of the problem, namely the hierarchical structure. In fact, the students/graduates are nested into the institutions and, while the aim is to measure the effectiveness of the institutions, the outcomes are defined at the student level. From a statistical viewpoint, standard regression models make unsuitable assumptions on the variance-covariance structure since they assume independence of the observations, while the results of the students/graduates of the same institution are positively correlated as they share several unobserved factors at the institution.
level. The consequence is a poor quantification of uncertainty (and in nonlinear models also a systematic attenuation of the estimates of the regression coefficients).

Multilevel models are well suited for assessing the relative effectiveness of institutions, in particular they allow to

- specify distinct sub-models for the behaviour of the institutions and the behaviour of their users;
- represent adequately the variance-covariance structure, a necessary step to achieve a good quantification of the uncertainty;
- represent explicitly the concept of effectiveness by means of a random effect added to the linear predictor.

Considering for simplicity the linear specification (apt for a continuous outcome variable), a multilevel model can be written as (Snijders & Bosker, 1999; Goldstein, 2003)

$$Y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + \gamma_{01} w_j + u_j + e_{ij},$$

where $j$ indexes the level 2 units (clusters) and $i$ indexes the level 1 units (subjects). In terms of the Principal-Agent-User model, the clusters are the Agents (i.e. the institutions under evaluation) and the subjects are the Users. The variables in the model are:

- $Y_{ij}$, the outcome variable of User $i$ of Agent $j$ (a raw measure of performance);
- $x_{ij}$, a vector with the features of User $i$ of Agent $j$;
- $w_j$, a vector with the features of Agent $j$ and the context in which it operates.

Then, $u_j$ is the random effect of Agent $j$, i.e. an unobservable quantity characterizing such an Agent and shared by all its Users. The term $u_j$ is an adjusted measure of performance: in fact, it is a residual component that captures all the relevant factors at the Agent level not accounted for by the covariates and thus its meaning depends on which covariates enter the model. The effect $u_j$ is called “random” because it is a random variable, assuming independence among the Agents. For consistency of the estimates, the crucial assumption is that the expectation of $u_j$ conditionally on the covariates is null (exogeneity). Less crucial, but standard assumptions are the homoscedasticity, i.e. $u_j$ has constant variance $\sigma_u^2$, and the normality of the distribution.
Finally, the level 1 errors $e_{ij}$ are residual components taking into account all the unobserved factors at the User level that render the outcome $Y_{ij}$ different from what predicted by the covariates and the random effect. The $e_{ij}$ are assumed independent among Users and independent of $u_i$. The other standard assumptions are similar to those on $u_i$, i.e. exogeneity, homoscedasticity (with variance denoted as $\sigma_e^2$) and normality.

The parameters of a linear multilevel model can be easily estimated with Maximum Likelihood through algorithms such as IGLS, Fisher scoring and EM (Snijders & Bosker, 1999; Goldstein, 2003).

A multilevel version of a Generalized Linear Model (e.g. a logit model) is obtained simply by adding a random effect to the linear predictor. However, outside the linear realm, multilevel models sometimes become quite complex, in terms of both interpretation and estimation.

4. Discrete-time survival models

An outcome variable often considered in the analysis of external effectiveness of schools or universities is the time elapsed from graduation to the beginning of the first job. While the time of graduation can be easily obtained from administrative archives, the beginning of the first job usually can be determined only through a survey and requires to choose among a set of possible definitions of “first job”. The survey can be retrospective (a single interview) or longitudinal (repeated interviews), each with well-known pros and cons.

In any case, the time elapsed from graduation to the beginning of the first job has two important features:

- the time is discrete, as it is recorded in months or larger time units, so many graduates have the same recorded time (so-called ties);
- the time is censored for those who did not get the first job by the end of the observation period.

The appropriate statistical tool is thus a discrete-time survival model (e.g. Allison, 1982; Singer & Willet, 2003), where the time for subject $i$ is represented by a discrete random variable $T_i$ assuming values in the positive integers. In survival modelling, it is convenient to represent the distribution by means of the hazard (risk) function:
At a given time point $t$, the hazard is the probability of experiencing the event of interest at time $t$ conditional on being still at risk. In the setting of job search, the hazard is the probability of getting the first job in month $t$ given that it was not yet obtained.

