The Effect of Misclassification in Estimating Transition Models

Nicola Torelli
Department of Economics and Statistics, University of Trieste
P.le Europa 1
Trieste, Italy
nicola.torelli@econ.univ.trieste.it

Adriano Paggiaro
Department of Statistics, University of Padua
Via C. Battisti 241
Padua, Italy
paggiaro@stat.unipd.it

1. Introduction

Longitudinal survey data are widely used to study micro-level dynamics of social and economic phenomena. The interest is often on the use of longitudinal data to study the transition of units among a finite set of states. For instance, when studying labor market dynamics, the estimation of some descriptive measures, like gross flows among labor force states, is crucial. The effect of misclassification of labor force states in estimating gross flows have been analyzed by many authors and it is well known that estimates could be affected by a severe bias leading to a totally wrong picture of labor force dynamics (for a review, see Skinner and Torelli, 1993). Statistical models have been proposed to estimate gross flows taking into account classification errors; these models can either use validation data from re-interview studies or use auxiliary variables (see, among the others, Chua and Fuller; 1987, Pfefferman, Skinner and Humphrey, 1998).

Similar type of data can be used to estimate transition models. By transition models we mean models aimed to explain how the time spent in a state affects the probability of exiting it. Analysis of unemployment duration is a classical example of application of these models in the context of studying labor force dynamics. If labor force states are misclassified, observed transitions could correspond to situations where actually the unit is still in the same state, or vice versa, thus one should expect that estimates obtained from the application of transition models could be biased.

The problem is obviously more general, and similar consequences could be expected whenever longitudinal data from follow-up studies or from panel surveys are available.

In this paper the effect of misclassification in estimating transition model is assessed. A strategy is then proposed to obtain estimates of the model parameters along with estimates of misclassification probabilities. In section 2, a simple version of a transition model polluted by misclassification in the destination state is presented and the effect of ignoring misclassification is evaluated by means of a simulation exercise. In section 3, an estimation strategy to correct for misclassification is specified. Some final comments are in section 4.

2. A transition model with misclassification in the destination state

Let us start by considering the simplest situation, where data are obtained from two waves of a panel survey. Our interest is in estimating the probability of transition between two states (denoted in the sequel E and U); more precisely, we will consider the probability of occupying the state E for those in the state U at the first survey. It is also assumed that data on the time T spent in the state U before the first interview are available and that the time between the two interviews (say k) is short enough to make negligible the probability that more than one transition occur between the surveys. The probability of transition from U to E depends on the time T spent in the state U and on a vector of covariates X. Let $S(t)$ and $f(t)$ denote respectively the survivor function and the density function of the time t spent in U at the moment of the first interview and let d=1 if a transition to E is

observed and δ=0 otherwise.
The dependence of the survivor function on x can 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; estimation of β and θ can be obtained by maximising the following likelihood function (assuming that a sample of n independent observations is available)

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

This situation closely corresponds to what is typically found when analysing unemployment duration data from Labour force surveys that adopt rotating sampling scheme (Trivellato and Torelli 1989), *i.e.*, considering a state-based sample with follow-up. These models have been frequently adopted for application by econometricians and other social scientists (see Lancaster, 1990).

Let us assume that (i) at the first interview the state is observed without error, (ii) at the second interview the state can be misclassified. This simple situation is not without some practical interest, often the follow-up interview is less accurate and obtained using different modes (telephone interviews instead of face to face, more proxy responses, etc.). The same assumption is made by Poterba and Summers (1995). The indicator δ is measured with error, then: (a) $P(\delta=1 \mid E)=p_E$, *i.e.* the probability that we observe a transition from U to E when the true state at the second interview is E, is not equal to 1 (but we expect it to be very close to 1), (b) $P(\delta=1 \mid U)=p_U$, *i.e.* the probability that we observe a transition from U to E but the true state at the second interview is U, is not equal to 0 (but it is likely that it is close to 0).

This simple misclassification mechanism can induce severe bias in parameter estimates. To appreciate this we report some results from a Monte Carlo study; state-based data with follow-up are simulated assuming that transition times are generated by a proportional hazard model with a Weibull baseline (with shape parameter denoted by α). We tried to match the simulation study to some real situation, so the size of the sample and data on the covariates (age, sex, education, marital status) are fixed at the same values actually observed in the Italian labour force survey for those unemployed in northern Italy in 1997, and the parameters used are such that the average duration is not far from those typically observed for unemployment duration in Italy.

