Contact Histories from the Survey of Grouped Individuals: Multivariate Multilevel Survival Analysis of Grouped Censored Data with Competing Risks

Hiroaki Minato
NORC at the University of Chicago, 55 East Monroe Street, 20th Floor, Chicago, IL 60603

Abstract

Suppose we have a sample of households. Each household contains at least one person. A survey of our interest is to interview multiple randomly selected eligible persons in each randomly selected eligible household in the sample. Define that a household completes the survey if every selected person of the household responds to his or her questionnaire. Extending the current theory of multivariate survival analysis of grouped censored data, we propose to model the survey completion time of households and the survey response time of persons simultaneously. More specifically, we describe group response outcomes and times as well as individual response outcomes and times in terms of within-group dependency structures of static and dynamic types. Our goal is to construct a statistical framework useful for characterizing and understanding contact histories in survey research.

Key Words: censored data, competing risks, contact history, dynamic modeling, multilevel grouped data, multivariate survival analysis

1. Introduction

We propose a research agenda for better understanding contact outcomes and histories when survey units are grouped, as in multistage surveys. A cluster sample is a special case where all eligible members are selected from each group. Units in a given group are thought, in general, to share certain behavior and attitudes and to interact among themselves. For example, a survey may be conducted with multiple selected members from households that are selected into the sample; here, the household forms a predefined group. Another example is a survey of students within selected schools.

Survey contact outcomes and histories in such grouped-individual surveys are much more difficult to analyze because of the within-group dependency and dynamism. Our current goal is to break down the survey contact histories into fundamental pieces and to propose a survival analysis method that can best assemble and model them back. Further, we lay down a statistical framework that may be useful for understanding key analytical elements in the survey contact process that is dependent and dynamic.

2. An Example

We begin with a simple example illustrating all critical ingredients in our survey contact outcome and history data. Suppose the following scenario. We are conducting a community survey. The community consists of households, each of which contains at least one eligible person for the survey. Let us assume we have a complete list of households and their members. In the first stage, we randomly select a sample of households. In the second stage, we randomly select up-to three eligible persons for the survey interview.

Assume that our contact rules are to attempt a contact with each selected person at most once a day for up-to three times. Each person-contact attempt can result in (C) we contacted the person and the person completed the interview, (R) we contacted the person but the person refused the interview and any further contact (e.g., hard refusal), or (U) we could not contact the person. (There may be a soft refusal case such that we could contact the person but the person requested a re-contact in the future time. For simplification, this incomplete but hopeful situation is included in (U).)

The (C) or (R) event would be the final event or the outcome for a given case and would close the case history. Since the two outcomes are mutually exclusive, they may be thought of as competing risks in the survival analysis terminology. The (U) event would be an intermediate event and the case stays active, unless it occurs at the final contact attempt, i.e., at the third contact attempt. In the latter situation, (U) would also signify the final event or the outcome for a given case and the case history would be terminated as incomplete. In survival analysis, such a case is said to be right censored.

Here, we are loosely defining that (1) an event to a case is some classified occurrence to the case at a specified time, (2) an outcome to a case is a result from all events to the case, and (3) a history is a set or series of events and their outcome along with the time information. A case can be at the individual level or at the group level, depending on the observation level of the case.

Table 1 shows an example of survey contact events, outcomes, and histories. The number in the parentheses ( ) indicates the count or the ordinal time of person contact attempt, while the number in the square brackets [ ] indicates the count or the ordinal time of household contact attempt. So, for instance, if two persons are selected from a household and each of the selected persons was contacted once, then the household was contacted twice. Here, it is assumed that multiple persons in a same household are not contacted for the interview simultaneously. That is, no more than one interview is conducted at a given household time. It is a reasonable assumption unless multiple interviewers contact a same household by multiple contact methods (e.g., through multiple phone lines). Therefore, there are no “ties” with respect to the household time. Meanwhile, the calendar time, if defined by dates, could be identical among multiple persons in a same household. For example, a telephone interviewer might successively ask for multiple persons in one connected call.