Inference methods for survival analysis allow for right censoring. The time of a subject is right censored at $t$ if the observation period of that subject ends at $t$ before experiencing the event of interest, so the observation is not $T_i \leq t$, but rather $T_i > t$. The end of the observation period may be determined by the subject itself or by the design of the survey. In a retrospective survey the observation period is ended by design at the day of the interview.

With right censoring, the observation of subject $i$ is represented with a couple $(t_i, d_i)$, where $t_i$ is the time recorded and $d_i$ is an indicator of the occurrence of the event of interest. Thus $d_i = 1$ means that $t_i$ is uncensored ($T_i = t_i$), while $d_i = 0$ means that $t_i$ is censored ($T_i > t_i$). To make an example in the setting of graduates’ job, the record $(t_i = 42, d_i = 1)$ says that graduate $i$ got job after 42 months, while the record $(t_i = 42, d_i = 0)$ says that graduate $i$ did not get job within 42 months.

Censored times cannot be deleted, nor treated as uncensored times. The standard estimation methods for survival models use the censored times under the assumption of non-informative censoring. Informally, censoring is non informative when, conditionally on the observed covariates, the end of the observation period does not depend on the hazard. This condition is usually met in retrospective surveys, while it likely to be more or less violated in longitudinal surveys.

Statistical inference is based on the likelihood. The contribution of subject $i$ to the likelihood is different if the time is uncensored or censored:

\[
h_i(t) = P(T_i = t | T_i \geq t) \quad .
\]

\[
\text{Uncensored (}d_i = 1): \quad P(T_i = t_i) = \prod_{u=1}^{t_i-1} [1 - h_i(u)] \times h_i(t_i)
\]

\[
\text{Censored (}d_i = 0): \quad P(T_i > t_i) = \prod_{u=1}^{t_i} [1 - h_i(u)] \quad .
\]
For example, the record \((t_i = 3, d_i = 1)\) leads to \(P(T_i = 3) = (1 - h_i(1))(1 - h_i(2))h_i(3)\). Since the building blocks of the likelihood are the hazards, a survival model is just a model for the hazard function.

To adjust for covariates one can specify a functional form for the conditional hazard

\[
h_i(t | x_{it}) = P(T_i = t \mid T_i \geq t, x_{it}),
\]

where the vector \(x_{it}\) includes all the covariates of subject \(i\) at time \(t\). The covariates can be time-invariant or time-varying. Time-varying covariates are extremely useful in building a proper model for the hazard, but they are rarely available in practice because of the difficulty of an accurate measurement, especially in retrospective surveys. A classical example is the military service, which is time-varying and strongly affects the occupational chances, especially when done after the degree: however, most surveys do not ask for the beginning and end of the service period, so only time-invariant covariates are available (e.g. service done before the degree versus after the degree, exempted from the service,…).

Since the hazard function is bounded between 0 and 1, a linear model for the hazard itself is not suitable, but one can apply a linear model to an appropriate transformation of the hazard:

\[
g(h_i(t | x_{it})) = \alpha_i + x_{it}'\beta,
\]

where the transformation \(g\), called link function, maps the \((0,1)\) interval onto the real line. On the right-hand side, \(\beta\) is the vector of regression coefficients and \((\alpha_1, \alpha_2, \ldots, \alpha_P)\) are time-specific intercepts representing the baseline hazard, i.e. the hazard for the hypothetical subject with all the covariates set to zero. The number of time-specific intercepts is \(P\), the maximum number of time points (intervals) in the data.

When the link \(g\) is the logit function \(\log\left(\frac{x}{1-x}\right), x \in (0,1)\), the corresponding model is called logit or proportional odds

\[
\log\left(\frac{h_i(t | x_{it})}{1 - h_i(t | x_{it})}\right) = \alpha_i + x_{it}'\beta,
\]
or, in terms of the hazard function,

\[ h_i(t \mid x_{it}) = \frac{1}{1 + \exp(-\alpha_i - x_{it}' \beta)} \]

The interpretation of the regression coefficients requires some care, since \( \beta_k \) is the change in the logit of the hazard following a unit increase in the \( k \)-th covariate.