Table 1 contains a selection of the simulation results, where γ_E and γ_U denote the logit transforms, respectively, of p_E and p_U.

Table 1. Monte Carlo simulations, Weibull hazard: number of replications=100, γ_E ***=4.5,*** γ_U***=-3.5.***

Parameter	**true**	**mean**	**st.dev**	**true**	**mean**	**st.dev.**	**true**	**mean**	**st.dev.**
log α	-0.5	-0.399	0.122	0	-0.011	0.094	0.5	0.422	0.087
Intercept	-2.5	-2.390	0.390	-4	-3.587	0.399	-6	-5.279	0.590
age	-1	-0.788	0.308	-1	-0.815	0.309	-1	-0.798	0.301
Sex (1=F)	1	0.845	0.160	1	0.834	0.158	1	0.821	0.184
Marit. (1=married)	0	0.013	0.244	0	-0.016	0.243	0	0.015	0.251
Educ. (1=Higher).	1	0.855	0.168	1	0.814	0.175	1	0.809	0.169

The evidence about the strong biasing effect due to misclassification has been clearly confirmed by larger simulation exercises (not presented here). It is interesting to note that the size of the β parameters is reduced showing a sort of attenuation effect.

3. Bayesian estimation of transition models with classification errors

If p_E and p_U were known it would be possible to obtain consistent estimation of the parameters of the model by maximising the following likelihood function:

$$L(\boldsymbol{b},\boldsymbol{q}; p_E, p_U) = \prod_{i=1}^{n} [P(\boldsymbol{d}_i = 1; \boldsymbol{b},\boldsymbol{q}, p_E, p_U)]^{d_i} [1 - P(\boldsymbol{d}_i = 1; \boldsymbol{b},\boldsymbol{q}, p_E, p_U)]^{1-d_i} \quad \text{where}$$

$$P(\boldsymbol{d} = 1; x) = P(\boldsymbol{d} = 1|\mathrm{E})P(\mathrm{E}; x) + P(\boldsymbol{d} = 1|\mathrm{U})P(\mathrm{U}; x) =$$
$$= p_E \left[1 - \frac{S(t_i + k; x_i, \boldsymbol{b},\boldsymbol{q})}{S(t_i; x_i, \boldsymbol{b},\boldsymbol{q})}\right] + p_U \left[\frac{S(t_i + k; x_i, \boldsymbol{b},\boldsymbol{q})}{S(t_i; x_i, \boldsymbol{b},\boldsymbol{q})}\right]$$

is the probability of observing a transition. In fact, given some reasonable restriction for allowing identification of the parameters, the same likelihood function can be used to obtain estimates of the parameters of the model also when misclassification probabilities are unknown (possibly using the E-M algorithm). Unfortunately, in practice maximising the likelihood function for this model leads very often to totally unreasonable values for some parameters and the algorithm converges to points at the border of the parametric space.

In applications to real data, it is not unreasonable to have some more or less vague ideas about which values are reasonable for the misclassification probabilities. At least we know that those probabilities should not be very far from 0, this encourage us to formulate the same model within a bayesian framework. In this case, the posterior distribution of the parameters is given by:

$$g(\boldsymbol{b},\boldsymbol{q}, p_O, p_D, \mathrm{A}|t, X, \boldsymbol{d}) = L(\boldsymbol{b},\boldsymbol{q}, p_O, p_D, \mathrm{A}|t, X, \boldsymbol{d}) g_0(\boldsymbol{b},\boldsymbol{q}, p_O, p_D, \mathrm{A})$$

where $L(\boldsymbol{b},\boldsymbol{q}, p_O, p_D, \mathrm{A}|t, X, \boldsymbol{d}) =$

$$\prod_{i=1}^{n} \left[1 - \frac{S(t_i + k; x_i, \boldsymbol{b},\boldsymbol{q})}{S(t_i; x_i, \boldsymbol{b},\boldsymbol{q})}\right]^{A_i} \left[\frac{S(t_i + k; x_i, \boldsymbol{b},\boldsymbol{q})}{S(t_i; x_i, \boldsymbol{b},\boldsymbol{q})}\right]^{1-A_i} \left[{p_O}^{d_i}(1 - p_O)^{1-d_i}\right]^{A_i} \left[{p_D}^{d_i}(1 - p_D)^{1-d_i}\right]^{1-A_i}$$

and A denotes a variable indicating the transition without error and g_0 denotes the prior distribution. The posterior distribution can be explored by using MCMC methods alternating the Gibbs sampler and the Metropolis-Hastings algorithm.

