Using PROC NLMIXED in Survival Analysis with Reversible Events

Carl Formoso, Division of Child Support, Olympia, WA

Abstract

A method is described where multiple event histories of reversible changes are simultaneously analyzed for the effects of independent variables. At present the method is limited to two states with free movement back and forth (such as employed or unemployed).

A likelihood expression is developed relating the observation of being in a particular state at a particular time to a set of covariates. The general maximum likelihood search in PROC NLMIXED is used to estimate the coefficients of the covariates.

Tests with simulated data with three covariates indicate the validity of the method. As an example with real data, welfare histories through 39 months are analyzed with eight covariates.

The mathematics of reversible events suggest that a standard survival analysis approach could produce artifactually time-dependent hazards. Analyzing welfare histories using PROC PHREG did, in fact, produce time dependent hazard rates, while the method presented here gave an adequate fit with time-independent hazard rates.

Introduction

The origins of survival analysis focused on death, a one directional change. This created an essentially one-state analysis centered on residence time in the state and exits from the state. While survival analysis has devised ways to deal with multi-state processes, it retains a one-state viewpoint. This leads to possible shortcomings, particularly in dealing with repeatable events where multiple exits from a state are possible (see Allison, 1995 & 1997).

In this paper we show that it is possible to devise a two-state survival analysis, where exits and entries are considered simultaneously with maximum likelihood estimation. The general mathematical constructs of multi-state reversible processes have previously been presented (Coleman, 1981; Tuma and Hannan, 1984; Allison, 1985), but the analytical approach presented here is new.

Extending the method described in this paper to a higher number of states will be more difficult, but appears to be possible.

Two-State Survival Analysis

Figure 1 shows our model for considering survival with two states. To simplify the work, we will assume that the hazard rates do not depend on time. N_1 is the fraction of the population in State 1 and N_2 is the fraction of the population in State 2, with $N_1 + N_2 = 1$. The rate of change in N_1 at time t can be written as

$$\frac{dN_1}{dt} = -h_1 N_1 + h_2 N_2 \quad ,$$

where h_1 is the hazard rate (number of events per individual per time unit) for exiting State 1 and h_2 is the hazard rate for exiting State 2. Substituting $N_2 = 1 - N_1$, we have a form which integrates to:

$$\ln\left\{\frac{h_2-(h_1+h_2)N_1}{h_2-(h_1+h_2)N_{1,0}}\right\}=-(h_1+h_2)t \quad ,$$

where $N_{1,0}$ is the residence level of State 1 at $t=0$. Assuming that $N_{1,0}=1$ will then give us

$$N_1=\frac{h_2+h_1e^{-(h_1+h_2)t}}{h_1+h_2} \quad .$$

If we let $a=h1/(h1+h2)$ and $b=(h1+h2)$ this will reduce to the form

$$N_1=1-a(1-e^{-bt}) \quad ,$$

and the expression for N_2 will be

$$N_2=a(1-e^{-bt}) \quad .$$

If $N_{1,0} \neq 1$ then the expression for N_2 will be

$$N_2=a(1-e^{-bt})+(N_{2,0})e^{-bt} \quad .$$

N_1 and N_2 are estimators of the probabilities of being in State 1 or State 2 at time t.

Now expressing the dependence of hazard rates on covariates $1,2,...i$ for individual j we have

$$h_{1j}=k_1e^{\sum_i \beta_{1i}x_{ji}} \quad , and$$

$$h_{2j}=k_2e^{\sum_i \beta_{2i}x_{ji}} \quad .$$

Time dependence in hazards could be incorporated by allowing k_1 and k_2 to depend on time.

Thus we can write likelihood expressions which relate the observation of being in a particular state at a particular time to a set of covariates, and can create an analysis similar to standard one state survival analysis. With an appropriate maximum likelihood estimation procedure we are able to fit individual level observed state residence, expressed as a dichotomous variable, to the probabilities, estimating the set of coefficients βmi. The NLMIXED procedure in SAS Version 8, allowing a general maximum likelihood search, was used to process individual level data.