When the maximum number of time points \( P \) is high, it is advisable to assume a smooth functional form for the time-specific intercepts, e.g. a polynomial of degree \( R \):

\[ \alpha_i = \sum_{r=0}^{R} \gamma_r t^r \]

thus replacing the \( P \) parameters \( (\alpha_1, \alpha_2, \ldots, \alpha_P) \) with the \( R+1 \) parameters \( (\gamma_0, \gamma_1, \ldots, \gamma_R) \).

Another key feature of survival analysis is the study of the dynamics of the covariates’ effects. In fact, a time-invariant covariate may have a time-varying effect, for example the hazards of getting job for males and females may be rather different immediately after the degree, but rather similar after three years. In a model with polynomial baseline hazard, to let a time-invariant covariate \( x_k \) have a time-varying effect is sufficient to insert interactions with time, i.e. \( x_k t \), \( x_k t^2 \), \( x_k t^3 \) …. To test if the effect is time-invariant amounts to test if the coefficients of the interactions with time are jointly null.

Estimation can be carried out using standard software for binary response models. In fact, the likelihood of a discrete-time survival model on the original dataset is the same as the likelihood of a binary response model on the person-period dataset. To obtain the person-period dataset, each original record \( i \) is replicated as many times as the observed time \( t_i \) and the new response variable is the indicator of the event of interest. For example, the record of a subject experiencing the event of interest at time 5, i.e. \( t_i = 5, d_i = 1 \), is replicated 5 times and the values of the new response variable are (0,0,0,0,1). Also for a subject censored at time 5, i.e. \( t_i = 5, d_i = 0 \), the record is replicated 5 times, but the values of the new response variable are (0,0,0,0,0).
5. Multilevel discrete-time survival models

A relevant aspect of the external effectiveness of an educational institution is the time needed by its graduates in order to obtain the first job. The assessment of this kind of effectiveness requires the use of statistical models lying at the intersection between discrete-time survival models and multilevel models: Barber, Murphy, Axinn & Maples (2000), Hedeker, Siddiqui & Hu (2000), Rabe-Hesketh, Yang & Pickles (2001), Reardon, Brennan & Buka (2002), Steele, Goldstein & Browne (2004). Multilevel discrete-time survival models are increasingly used in job search analysis: Biggeri, Bini & Grilli (2001), Grilli (2005), Paccagnella (2006), Windzio (2006).

Focusing on the proportional odds model with polynomial baseline hazard, the three-level random intercept version is written as follows

\[
\text{logit} h_{ijk}(t | x_{ijkt}, u_{jk}, v_k) = \sum_{r=0}^{R} \gamma_r t^r + x_{ijkt}' \beta + u_{jk} + v_k ,
\]

where \(i\) indexes level 1 units (e.g. graduates), \(j\) indexes level 2 units (e.g. course programmes), and \(k\) indexes level 3 units (e.g. universities). The standard assumptions on the random effects are: (i) the random effects \(u_{jk}\) are iid across level 2 units, while the random effects \(v_k\) are iid across level 3 units; (ii) random effects at different levels are independent; (iii) the random effects have zero mean for every value of the covariates (exogeneity assumption). Barber, Murphy, Axinn & Maples (2000) discuss the assumptions in detail.

As for the distribution of the random effects, the two most common choices are: (a) Gaussian distribution; (b) discrete distribution with unknown locations and mass points. With the first option the model is fully parametric and the random effects of a given level have an unknown variance to be estimated. For random effects with discrete distribution, the number of mass points can be treated as fixed or can be estimated: in the first case the model is called latent class or mixture, while the second case leads to a semi-parametric model estimable through Non Parametric Maximum Likelihood. See Heckman & Singer (1984), Vermunt (1997), Muthén & Masyn (2005), Skrondal & Rabe-Hesketh (2004).