Table 2: Sample from the posterior distribution: data simulated by MCMC strategy.

	Maximum likelihood			**Posterior distribution summaries and percentiles**						
	true	**MLE**	**s.d.**	**mode**	**mean**	**5%**	**25%**	**50%**	**75%**	**95%**
log α	-0.5	-0.432	0.112	-0.492	-0.563	-0.900	-0.662	-0.534	-0.434	-0.313
Intc.	-2.5	-1.899	0.353	-2.068	-2.266	-3.092	-2.607	-2.213	-1.928	-1.513
Age	-1	-0.850	0.299	-0.928	-1.004	-1.705	-1.272	-0.977	-0.709	-0.396
Sex	1	0.712	0.156	0.886	0.903	0.523	0.751	0.892	1.044	1.294
Marit.	0	-0.280	0.218	-0.176	-0.128	-0.576	-0.337	-0.149	0.050	0.391
Educ.	1	0.880	0.156	1.053	1.143	0.769	0.970	1.116	1.292	1.594
γ_E	4.5			4.167	4.011	2.810	3.565	4.032	4.458	5.121
γ_U	-3.5			-2.945	-3.321	-4.685	-3.737	-3.202	-2.791	-2.438

Table 2 contains some characteristics of the posterior distribution resulting from the use of MCMC applied to one simulated sample from a proportional hazard model with Weibull baseline. In the table (first three colums), one can also find the results obtained by estimating parameters by

maximum likelihood when misclassification is ignored. Once again, to gain realism, simulated data have been chosen to resemble data from two successive wave from the Italian labor force survey, with respect to the analysis of transition from unemployment to employment. The misclassification probabilities chosen for simulation are similar to those encountered in real situations, as far as classification of labor force states is concerned (γ_E=4.5 and γ_U=-3.5 that are approximately equivalent to set p_E=0.99 and p_U=0.05). The prior distributions (assumed independent) for the $\boldsymbol{b}$ parameters are gaussian *N(0,4),* for γ_E and γ_U the priors are respectively *N*(4,1*)* and *N*(-4,1). 100.000 Gibbs sampler iterations have been considered, after a burn in period of 20000 iterations, and a sample of size 1000 from the posterior is obtained selecting values every 100 iterations.

The results obtained seem to encourage the use of this strategy in order to obtain "good" estimates of the parameters of interests in presence of misclassification.

4. Concluding remarks

Often, misclassification probabilities are unknown, but we have good indications about their likely magnitude. The use of classical inferential methods is still possible but less natural and less practical than adopting a bayesian approach. By this approach one can obtain estimates both of the parameters of the model and of the misclassification probabilities. The simple model here considered can be extended in many directions: by considering the case of data coming from multi-wave panel surveys, to include the case where also at the first interview there is a substantial misclassification, and to consider movements among a set of more than two states.

REFERENCES

Chua, T. and Fuller, W.A. (1987). "A Model for Multinomial Response Errors Applied to labor flows". *Journal of the America Statistical Association,* 82, 46-51.

Lancaster, T. (1990), *The Econometric Analysis of Transition Data*, Cambridge University Press.

Pfefferman, D., Skinner, C.J. and Humphrey, S.K. (1998). "The Estimation of Gross Flows in the Presence of Measurement Error Using Auxiliary Variables*". JRSS*, Series A, 161, 13-32.

Poterba, J.M. and Summers, L.H., (1995). "Unemployment Benefits and Labor Market Transitions: a Multinomial Logit Model with Errors in Classification" *Review of Economics and Statistics*, 77, 207-216.

Skinner, C.J. and Torelli, N. (1993)."Measurement Errors and the Estimation of Gross Flows from Longitudinal Economic Data". *Statistica,* 3, 391-405.

Trivellato U. and Torelli, N. (1989). "Analysis of Labor Force Dynamics from Rotating Panel Survey Data". *Bulletin of The Internatinal Statistical Institute,* Vol. LIII, Book 2. 425-444.

SUMMARY

Estimation of models for transitions between a set of states could be severely biased if units are incorrectly classified. In the paper a bayesian strategy to deal with misclassification is proposed.

RESUME

Les estimations des probabilitès de transition pour systhèmes d' ètats finis ont un biais important dans le cas ou il y a des erreurs de classification. Dans cette article on propose une stratègie bayèsienne por corriger le biais.