Implications for Standard Analysis

It should be noted that the relationships developed above imply that a standard approach could produce an apparently time-dependent hazard, where the hazard is not, in fact, time dependent. As a simple approach we use N_1 as a survival function; then the apparent hazard is the negative derivative of the natural logarithm of the survival function:

$$h_1=-\frac{d\ln S_1}{dt}\cong-\frac{d\ln N_1}{dt} \quad .$$

Substituting the expression for N_1 and differentiating, the apparent expression for h_1 would be

$$h_1=ab\frac{e^{-bt}}{1-a(1-e^{-bt})} \quad .$$

Where, in fact, $h_1=ab$, and is not time dependent.

Method

Since there are only two states, monitoring one state is sufficient. For each individual we use observation of residence in state 2 at each time. Each time observation, along with the appropriate covariates, becomes a separate record in the data file input to NLMIXED. Thus analysis occurs for each discrete time, and the method is directly adaptable for time-dependent explanatory variables.

Two-State Estimation with Simulated Data

A two-state system, with three covariate factors, was simulated for 10,000 individuals. Figure 2 shows the simulated time dependence of the number of individuals in state 2, and the calculated fit using the best parameters from NLMIXED. Table 1 shows the input simulation parameters compared to the best NLMIXED parameters.

Overall the method produces good results with the simulated data. Increasing the number of individuals improves the results; our first attempt with 1000 individuals was only marginally satisfactory.

Two-State Estimation with Welfare Data

Using data from other work (Formoso, 1999), we follow 116,377 welfare-using adults through 39 months. This cohort had been selected as all adults who had used welfare in 4th quarter of calendar year 1993, with thirteen quarters of follow up data. State 1 is defined as "on welfare, " and State 2 as "off welfare." We limited covariates to the five generally most significant (largest chi-square) from a standard survival analysis, plus three program indicators (Formoso, 1999).

Since we have 39 time observations to fit for each individual, this produces 4,538,703 records to be analyzed by the NLMIXED procedure. To reduce time and space requirements we produced files for analysis which were random samples of the 4,538,703 records. Repeated random samples gave nearly identical results.

Figure 3 shows results for an approximate 2% random sample, 91,033 records. The overall fit is quite good. Because of the translation between a quarterly cohort and monthly data, $N_{2,0}$ is not zero. 9.63% of the cohort did not use welfare in the last month of 4th quarter 1993. Thus we used the second expression for N_2 given above.

Table 2 shows the best parameter estimates from NLMIXED and calculated risk ratios. These values must be viewed with caution because this is not a complete analysis. The purpose here is simply to show that it is possible to obtain results in a two-state survival analysis with real data.

We have previously done a standard one-state survival analysis for this cohort, analyzing exit events from State 1 separately from exit events from State 2. The results are represented in Figure 4, where it can be seen that the slopes of the plots (which are the negative of the hazards) are not constant with time. However, the two-state analysis appears to adequately fit the data with hazards which are not time dependent.

References

Allison, Paul D., *Survival Analysis of Backward Recurrence Times*, J Am Stat Assoc, vol 80, p315-322, 1985.

Allison, Paul D., *Survival Analysis Using the SAS® System: A Practical Guide*, Cary, NC: SAS Institute Inc.,1995.

Allison, Paul D., *Event History & Survival Analysis*, Course Notes,1997.

Coleman, James S., *Longitudinal Data Analysis*, Basic Books, NY, 1981.

Formoso, Carl, *The Effect of Child Support and Self-Sufficiency Programs on Reducing Direct Support Public Costs*, submitted to the Lewin Group under Subcontract #TLG98-003-1898.02, 1999.

Tuma, Nancy B. and Hannan, Michael T., *Social Dynamics*, Academic Press, 1984.

