24th International Workshop on Household Survey Nonresponse. August 2013.

Generating Nonresponse Adjustment Variables using Sequence Analysis of Call Record Data

Mark Hanly
PhD Student
University of Bristol
mark.j.hanly@bristol.ac.uk

1. Introduction

As a technique to analyse call records, sequence analysis is gathering increasing interest. Kreuter and Kohler (2009) have proposed its use as a method to generate nonresponse adjustment variables from call records. Pollien and Joye (2011) have used sequence analysis to create a typology of households which differed in response likelihood and other characteristics of interest. Durrant et al (2013) have suggested using sequence analysis to analyse interviewer calling patterns. To date however, there has been no incisive application of the method which promotes sequence analysis as a valuable addition to the survey methodologist's toolkit. The aim of this mini-paper is to demystify the sequence analysis method, to clarify what it can and can not do and to promote discussion on specific research questions which may benefit from this method.

2. The Motivation for Exploring Sequence Analysis

The motivation for exploring SA is clear. Paradata are potentially useful for nonresponse analyses, being available for respondents and nonrespondents alike, and call record data such as those presented in Table 1, are a particularly rich subset of paradata. However, call records are distinct from other interviewer observations in that they are time variant, and this longitudinal structure necessitates an appropriate analytic technique.

Table 1: Example of call records for one household.

Household ID	Call No.	Time	Date	Outcome
101	1	11:00	04/08/2013	Noncontact
101	2	14:30	04/08/2013	Noncontact
101	3	18:45	05/08/2013	Contact - Appointment Made
101	4	10:00	11/08/2013	Interview

The optimal method to summarise or model these complex data remains an open research question. One solution is to aggregate over all calls to generate summary variables at the household level, such as the total number of calls made to a household or an indicator for whether an evening call was ever made. However, aggregating in this way potentially ignores much of the information contained in call records and it is easy to imagine how the full call history at a household may be more revealing than summary statistics such as the overall number of calls. For example, three noncontact calls at a household followed by a contact producing an appointment suggests a different disposition than at a household where an appointment on the first call is followed by three noncontacts.
An alternative to aggregating over all calls at a household is to model the likelihood of an event occurring (such as making contact) given what has previously occurred, using event history analysis. This approach allows the inclusion of time-variant call records in the nonresponse model and has been used to predict best times of contact and the likelihood of participation, given previous call outcomes (Durrant et al 2011; 2013).

Sequence analysis is simply an alternative method to analyse data with this longitudinal or sequential structure. Here, instead of focusing on transitions between outcomes, the unit of analysis becomes the full series of calls at each household. This shift is motivated by the

premise that important information is potentially contained in the full sequence which cannot be captured either by aggregating the call records or analysing them on a call-by-call basis.

3. Preparing Data for Sequence Analysis

In the context of call records, analysis to date has focused on the sequence of call outcomes at a household (Pollien and Joye 2013; Kreuter and Kohler 2009). Thus the elements of the sequences are the actual outcomes recorded by the interviewer. The full set of potential elements is known as the sequence alphabet and in this context comprises all the possible calls outcomes at a household. Adopting this approach, the unit of analysis for the household described in Table 1 would be the sequence

(a) Noncontact – Noncontact – Appointment – Interview

The elements which make up the alphabet are of course at the discretion of the researcher and there is no reason why the above call outcomes could not be recoded as a binary indicator for contact, so that the alphabet has only two possible elements, contact (1) and noncontact (0), making the above sequence

(b) 0 – 0 – 1 – 1

In fact the focus on the call outcome is not necessary. Depending on the research question it may be preferable to focus on sequences of call times, in which case the sequence for analysis becomes

(c) Morning – Afternoon – Evening – Morning

Or perhaps the interaction of call timing and call outcome is of interest in which case the alphabet might be coded to reflect both domains

(d) <5pm noncontact – <5pm noncontact – >5pm contact – <5pm noncontact

4. Summarising Sequences

Once the call sequences are defined for each household, the next step is to summarise the set of all sequences in some way. Kreuter and Kohler (2009) examine six summary indicators in their search for post-survey adjustment variables. In order to make this step explicit, Table 2 below reproduces presents the value of these six summary measures for the call records presented in Table 1.

