PARAMETRIC TRANSITIONAL MODELS FOR THE ANALYSIS OF TIME SEQUENCES: TWO EXAMPLES

Marco Bonetti and Gaia Salford

Istituto di Metodi Quantitativi
Università Bocconi
V.le Isonzo 25, 20135 Milano, Italy
(e-mail: marco.bonetti@unibocconi.it, gaia.salford@phd.unibocconi.it)

ABSTRACT. We illustrate the use of parametric transitional models to the analysis of time sequences. After recalling an earlier application to medical research, we describe the preliminary results of an implementation to individuals' life sequences in demography. This modeling approach is useful when the underlying structure of the data-generating process can be envisioned with enough precision, and it allows for the description of rather complicated patterns of observed data, such as the phenomenon of masking of one event by another event, anticipation effects, censoring and in general possibly nonignorably missing data. Covariate effects are included through regression components in the models, and inference is performed by standard asymptotic likelihood theory. The analyses are based on motivating datasets from a breast cancer clinical trial and from the Dutch Fertility and Family Surveys (FFS).

1 INTRODUCTION

We discuss two applications of parametric transitional models to time sequences. The approach does not pose particular theoretical problems, and its flexibility allows for the description of rather complicated underlying processes that may give rise to interesting observed data structures. We illustrate an earlier application motivated by clinical research in oncology and a novel one based on a dataset collected for demographic research.

The general setting is that of a categorical outcome observed over time. The approach consists of working out a complete parametric data model which may provide a reasonable description of the stochastic data generating mechanism. This underlying model may then be extended to handle missing data either directly or through the description of truncation effects due to selection mechanisms of individuals in the dataset, masking effects among different kinds of events, or censoring, as well as structural effects such as "anticipation" effects (see below), as well as covariate effects via regression components. (For a specific application to a situation of nonignorably missing data patterns as defined in Little and Rubin (1997) in ordinal categorical data collected over time see for example Cole *et al.* (2005)).

The observed data likelihood may then be (carefully) maximized, and the usual techniques utilized for model selection and for point and interval estimation. Given the parametric structure of the model, additional quantities may be easily calculated either analytically or by generating sequences from the fitted model. Goodness of fit may be assessed by comparing the distribution of some observed quantities with those generated by the fitted model.

Below we describe two applications. In the interest of space, in particular for the first one we refer to the corresponding reference for details and for a list of references.

2 BINARY OUTCOME SUBJECT TO MASKING BETWEEN EVENTS

The first implementation is motivated by the oncology clinical trial setting, and it can be described as the estimation of time to event distributions in presence of a non-distinguishable competing event and censoring. The objective is the clinically relevant description of the process by which menses discontinue and resume (a binary outcome over time) after the administration of potentially ovarian function suppressing adjuvant treatment for the disease. This process is complicated by the fact that natural menopause also occurs in the patient population, and that treatment-induced amenorrhea (TIA) is not distinguishable from menopause unless menses are observed to resume after treatment completion. Also complicating the process is the partial observation due to censoring, and the fact that only pre-menopausal patients were allowed to enter the clinical trial, that randomized patients to the following four arms: LH-RH analogue (goserelin) x 24 months (A), CMF chemotherapy x 6 months (B), CMF x 6 months followed by LH-RH analogue x 18 months (C), and No adjuvant treatment (D). In Szwarc *et al.* (2006) the authors develop a parametric model for this problem. The model

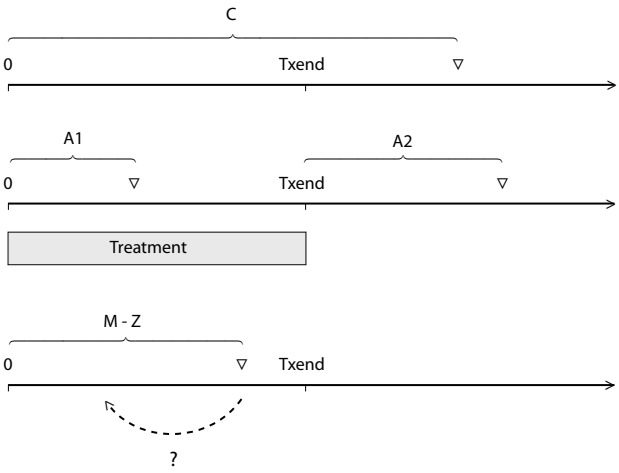

Figure 1. Illustration of data-generating process.