In survival models a crucial task is to properly account for the unobserved heterogeneity, i.e. the variance not explained by the covariates, since it causes a duration bias. The model with
random effects controls for unobserved heterogeneity at the course programme and university levels. A possible residual unobserved heterogeneity at the graduate level can be adjusted for by adding a further random effect at level 1.

6. Application: time to obtain the first job for Italian graduates

To illustrate the methodology of multilevel discrete-time survival models, I briefly review the application in Biggeri, Bini & Grilli (2001). The aim of their analysis is to characterize the factors that affect the transition from university to work (e.g. gender, age, socio-economic status, curriculum studiorum) as well as to evaluate the effectiveness of universities and course programmes with respect to the labour market outcomes of their graduates (external relative effectiveness).

The dataset (13511 interviewed graduates) comes from a retrospective survey on job opportunities of the 1992 Italian graduates conducted by the Italian National Statistical Institute (Istat) by means of a postal questionnaire in December 1995. The observation period thus ranges from 37 to 48 months after graduation and it is possible to determine, for each graduate, the time needed to get their first job (from 1 to 48 months) or the censored time for those still unemployed at the date of the interview. However, in order to reduce the dimension of the person-period dataset to be constructed, the analysis is conducted with the time collapsed into quarters (from 1 to 16). A sensitivity analysis shows that the collapsing has minor consequences.

The outcome variable is the time to get the first job. Unfortunately, the data do not allow to distinguish between temporary and permanent jobs, or part-time and full-time jobs (this classification is available only with respect to the job held at the date of the interview, which is not necessarily the first one). Such an outcome variable has several limitations. Mainly, in the light of the high flexibility of the youth labour market it would be important to take into account the features of the job, in particular the quality (position, salary, potentialities…) and the consistency with the curriculum. In addition, an observation period of 3-4 years after graduation is likely to be too short for a fair comparison among the course programmes, since in some cases, such as medicine or law, the graduates usually keep on studying for several years (though, on the other hand, policy makers cannot wait too much for the results of the evaluation).

The dataset of 13511 records has been reduced to 10338 records by eliminating the
graduates who, at the date of the interview, (i) had the same job as before the degree or (ii)
declared that they were not interested in getting job. The exclusion of the graduates not interested
in getting job aims at increasing the comparability of the course programmes. However, it is not
quite fair anyway to compare the (many) graduates in economics looking for job immediately
after the degree with the corresponding (few) graduates in law. In addition, the exclusion of the
graduates not interested in getting job has some (almost inevitable) limitations: (a) the condition
is established on the basis of a declaration of the graduate and thus subject to measurement error;
(b) even if the condition is likely to change during time, it is measured only at the date of the
interview; (c) the condition may be endogenous, i.e. correlated with the chances of getting job,
for example in the case of “discouraged” job seekers or in the case of graduates who looked for
job in vain for a couple of years and then decided to keep on studying.

The 10338 graduates of the subsample of interest are nested in 766 course programs
which are grouped into 64 universities, so the dataset has a hierarchical 3-level structure. Here a
course program is characterized by a subject/university combination, so that, for example,
Economics in Florence is distinguished from Economics in Pisa. Another possibility would be to
adopt a cross-classified structure, crossing the type of course programme with the university.
However, the curricula of a given type of course programme are quite different among the
universities, so a nested structure seems preferable.

The available explanatory variables include demographic features, family background,
educational history and work experience. They are all measured at the graduate level and at a
fixed time point (i.e. they are level 1 time-invariant covariates), so the potentialities of the
multilevel survival model are only partially exploited.

The data are analyzed by means of the 3-level logit survival model of equation (1) with
Gaussian random effects. In order to perform estimation using algorithms for binary response
models, the person-period dataset is generated (71143 records). The estimation method employed
is the quasi-likelihood PQL2 method of MlwiN, but many other programs can do this task
(usually with higher computational times, sometimes with higher accuracy). The model selection
is based on a backward strategy with 95% Wald tests.

