

The Effect of Classification Errors in Survival Data Analysis

L'effetto di Errori di Classificazione nell'Analisi di Sopravvivenza

Adriano Paggiaro

Dipartimento di Scienze Statistiche
Università di Padova

Nicola Torelli

Dipartimento di Scienze Economiche e Statistiche
Università di Trieste

Via C. Battisti 241, 35121, Padova
E-mail: paggiaro@stat.unipd.it

Piazzale Europa 1, 34127, Trieste,
E-mail: nicola.torelli@econ.univ.trieste.it

Riassunto: l'analisi di sopravvivenza studia il processo di transizione verso un definito stato di destinazione, condizionatamente al tempo trascorso in uno stato di origine. Per le applicazioni di tale analisi in ambito socio-economico si dispone, tipicamente, di dati relativi ad un campione estratto dalle unità presenti nello stato di origine in un dato momento, per il quale sono poi previsti uno o più follow-up. In questo lavoro, si discute l'effetto di errori di classificazione nella registrazione della transizione ad altro stato sulla stima dei parametri di un modello di sopravvivenza. In particolare, si propone un approccio bayesiano per stimare i parametri del modello quando gli errori di classificazione sono assunti dipendenti da variabili esplicative.

Keywords: Misclassification, Transition Models, MCMC.

1. Introduction

Survival data analysis aims at collecting data on durations spent in a state by a sample of units, in order to analyse the process of transition to a different state. When interest focuses on survival analysis of social and economic phenomena, the commonest way of collecting information on time spent by people before experiencing a given event is to select a set of individuals who enter a given original state during a given period (e.g., those starting a spell of unemployment, those who get married, patients with a given disease, etc.), and then to collect data on possible transitions to another state in one or more follow-ups. In this paper, we explore the effect of misclassification of the transition indicator (formally equivalent to the censoring indicator).

Our motivating example is the analysis of data on unemployment duration, but the problem may be of more general interest. When analysing transition to employment longitudinal, survey data are widely used, allowing the estimation of some descriptive measures, like mass flows among national labour forces. The effects of misclassification in estimating gross flows have been analysed by many authors, and it is well-known that estimates may be affected by severe bias, leading to a totally erroneous picture of labour force dynamics. The same type of data are used to estimate transition models aimed at explaining how time spent in one state affects the probability of leaving it for another. Analysis of unemployment duration is a classic example of application of these models in the context of labour force dynamics. Lancaster (1990) reviews the uses of data from different observation schemes in estimating transition models.

Data on time spent in each state are often obtained by observing the states occupied by the same unit in successive waves of a panel survey. Looking at analysis of labour market dynamics,

if labour force states are misclassified, either, observed transitions actually refer to situations in which the units are still in the same state, or those observed in the same state have in fact changed. Parameter estimates of transition models when states are misclassified, may be expected to be biased. Torelli and Paggiaro (2002) assess the effects of misclassification of the censoring indicator with data obtained from a scheme common to labour force surveys, and propose a Bayesian estimation strategy to obtain estimates of both the parameters of the duration model and the misclassification probabilities (Bayesian estimation strategies with misclassified binary responses are also assessed in Rekaya *et al.*, 2001). In the present paper, those results are extended in two directions: (i) introducing the possibility of considering explanatory variables of misclassification ; (ii) discussing some assumptions about the role of misclassification in forming the relevant sample for analysis are discussed.

2. The general model for misclassification in the censoring indicator

Our interest focuses on transitions from a state U (in our example, unemployment) to the destination state E (employment). We assume that data t , δ and x are available and that, on a given sample of units, they respectively record time T spent in U, a binary destination indicator, with $\delta = 1$ if a transition to E is observed and $\delta = 0$ otherwise, and the values of covariates X . Our goal is to assess how T and X affect the probability of transition to E. Let $S(t)$ and $f(t)$ denote respectively the survival function and the density function of T . The dependence of survival on X may be modelled in different ways (for instance, by assuming a proportional hazard specification), so that the effect of the covariates is measured by a set of parameters β . We assume that S belongs to a given parametric family S_θ .