distinguishes between the two kinds of cessation, incorporating the important age effect and allowing for the (scientifically plausible) possibility that treatment induce an anticipation of natural menopause. We illustrate the complete-data generating process in Figure 2. Quantities shown are: C=time to censoring; A_1=time to treatment-induced amenorrhea (both measured from entry into the study); A_2=time to recovery of menses, measured from end of treatment ($Txend$; only for patients who experienced TIA); Z=Age of patient at entry; M=Age at natural menopause (so that M-Z is the time from entry to natural menopause). The question mark indicates the possible anticipation effect on M-Z by treatment (modeled deterministically via shrinkage of $M-Z$ by a constant $k<1$). The model uses cure-rate submodels both for A_1 and A_2 (see for example Peng *et al.* (1998)) since both TIA and recovery are not experienced by all patients. Due to the masking of the two kinds of amenorrhea (treatment-induced and

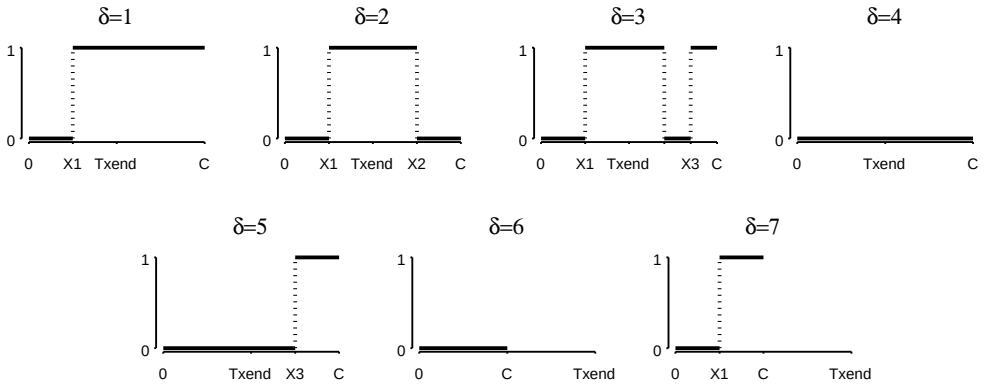

Figure 2. The six possible observed data configurations.

natural) and to the effect of censoring, the observable data is (X_1, X_2, X_3, C), where X_1 is the first cessation observed between entry and $TXend$, X_2 is the time from $Txend$ to resumption of menses, and X_3 is the time from X_2 to the final cessation. The observed data may take one of 7 possible forms, shown in Figure 2. These must be considered separately in the observed data likelihood.

After performing model selection and fitting the final model to the data, other quantities of interest may be studied. For example, in this setting it seems relevant to study the age-adjusted survival distribution $Pr(S > s \mid \text{ Age at entry} = z)$ of the random variable S=time of the (last) menses interruption that will *not* be followed by a resumption (see Figure 3). This variable may be relevant in the choice of the treatment, as it indicates the moment when the permanent interruption of the patient occurs. Goodness of fit techniques included the comparison of the pre-entry distribution of age at menopause with known historic data as well as comparison of the model-based distribution of observable quantities with the frequency distributions observed in the data. Interestingly, the model confirmed the clinical researchers' hypothesis of an anticipation effect of CMF treatment on natural menopause.

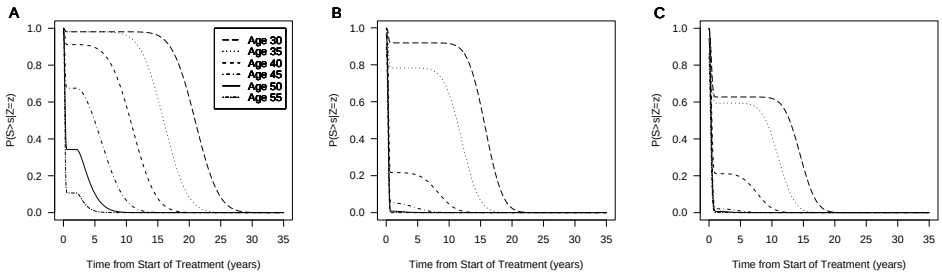

Figure 3. Survival function of time to last menses cessation *not* followed by resumption, for the three arms (left to right) goserelin only, CMF only, and CMF+goserelin.

3 LIFE SEQUENCES IN DEMOGRAPHY

The description of life courses, and in particular transition to adulthood, has attracted increasing interest in the demographics literature. A number of methods have been proposed to study trajectories of events, mainly with two goals: (i) to find common patterns among a set of sequences; and (ii) to describe how they are generated. In this work we deal with the second objective by proposing a model that generates sequences whose properties resemble those of the original data. (See for example McVicar and Anyadike-Danes (2002)).