Author

Carl Formoso
Division of Child Support
PO Box 9162, Olympia, WA 98507
(360)664-5090, FAX (360)586-3274
cformoso@dshs.wa.gov

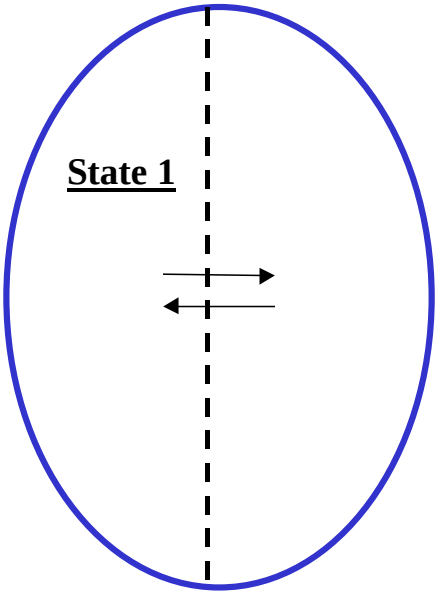

Figure 1
Model for Two-State Dynamical Analysis

Table 1: Simulation Parameters and Two-State Maximum Likelihood Estimates*

Parameter	Input Value	Estimate Value
k_1	0.03	0.0305
k_2	0.01	0.0093
β_{11}	0.1	0.151
β_{12}	1.2	1.212
β_{13}	-0.3	-0.295
β_{21}	0.2	0.416
β_{22}	-1.4	-1.379
β_{23}	0.6	0.530

* All estimates have $p < 0.0001$. Results from the y ~ binary(p) option of NLMIXED.