In stock sampling, one randomly samples the set of people who occupy a state U at a specified time. This is typically the case for data currently collected in the Italian labour force survey (and in fact in many other national labour force surveys). Elapsed duration t and the state after a fixed time interval of length k are observed. To avoid the problem of length bias, as longer spells have a greater probability of being in the sample (see Lancaster, 1990, for details), the following likelihood is appropriate:

$$L(\beta, \theta) = \prod_{i=1}^n \left[1 - \frac{S(t_i + k; x_i, \beta, \theta)}{S(t_i; x_i, \beta, \theta)} \right]^{\delta_i} \left[\frac{S(t_i + k; x_i, \beta, \theta)}{S(t_i; x_i, \beta, \theta)} \right]^{1-\delta_i}.$$

Torelli and Paggiaro (2002) analysed data from stock sampling with follow-up, assuming independent misclassification for the second interview. Here, we generalise the framework, allowing classification errors to depend on covariates Z . Let:

$$(a) \ P(\delta = 1|E; Z) = p_E(Z, \gamma_E) = \frac{\exp(z\gamma_E)}{1 + \exp(z\gamma_E)} \text{ and}$$

$$(b) \ P(\delta = 1|U; Z) = p_U(Z, \gamma_U) = \frac{\exp(z\gamma_U)}{1 + \exp(z\gamma_U)}$$

be the probabilities, depending on Z , that we observe a transition from U to E when the true states at the second interview are respectively E and U. Misclassification is implied by assuming that (a) is not equal to 1 (but is very close to 1) and (b) is not equal to 0 (but is very close to 0).

To see the effect of such errors on estimating the duration model, we analysed the results of some Monte Carlo studies. We simulated data assuming that transition times are generated by a proportional hazard model, with different assumptions for the baseline hazard function. Here we briefly describe the main results:

- as noted in Torelli and Paggiaro (2002), when errors are not dependent on covariates, the estimates of both shape parameters θ and proportional effects β are biased;
- with low (high) transition rates, the magnitude of the bias seems to be mainly due to the value of $\gamma_U(\gamma_E)$, while that of $\gamma_E(\gamma_U)$ is less relevant;
- the bias of the estimate of θ does not seem to depend on Z , whereas the estimates of β depend on the form of the effect of Z on misclassification probabilities. If X and Z share some common covariates, the bias is attenuated (amplified) when the effects of X and Z are of the same (different) sign. If X and Z are sets of entirely different variables, the bias is related to the correlation between them. Lastly, an effect in the precision of estimates is observed.

3. Misclassification and stock sampling

The effect of misclassification is examined by assuming that no classification errors are present when the units are observed for the first time. When misclassification is assumed in the first interview, the problem becomes that of correctly defining the sample, as it is composed of units which are in state U at the first observation. Thus, two kinds of error may arise:

- (a) units are included in the sample which are in fact not in state U , the problem of unobserved heterogeneity;
- (b) units are excluded which in fact are in state U , the problem of possible non-random selection.

Analysis of this kind of error must be studied within a completely different framework. Nonetheless, in many practical studies it is reasonable to assume that misclassification in the first interview is less important. This is often either because data are collected more accurately and with more reliable modes in the first survey, or because classification in a particular state can be often obtained from registry data, which are usually more reliable.

4. Estimation Strategy

Torelli and Paggiaro (2002) propose a Bayesian strategy to estimate jointly the parameters of interest and misclassification probabilities, in the case of stock sampling with follow-up and misclassification of the censoring indicator in the second interview. Here, we generalise the joint likelihood of the transition model with classification errors which are dependent on the covariates:

$$L(\beta, \theta, \gamma_E, \gamma_U; t, X, Z) = \prod_{i=1}^n \left[1 - \frac{S(t_i + k; x_i, \beta, \theta)}{S(t_i; x_i, \beta, \theta)} \right]^{E_i} \left[\frac{S(t_i + k; x_i, \beta, \theta)}{S(t_i; x_i, \beta, \theta)} \right]^{1-E_i} \times \left\{ p_E(Z, \gamma_E)^\delta [1 - p_E(Z, \gamma_E)]^{1-\delta} \right\}^{E_i} \times \left\{ p_U(Z, \gamma_U)^\delta [1 - p_U(Z, \gamma_U)]^{1-\delta} \right\}^{1-E_i}.$$