The motivating data for our analysis originate from the Family and Fertility Surveys (FFS) conducted between 1988 and 1999 in 25 countries by the National Statistical Offices (see Scheonmaeckers and Lodewijckx (1999)). In particular, we study the transition to adulthood in the Dutch data. The retrospective histories of 1897 women about childbearing and union formation were collected on a monthly time scale. Of these, 915 women belong to the 1953-1958 birth cohort and 982 belong to the 1958-1962 birth cohort. Information on respondent's family status between the age of 18 and 30 were collected, leading to a trajectory of 144 consecutive states. Figure 4 shows three sample trajectories.

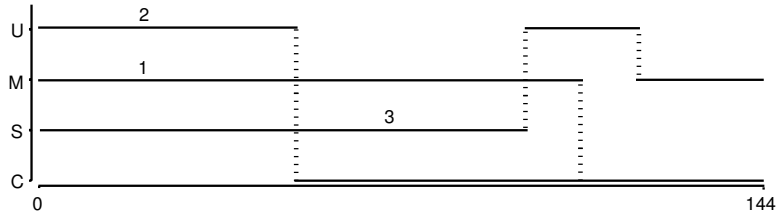

Figure 4. Three examples of possible observed life sequences (see text for the definitions of the states U, M, S, and C).

We model the whole process that generates the paths with a combination of (discrete) time-to-event distribution and transition probabilities. These parametric models are allowed to depend on covariates through generalized logit models. We estimate the transition probabilities and the duration distributions between subsequent transitions, as well as compute the probability that a given individual experiences a certain transition for each combination of characteristics.

Consider an individual who has visited a total of $r+1$ states, not all necessarily different (i.e. who has experienced r state transitions), with S_0 the baseline state. Let $\{S_k, k = 0,1,\ldots,r\}$ indicate the states that the individual experiences, in the order in which they have been visited, and $\{T_k, k = 1,\ldots,r+1\}$ be the corresponding times spent in the $r+1$ states. Let Z denote a vector of covariates. These covariates can in principle be time-varying (and indeed in the motivating application this is the case), but to simplify notation here we simply indicate them as Z. We assume the following:

1. T_k follows a geometric distribution with parameter p that depends on covariates through the logit link: $p = (\exp(\beta Z))(1+\exp(\beta Z))^{-1}$.
2. The probability of transitioning from state i_k to state j_k at the k-th transition, conditionally on the covariate vector Z, is the quantity $P_{i_k j_k k} = Pr(S_k = i_k | S_{k-1} = j_k, Z)$, where

$i_k, j_k \in \{1,...,J\}$ and $k = 1,...,K$, with K representing the maximum number of possible transitions, J the total number of states, and $P_{iik} = 0\ \forall i,\ \forall k$. These transition probabilities are modelled through generalized logits (see for example Agresti (1990)). Using the Jth category as a reference, the logits can be parametrized as $P_{i_k j_k k} = (\exp(\gamma_{i_k j_k} Z))(1 + \sum_{h \in \{(1,...,J-1)-(i_k)\}} (\exp(\gamma_{i_k h} Z)))^{-1}$ with $j_k \in \{(1,...,J-1)-i\}$. For the last state one has $P_{i_k Jk} = (1 + \sum_{h \in \{(1,...,J-1)-(i_k)\}} (\exp(\gamma_{i_k h} Z)))^{-1}$.

Each individual's contribution to the loglikelihood depends on the number of (not necessarily different) states visited. For a woman who visits $(r+1)$ states (r transitions) for durations $\{t_k, k = 1,\ldots,(r+1)\}$, the contribution is: $l(\theta|S,T,Z) = \sum_{k=1}^{r} [\log(P_{i_k j_k k}|Z)) + \log(P(T_k = t_k|Z))] + \log(P(T_{r+1} > t_{r+1}|Z))$ to account for the fact that the last duration is always censored. Here we called $\theta = (\beta,\gamma)$ the full parameter vector. Let $OT \in \{0,...,K\}$ be the random variable that counts the number of observed transitions. The parameter θ can be estimated from the observed data on n women by maximizing the observed data log-likelihood $l(\theta|S,T,Z) = \sum_{p=1}^{n} [1(OT_p = 0) l_0(\theta|S_p,Z_p) + ... + 1(OT_p = K) l_K(\theta|S_p,T_p,Z_p)]$ where for a woman p who has experienced r transitions we have $l_r(\theta|S_p,T_p,Z_p) = \sum_{h=1}^{r} [\log(P_{i_h j_h h}|Z_p) + \log(P(T_h = t_{hp}|Z_p))] + \log(P(T_{r+1} \geq t_{r+1}))$ and $l_0(\theta|S_p,Z_p) = \log(P(T_1 \geq 144|Z_p))$ to take into account the fact that the last duration is always censored (this occurs at time 144 months, the end of the data collection period).