Table 1: An Example of Survey Contact Events, Outcomes, and Histories

Community	*Household*	*Person*	*Calendar Time*				
			1	2	3	4	5
A	1	1	**C**(1)[1]				
A	1	2		**C**(1)[2]			
A	2	1	**C**(1)[1]				
A	2	2		**R**(1)[2]			
A	3	1	**C**(1)[1]				
A	3	2					**C**(1)[2]
A	4	1	**U**(1)[1]	**U**(2)[2]	**U**(3)[3]		
A	4	2				**U**(1)[4]	**C**(2)[5]
A	5	1			**U**(1)[1]	**C**(2)[4]	
A	5	2			**U**(1)[2]	**R**(2)[5]	
A	5	3			**U**(1)[3]	**U**(2)[6]	**U**(3)[7]

C = Contacted with completed interview
R = Contacted without completed interview and no further contacts allowed
U = Could not contact or finalize
( ) = Person contact attempt; [ ] = Household contact attempt

2. Data Structure

Table 1 depicts the so-called contact history data for a survey. Call histories are common data recorded during the survey data collection stage, and they are usually delivered along with the main survey data. However, they are rarely used or analyzed except for computing the average number of contacts. However, the data are quite rich and come in a rather unusual form for survey researchers. In this section, we are going to dissemble the contact history data to see what may be going on.

First, there are two levels of outcomes—the person level and the household level. For example, both Persons 1 and 2 in Household 1 have the person level outcome of C. Here, we define that a household completes a survey if all selected members of the household complete the survey. Then, Households 1 is said to have completed the survey with two contact attempts to the household. Meanwhile, Household 2 did not complete the survey with two household contact attempts. In Household 4, Person 1 could not be contacted within the limit of the contact rules; thus, we do not know if the person would complete the survey, nor do we know if the household would complete the interview. Household 5 also contains a person with the final outcome of U, but the household outcome is R because its Person 2 was R. Of course, the household outcomes depend on their definitions. If we are to define that a household completes the survey if at least one selected member completes the survey, then Households 4 and 5 would be considered C.

Second, we could have three levels of covariates or explanatory variables—the person level, the household level, and the community level. At the person level, basic demographic information such as age, gender, and family role could be available before the interview from the sampling frame or through screening. More personal information is usually difficult to acquire a priori. At the household level, the race-ethnicity, size, and composition of each sample household may be known. At the community level, some aggregated or summarized information of the community or the geography containing the community may exist, e.g., population density, median income, and other Census-type information.

Third, it would be natural to treat the grouped outcomes as multivariate or parallel outcomes. Although the number of selected persons varies from household to household, our data are clearly multivariate at the household level and should be modeled as such.

Fourth, the unique nature of our data is the time dimension. It is possible to measure the time from the beginning of the data collection to the moment of each contact in terms of physical time such as weeks, days, hours, minutes, and seconds. However, the counting time is often generated for each case by the contact attempt. This is what we are considering in this paper. Thus, in our data, the time for each member as well as for each household is positive and discrete, restricted only by the contact rules. Meanwhile, the calendar time for each contact is normally included in the contact history data, so year, month, week, day of the week, and time of the day may be used as person level covariates.

Fifth, a case history consists of outcome and time, and as we have seen in the example, a person history would be incomplete or right censored if the person's event at the final attempt is still U, i.e., the person's outcome is U with the person time 3. This person is said to be right-censored at the person time 3. Since the contact rules are the same for all persons, we have a homogenous censoring mechanism. On the other hand, households could be censored at various household times—from three to nine. Further, the household censoring is not expected to occur independently of the household characteristics. These are major violations of the usual assumption of homogenous random censoring.

Sixth, if we look at the person outcomes within a given household, at least two kinds of dependency are evident. One is within-group dependency due to some commonality of people who belong to the same group. The members in the same household are likely to possess common biological traits or genetic make ups. They might also share similar environments and experiences. For example, in Table 1, all of Household 3 might go out for dinner on the calendar day 3 and the same day every week. This sort of dependency may be thought of as a positive, constant, and long-term driving force for grouped members' attitudes and behavior.