The posterior likelihood is given by the joint likelihood and independent gaussian priors. Estimation derives from investigating posterior distribution by MCMC methods, with a straightforward generalisation of methods described in Torelli and Paggiaro (2002).

Some results on a simulated sample are shown in Table 1. In designing the simulation study, we matched a real situation, so that sample size (999 individuals) and data on covariates were fixed at the values actually observed in the Italian labour force survey for those unemployed in northern Italy in October 1997. Moreover, the parameters were such that the average duration was not far from the durations typically observed in Italy. Data were simulated assuming a proportional hazard model with a Weibull baseline (α indicating the shape parameter) and gender as the only covariate X . Misclassification is assumed to depend on who responds (self or proxy).

Table 1 compares some summaries of posterior distributions obtained when the effect of Z is taken into account or not. The main evidence is on estimation of parameters of interest, which in both models works better than MLE. Estimation of α and the intercept are close in the two models, while some differences emerge in the effect of gender: the marginal posterior distribution estimated by the model taking into account the covariate in Z is more symmetric and less dispersed than that from the other model.

Regarding estimation of misclassification probabilities: (a) the posterior distribution of γ_E is highly dispersed but, as discussed in Torelli and Paggiaro (2002), this does not seem to affect the estimation of other parameters; (b) the introduction of the explanatory variable self/proxy seems to help in estimating γ_U parameters, for both intercept and effect of respondent.

Table 1: *Some summaries of samples obtained (by MCMC) from posterior distribution of a transition model, misclassification with and without covariates*

	Maximum likelihood			Posterior distribution summaries and percentiles						
	True	MLE	s.d.	Mode	Mean	5%	25%	50%	75%	95%
Misclassification with covariates										
Log α	-0.5	-0.330	0.126	-0.416	-0.532	-1.106	-0.662	-0.478	-0.332	-0.161
Intc.	-2.5	-2.239	0.387	-2.531	-2.727	-3.921	-3.164	-2.641	-2.251	-1.682
Gender	1	0.631	0.177	0.928	1.099	0.435	0.746	1.030	1.343	2.036
γ_E Intc.	3.5			3.763	4.007	1.924	3.105	3.939	4.872	6.260
Resp.	1			0.009	-0.100	-2.358	-1.144	-0.091	0.901	2.164
γ_U Intc.	-2.5			-2.152	-2.316	-3.249	-2.512	-2.213	-1.989	-1.734
Resp.	-1			-0.987	-1.287	-2.467	-1.659	-1.160	-0.848	-0.462
Misclassification without covariates										
Log α	-0.5	-0.330	0.126	-0.395	-0.527	-1.135	-0.648	-0.462	-0.326	-0.171
Intc.	-2.5	-2.239	0.387	-2.510	-2.698	-4.049	-3.092	-2.609	-2.194	-1.655
Gender	1	0.631	0.177	0.850	1.171	0.457	0.738	0.985	1.500	2.391
γ_E				3.992	4.087	1.917	3.136	4.056	5.008	6.358
γ_U				-2.479	-3.134	-4.892	-3.612	-2.907	-2.467	-2.182

References

- Lancaster T. (1990) *The Econometric Analysis of Transition Data*, Cambridge U. P., Cambridge.
- Rekaya R., Weigel K.A., Gianola D. (2001) Threshold model for misclassified binary responses with application to animal breeding, *Biometrics*, 57, 1123-1129.
- Torelli N., Paggiaro A. (2002) Estimating transition models with misclassification, in: *IASS Topics, 53rd Session of the International Statistical Institute, Proceedings*, IASS (forthcoming).