In order to improve the optimization process (we used the OPTIM function in *R*) we provided the function with the vector of the partial derivatives of the log-likelihood with respect to the parameters. The number of parameters of this model can be very high, but some of these can be set to 0 to obtain more parsimonious models. Traditional likelihood-theory-based hypothesis testing can be used to select the "best" model (when estimation is possible, that is).

In the Dutch FFS dataset the states were: Single (S); Married (M); Unmarried cohabitation (U); Single with at least one child (SC); Married with at least one child (MC); Unmarried cohabitation with at least one child (UC). Out of the 1897 women, 4 were excluded from this analysis because they are the only ones who experienced transition from S to MC, from S to UC, and from U to MC. The number of observations for the parameter estimates of these transitions was not sufficient for estimation, and therefore we removed these sequences. (We forced the corresponding transition probabilities to zero in the model.) In addition, transitions from states with children to states without children are not possible and therefore we also set these parameters to zero. Parameters of other transitions that never occurred in the data were also set to zero (i.e. from M to SC, from M to UC, from U to SC, and from U to MC). The marginal counts of all transitions in the observed data are shown in Table 1. Note that because of the structure of the model, the table refers to transitions conditionally on the fact that a transition did occur. The most frequent marginal transition from M was MC, from U was M, from SC was UC, from UC was SC, from UC was MC, and from S women opted more for U or M. The sparseness of the data did not allow for the fitting of the model on the various "Children" states. As a consequence, we grouped the last three states (SC, MC, UC) into a unique absorbing state "C."

Some baseline information was available on individual and family characteristics: we included in particular the woman's level of education (1,2,3 if the woman had respectively no education, from 0 to 3 years of education, more than 3 to 5 years of education after age of

15). In the model we inserted level of education as two dummy variables for levels 2 and 3. Two indicator variables were included for religion belief (1 if the women believd in any religion, 0 otherwise) and for parental divorce status (1 if parents of respondent were separated or divorced). All three covariates were used to model the permanence times and transition probabilities. Moreover we inserted, among the covariates that explain the permanence times in each state, the (rescaled) age in months of each woman before she entered that state, as well as the state being visited. Among the covariates that explain the transition probabilities we also included the age at the time of the transition and the time spent in the state before the transition. Note that these covariates change at each visited state.

	S	**M**	**U**	**SC**	**MC**	**UC**
S	0	920	911	40	0	0
M	32	0	8	0	1140	0
U	178	554	0	0	0	47
SC	0	0	0	0	19	68
MC	0	0	0	67	0	7
UC	0	0	0	16	52	0

Table 1. Number of transition in FFS Dutch data

We applied the model described above to this data, using sequential Wald tests to select a final parsimonious model (results to be described elsewhere). A simulation study of the model was performed to ensure accuracy of the coverage probabilities of the asymptotic confidence intervals of the parameters (data not shown). Goodness of fit was assessed empirically by generating data according to the estimated parameters. Such sequences were censored at 144 months as in the original data, and summary statistics of the frequencies of the transitions among states were compared overall and within strata defined by the baseline covariates. Results were reproducing the original data satisfactorily. In addition to this comparison, an alternative permutation-distribution procedure is being developed to test for sufficient goodness of fit. This procedure will be presented elsewhere and will be applied to this data (work in progress).

REFERENCES

AGRESTI, A. (1990): *Categorical Data Analysis*. Wiley, New York, NY.

COLE, B.F., BONETTI, M., ZASLAVSKY, A.M., GELBER, R.D. (2005): A Multistate Markov Chain Model for Longitudinal Quality-of-Life Data Subject to Non-Ignorable Missingness. *Statistics in Medicine, 24*, 2317–2334.

LITTLE, R.J.A., RUBIN, D.B. (1987): *Statistical Analysis with Missing Data*. Wiley, New York, NY.

MCVICAR, D., ANYADIKE-DANES, M. (2002): Predicting successful and unsuccessful transitions from school to work by using sequence methods. *J. Royal Statistical Society A, 165*, 317–334.

PENG, Y., DEAR, K.B.G., DENHAM, J.W. (1998): A generalized F mixture model for cure rate estimation. *Statistics in Medicine, 17*, 813–830.

SCHEONMAECKERS, R., LODEWIJCKX, E. (1999): Changes in demographic behaviour in Europe: some results from FFS-country reports and suggestions for further research. *European Journal of Population, 15*, 207–240.

SZWARC, S., BONETTI, M. (2006): Modeling Menstrual Status During and After Adjuvant Treatment for Breast Cancer. *Statistics in Medicine, 25*, 3534–3547.