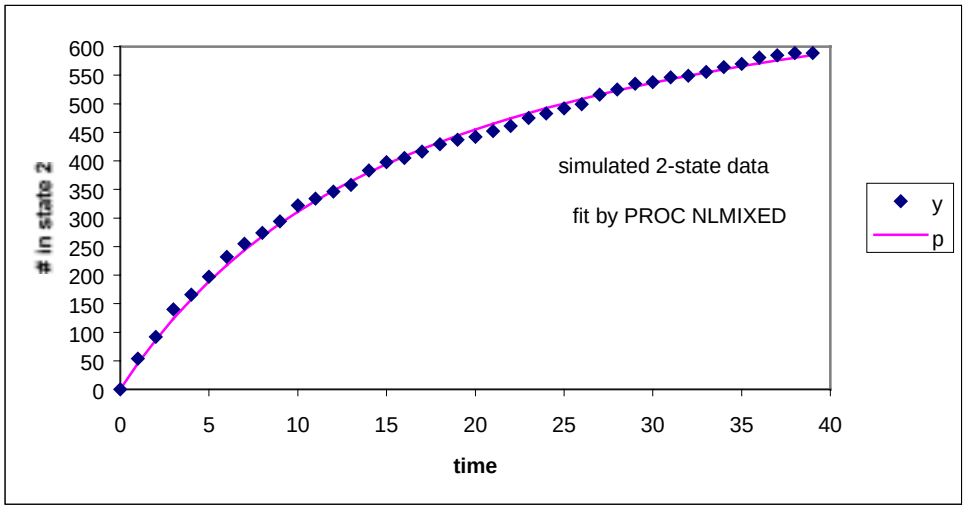

Figure 2: Simulated Observations and Two-State Maximum Likelihood Fit

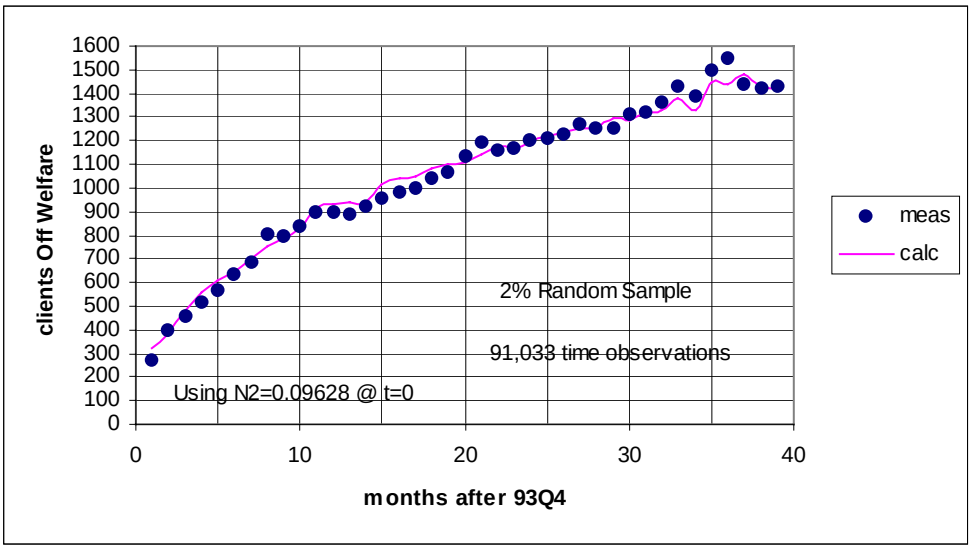

Figure 3: Observed State Residence "Off Welfare" and Two-State Maximum Likelihood Fit

Table 2: Two-State Welfare Maximum Likelihood Estimates*

Parameter	For					
	h1			h2		
	Estimate	Risk Ratio	p	Estimate	Risk Ratio	p
k	0.047	-	<.0001	0.008	-	<.0001
NumFam	-0.271	0.763	<.0001	0.000	1.000	ns
PrevStat	1.325	3.762	<.0001	1.884	6.582	<.0001
PreWelf	-0.658	0.518	<.0001	-0.478	0.620	<.0001
Hisp	0.782	2.185	<.0001	0.000	1.000	ns
Ref	0.811	2.249	<.0001	-0.626	0.535	0.0527
CPJN	-0.269	0.765	<.0001	0.000	1.000	ns
CPJY	0.000	1.000	ns	0.464	1.590	<.0001
CGJN	-0.146	0.864	0.0037	0.000	1.000	ns

* Results from the y ~ binary(p) option of NLMIXED. Covariates are all dichotomous; values are zero except: NumFam=1 if number in family is greater than 2, PrevStat=1 if any wages earned in 93Q4, Prewelf=1 if more than 12 months of welfare used in the two years prior to 93Q4, Hisp=1 if Hispanic, Ref=1 if refugee, CPJN=1 if child support collections are poor and no entry into JOBS program prior to 93Q4, CPJY=1 if poor child support collections and JOBS entry prior to 93Q4, CGJN=1 if good child support collections and no JOBS entry prior to 93Q4.

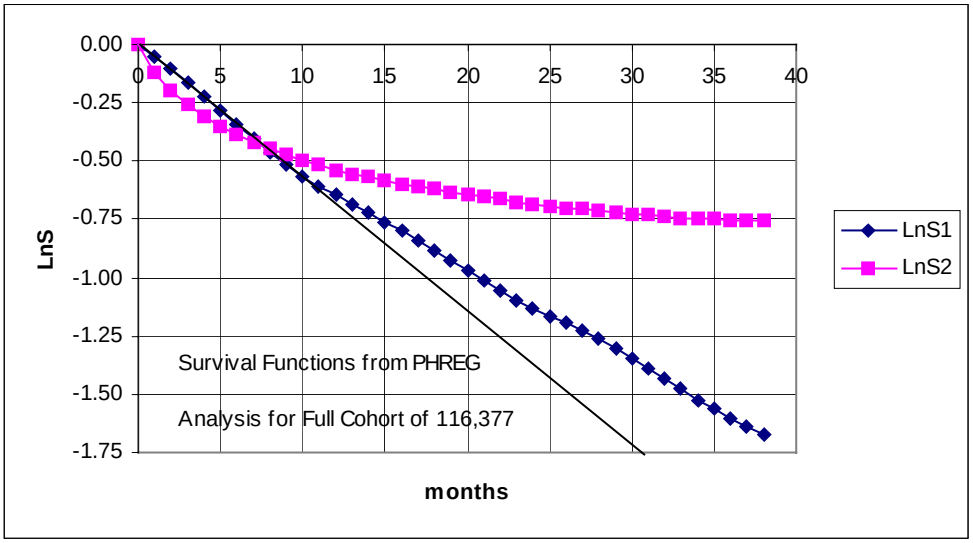

Figure 4: Results from Standard Survival Analysis*

* Output with Baseline option of PROC PHREG. Both curves are for average welfare clients with CPJN=1 (see footnote to Table 2). 95% confidence limits on these curves are approximately represented by the size of the markers.