The other kind of dependency in our data is event-related or event-induced dependency (Hougaard, 2000). This is rather unique to our contact history data because of the dynamic and interactive nature of households whose multiple members are possibly contacted for the survey multiple times. For example, in Household 1 of Table 1, Person 2 was available and cooperative possibly because Person 1 suggested or recommended Person 2 to take the survey, while in Household 2, Person 2 might have felt that the response of Person 1 on the previous day was "enough" for the survey. Thus, unlike the first kind of dependency, the effect of event-induced dependency could be positive or negative.

Further, the strength of event-induced dependency effect may not stay constant over time. That is, the magnitude and duration of the effect may show a non-constant, probably decreasing, functional form. Person 2 in Household 2 could have produced C, if Person 2 had been contacted several calendar days after Person 1—as in Household 3. The effect could also be accumulated or inflated in case of multiple events. (Note that dependency inducible events are effectively either C's or R's, not U's.) In short, recency and frequency of events may matter. Seventh, as defined earlier, our outcome can take one of the three values C, R, and U. In the context of survival analysis, we have two competing risks C and R in addition to censoring U, which could also be taken as a competing risk. In univariate survival data with no dependency structure, competing risks could be nicely decomposed. However, in our current data, competing risks are deeply imbedded in the dependency structure within household, even though we might be able to assume the censoring to be random and to separate it out from the structure.

Finally, in the multivariate survival analysis literature (e.g., Hougaard, 2000), the notion of states has been used to describe changing conditions or stages of a given group. For example, in Table 1, Household 1 may start in the state with no C, making a transition to the state with one C at the calendar time 1 and another transition to the state with two C's at the calendar time 2. Each household state is distinguished in terms of the number of C's in the household. In fact, this is an example of progressive states.

We can extend these ideas in our contact history data. Multiple contact modes are sometimes used in surveys in order to maximize the coverage rate of the target population and the contact success rate. The same mode is usually used for all selected members of a given household; however, if a household is non-contacted under one mode, e.g., CATI, the household might be given another chance under a different mode, e.g., CAPI. With such expanded contact rules, we can think of contact attempts under one mode forming one state and contact attempts under another mode forming another state. Now, each state becomes like an event, and a sequence of states produce the final outcome for the household case.

3. Statistical Models and Problems

Rich ingredients and their complex structures are found in the contact history data. However, the conceptual framework of survival analysis methods is there to model them (see, for example: Cox and Oakes, 1984; Fleming ad Harrington, 1991; Hougaard, 2000; Kalbfleisch and Prentice, 2002; Kleinbaum and Klein, 2005; and Lawless, 2003). A survival analysis model has two simple building blocks—a survivor function, *S*(*t*), and a hazard rate, *h*(*t*), which have the following relation:

$$h(t) = -\frac{dS(t)/dt}{S(t)},$$

where *S*(*t*) = Pr(*T* ≥ *t*) and t is a specific value of nonnegative survival time random variable *T*. In our case, *T* is actually a random matrix with some additional structures for multivariate grouped outcomes, right censoring, and competing risks as well as for multilevel outcomes and times.

To allow for multilevel multiple regression type analyses, *t* may be conditioned on a matrix covariate, *X*:

$$h(t \mid X) = -\frac{dS(t \mid X)/dt}{S(t \mid X)}.$$

The existence of covariates seems to limit our approach to be either parametric or semiparametric, as standard non-parametric models are not well equipped to handle covariates. Without any breakthroughs in nonparametric modeling, we will not pursue this route here.

The most interesting ingredients in our data are the two kinds of dependency, which we now characterize as group-specific random effects and event-induced effects. The former type of dependency has been well studied, and it is often described by so-called frailty models, which assume that all individuals within a group share a common unobserved characteristic, called the frailty (Hougaard, 2000). In other words, group-wise random effects are multiplied to the hazard functions. These random effects are usually assumed to be generated from some standard distribution such as a gamma or Weibull distribution. But, the specification is mostly based on theoretical convenience, so the assumptions tend to be too strong.

Substantively, the event-related effects are more important to understand, because they can be reduced or eliminated by carefully adapted contact rules. For example, we have seen in Table 1 that if there is a strong short-term negative event effect, increasing the calendar time interval between person contact attempts might prevent the effect. However, the event-induced dependency resides in the deepest area of our data. Extracting its pure mechanism is a very difficult statistical task, and measuring the dependency is even more difficult.