The final model has only two random effects (one for course programmes and one for
universities), both on the intercept. No significant random slopes are found. Also, the addition of
a further random term at the graduate level to account for unobserved heterogeneity does not significantly improve the fit of the model.

In model (1) a positive regression coefficient means that an increment of the covariate is associated with an increment of the hazard of getting job. However, the magnitude of the effect is not evident from the value of the regression coefficient since the hazard is a function and, in addition, it is on the logit scale. Therefore, after reporting the estimates of the coefficients, it is important to show the impact of the covariates on the hazard function using numerical examples and graphs.

With the exception of the final mark (whose values can be all the integers between 66 and 110), all the covariates in the final model are binary, with the 0 category as the reference one.

I first report the estimated regression coefficients for the covariates with a time-invariant effect in the final model. In this case the regression coefficient is the change in the logit of the hazard (at any time point \(t\)) following a unit increase in the covariate, all other covariates held fixed.

The covariates with a positive time-invariant effect are

- **FINAL MARK** (first quartile 100, median 106, third quartile 110; centered on 100): +0.0063 for males and +0.0165 for females (this is the only covariate having an interaction with the gender);
- **INSTITUTIONAL TIME** (1= degree obtained within the institutional time established for the course): +0.1157;
- **OCCUPATIONAL STATUS WHILE ATTENDING UNIVERSITY** (1= graduate held at least one job during university studies): +0.1874;
- **EDUCATIONAL LEVEL OF THE PARENTS** (1= at least one with secondary school certificate or degree): +0.0913;
- **OCCUPATIONAL STATUS OF THE PARENTS** (1= at least one working): +0.1129.

The only covariate with a negative time-invariant effect is:

- **AGE AT DEGREE** (1= over 30 years): −0.1258.
In the final model the degree of the polynomial for the baseline hazard is $R=3$. Replacing the parameters with their estimates, the *predicted baseline hazard* is

$$
\text{logit } \hat{h}_{ijk}(t \mid x_{ijk} = 0, u_{jk} = 0, v_k = 0) = -1.70 + 0.04t - 0.02t^2 - 0.0001t^3 ,
$$

where $t$ is the time in quarters, coded with the integers from 0 to 15. Baseline refers to a graduate with all the covariates set to zero and who belongs to a mean course programme ($u_{jk} = 0$) in a mean university ($v_k = 0$). For example, in the first quarter ($t = 0$) the hazard of getting the first job for the baseline graduate is $-1.70$ on the logit scale, corresponding to a probability of 15.4%. Computing the predicted hazard for every admissible $t$ reveals that, for the baseline graduate, the hazard of getting job slightly increases in the second quarter and then decreases monotonically as time elapses, as shown by the circles in Figure 1. Actually, the hazard has the same pattern for any graduate who differs from the baseline only for covariates with time-invariant effects.

![Figure 1 – Estimated hazard functions: ●, males who did military service before their degrees or were exempted from it; □, females; ▲, males who did military service after their degrees (the other covariates and the random effects were set to zero). Reprinted from Biggeri, Bini & Grilli (2001).](image-url)
Two covariates have time-varying effects in the final model, namely FEMALE (1=female) and MILITARY SERVICE (1= done after degree, 0= done before degree or exempted). Their contribution to the linear predictor of model (1) is

$$\text{FEMALE: } \ldots -0.32 x_{jk} - 0.08 x_{jk} t + 0.006 x_{jk} t^2 + \ldots$$

$$\text{MILITARY SERVICE: } \ldots -1.22 x_{jk} - 0.47 x_{jk} t + 1.18 x_{jk} t^2 - 0.01 x_{jk} t^3 + \ldots$$

The estimates for the covariates with time-varying effects must be interpreted with care. For example, the estimate –0.32 for FEMALE merely refers to the difference between females and males in the first quarter. To appreciate the temporal evolution of such difference it is necessary to compute the effect for each $t$ and plot the predicted hazard functions for the two genders. A comparison between the circles and the squares of Figure 1 allows to appreciate the magnitude and evolution of the disadvantage of the females with respect to the males without military service. It is worth to note that such difference diminishes as time elapses and vanishes at the end of the observation period.