Once all the key ingredients and their structures are modeled, our next task is to estimate (1) the survivor function and the median completion time of the households and (2) the survivor function and the median response time of the persons. A median is preferred to a mean, because our data are usually skewed to the right and also because right censoring could lump up the cases at the last time point. Needless to say, their statistical properties, exact or approximate, must be examined and established. Here we note that large sample properties may not be very useful for our data, as the time is finite and normally short. The sample size dimension can be pushed to infinity, but it would not help control the within-group dependency as only the number of groups can go to infinity, keeping the group size bounded by a small number. (However, it could still bring some insight if we are to segment the data by the group size and to develop a separate model for each size class.)

To summarize, we have proposed the following research problems.

I. Identify and specify multivariate and multilevel survival analysis models of grouped censored data with competing risks—including evaluation and selection criteria for various types (parametric, semiparametric, and nonparametric) and forms (additive, multiplicative, and mixture) of models.
II. Specifically, incorporate non-homogenous right censoring and competing risks. (There are some recent theoretical works on the latter topic: e.g., Chen, Kramer, Greene, and Rosenberg, 2008; Cheng, Fine, and Kosorok, 2009; Naskar, Das, and Ibrahim, 2005; and Scheike, Sun, Zhang, and Jensen, 2010.)
III. Formulate, measure, and test the within-group event-induced dependency.
IV. Construct statistically and substantively desirable estimators—e.g., consistent, unbiased, efficient, and computable estimators—for the median survival (or response) time of the individuals and the median survival (completion) time of the groups.
V. Expand or generalize the above models so that multiple states can be analyzed.

Many meaningful applications in survey research await the solutions.

Acknowledgements

I thank Rachel Harter for her helpful comments on my drafts.

References

Cai, T., Cheng, S. C., and Wei, L. J. (2002). Semiparametric mixed-effects models for clustered failure time data. *Journal of American Statistical Association*, **97**, 514-522.

Chen, B. E., Kramer, J. L., Greene, M. H., and Rosenberg, P. S. (2008). Competing risks analysis of correlated failure time data. *Biometrics*, **64**, 172-179.

Cheng, Y., Fine, J. P., and Kosorok, M. R. (2009). Nonparametric association analysis of exchangeable clustered competing risks data. *Biometrics*, **65**, 385-393.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Crowley, J. and Johnson, R. A. (Eds.). (1982). *Survival Analysis*. Institute of Mathematical Statistics, Hayward, CA.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag, New York.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, New York.

Gustafson, P. (1997). Large hierarchical Bayesian analysis of multivariate survival data. *Biometrics*, **53**, 230-242.

Hinkley, D.V., Reid, N., and Snell, E. J. (Eds.). (1991). *Statistical Theory and Modelling*. Chapman & Hall, London.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

Hougaard, P., Harvald, B., and Holm, N. V. (1992). Measuring the similarities between the lifetimes of adult Danish twins born between 1881-1930. *Journal of American Statistical Association*, **87**, 17-24.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd ed., John Wiley & Sons, New York.

Kleinbaum, D. G. and Klein, M. (2005). *Survival Analysis*, 2nd ed., Springer-Verlag, New York.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd ed., John Wiley & Sons, Hoboken, NJ.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

Naskar, M., Das, K., and Ibrahim, J. G. (2005). A semiparametric mixture model for analyzing clustered competing risks data. *Biometrics*, **61**, 729-737.

Ross, E. A. and Moore, D. (1999). Modeling clustered, discrete, or grouped time survival data with covariates. *Biometrics*, **55**, 813-819.

Scheike, T., Sun, Y., Zhang, M.-J., and Jensen, T. K. (2010). A semiparametric random effects model for multivariate competing risks data. *Biometrika*, **97**, 33-145.

Williams, R. L. (1995). Product-limit survival functions with correlated survival times. *Lifetime Data Analysis*, **1**, 171-186.

Yau, K. K. W. (2001). Multilevel models for survival analysis with random effects. *Biometrics*, **57**, 96-102.

Yue, H. and Chan, K. S. (1997). A dynamic frailty model for multivariate survival data. *Biometrics*, **53**, 785-793.