The predicted hazard function for the males involved with the military service after their degree (MILITARY SERVICE), represented by the triangles in Figure 1, shows a complex pattern. This pattern seems sensible, but its interpretation is quite contrived: in fact, the military service should be a time-varying covariate, but with the available data it can only be treated as a time-invariant covariate with a time-varying effect.

The estimated standard deviation of the random effect $v_k$ at the university level is 0.3444, while the estimated standard deviation of the random effect $u_{jk}$ at the course programme level is 0.4787. Such estimates, that are both statistically significant, indicate that there is more unexplained variability at the course programme level than at the university level. Given that the random effects are assumed to be normally distributed, multiplying the estimated standard deviation by 2 yields the difference, on the logit scale, between an “average” unit (i.e. on the 50th percentile) and a “high” unit (i.e. on the 97.5th percentile). For example, if graduates A and B have the same values in the observed covariates and obtained their degrees in the same university, but A comes from an “average” course programme, while B comes from a “high” course programme, then on the logit of the hazard of getting job the difference between B and A
is $2 \times 0.4787 = 0.9575$. To get an idea of how large such difference is, imagine that A is the baseline graduate, whose probability of getting job in the first quarter is 15.4%; then, the corresponding probability for B would be 32.2%.

The empirical Bayes residuals at levels 2 and 3 are used to assess the effectiveness of the course programmes and universities with respect to job opportunities. Since the model only controls for factors at the graduate level, the residuals should be interpreted as measures of Type A effectiveness.

Any residual should be used in conjunction with its comparative standard error (Goldstein, 2003), in order to take into account the uncertainty associated with the assessment of effectiveness.

![Figure 2 – Simultaneous confidence intervals for level 3 residuals. Adapted from Biggeri, Bini & Grilli (2001).](image)

Figure 1 shows a simultaneous confidence interval plot which compares couples of level 3 (i.e. university) residuals: two residuals are statistically different at an approximate 95% confidence level if and only if their intervals are disjoint (Goldstein, 2003). Unfortunately, in
most cases the couples are not statistically different, so one should be very careful when trying to rank universities. Anyway, all the largest residuals refer to universities that are in the north of Italy, which means that those are the best with respect to the time to obtain the first job (controlling for the features of the graduates).

As a general comment on this application, it appears that:

- the statistical methodology of multilevel discrete-time survival models is quite effective, as it allows to model in a satisfactory way the main features of the phenomenon under study, while the analysis can be performed using standard software;
- as I noted at various points, the limitations concern the appropriateness of the available covariates, which narrow the scope of the analysis and the interpretability of the results.

The deficiencies in the variables concerning the graduates (e.g. the features of the jobs held after graduation, the condition of actually looking for job) could be remedied by a careful design of the questionnaire. Anyway, an accurate measurement of the times at which the relevant events occur requires complex and expensive longitudinal designs. Notwithstanding, other key variables are inherently difficult to measure, for example the variables defining the social and economic context in which the universities operate, needed for assessing Type B effectiveness.

In a later work Grilli (2005) considers data from a survey on the high school graduates of the year 1995, carried out by the Italian National Statistical Institute (Istat) three years later. The graduates are nested in schools, so a 2-level model is adopted. Two grouped-time versions of Cox proportional hazards model are compared, showing the higher flexibility of the so-called “continuation ratio” version.

The set of available explanatory variables is wide, including many variables at the graduate level and a few variables at the school level (e.g. technical vs. gymnasium, public vs. private). Moreover, an attempt to control for local labour market conditions is made by using an external source to derive further explanatory variables for the regional-level youth unemployment rate and its trend.

Compared with university graduates, high-school graduates have hazards of getting job with more complex patterns, since most of them do not immediately search for job, but attend
training or educational activities (often of short length). Unfortunately, the dataset does not include information on the beginning and ending times of such activities, so they must be treated as time-invariant covariates, even if their slopes are allowed to vary with time in order to achieve more flexibility.

References