Table 2. Reproduction of the summary measures defined by Kreuter and Kohler (2009) for one household[1]

Household	Measure	Definition	Value
101	1	Number of Contact Attempts	4
101	2	Proportion of Noncontacts	0.5
101	3	Number of different elements / length	0.75
101	4	Number of different episodes / length	0.75
101	5	Disturbed interaction / length	0
101	6	Multidimensional scaling of distance matrix	0.32

[1] This is just an example rather than an exact replication - Kreuter and Kohler specify a different alphabet and perform other data cleaning. Please see the original paper for exact definitions of the summary variables.

The first thing that should be noted from Table 2 is that the output from SA is simply a variable (or series of variables) defined at the household level, summarising the sequence of calls to the household in some way. It is worth clarifying that the first five of the six indicators examined above are, strictly speaking, aggregations over all calls to the household, rather than outputs of full SA. For example no specialist knowledge or software is required to calculate the proportion of calls at a household which were noncontacts. The SQ-Ados in Stata (Brzinsky-Fay et al 2006) undoubtedly simplify the calculation of these measures but there is no reason why they could not be calculated directly using any basic statistical software. The sixth measure is somewhat different, and can be considered a true output of SA. The following sections will describe the process used to generate this measure.

5. Understanding the Sequence Analysis Algorithm

Sequence analysis uses a metric to quantify the similarity or "distance" between any two sequences. Calculating this distance for all pairs of observed sequences produces a distance matrix of order $N \times N$, where N is the number of unique observed sequences. This will be a zero-diagonal, symmetric matrix as $dist(A,A) = 0$ and $dist(A,B) = dist(B,A)$.

This matrix is the raw output from SA and on its own it is not much use. Some multivariate technique is needed to reduce this matrix to a manageable number of dimensions. The most common techniques for data reduction are cluster analysis and multidimensional scaling. The sixth measure presented by Kreuter and Kohler (2009) is the first dimension extracted using multidimensional scaling of the distance matrix generated from SA. It is up to the analyst to interpret what aspects of the call records (or more accurately the distance matrix) are being measured by the extracted dimension(s). If cluster analysis is chosen as a method to reduce the distance matrix, the extracted clusters groups similarly need to be interpreted and labelled.

I would now like to focus on the metric used to generate the distance matrix. The most common SA metric is based on a process known as Optimal Matching (OM). To calculate the distance between two sequences using OM the elements of the two sequences are aligned and one sequence is edited such that the aligned sequence elements match identically. Two edit operations are allowed in the process, referred to as substitution and indel. A substitution is a direct swap of an element in one sequence to match its counterpart in the other. Indels (a contraction of insertion and deletion) occur when missing elements are inserted in one sequence, or equivalently superfluous elements are deleted in the other. Each edit required in the matching process incurs a penalty cost, and the sum of these costs form the basis for the distance between the sequences. OM uses an algorithm to find the alignment of the two sequences which offers the cheapest combination of substitutions and indels in this matching process.

The structure of the distance matrix generated through SA depends on the costs assigned to the indel and substitution edits. In the social sciences, it is increasingly understood that the cost settings carry the substantive meaning of the sequences (Halpin 2008). The costs dictate what constitutes a similar trajectory or sequence of events. The role of substitutions and indels in the alignment process differs. Substitutions allow for the swapping of elements, thus the magnitude of the cost typically reflects the relative similarity between the elements to be swapped. Indel costs, on the other hand, are closely linked to temporality (Lesnard 2010). Insertions decelerate time, while deletions imply acceleration.

As an example of how edit costs carry contextual information, consider the distance calculation for two sequences of different length. Insertions or deletions will always be necessary in this situation, with the number of required indels equivalent to the disparity in length of the two sequences. As a result, increasing indel costs exaggerates the distance

between sequences of different length. Of course, whether differences in length between sequences should be reflected in the distance measure depends on the research question and the perspective of the analyst. The sequence length may be an important characteristic from the perspective of survey management, who accrue expenses for each additional interviewer visit. On the other hand, for the statistician primarily concerned about response rates, the number of calls may be a less important characteristic compared to the actual call outcomes. The substantive meaning of call outcomes can be reflected in the substitution costs. For example if a weekday morning call is considered relatively similar to a weekday afternoon call compared to a weekend afternoon call, the costs for substituting these elements can be appropriately set to reflect these differences.

6. Generating Nonresponse Adjustment Variables using Sequence Analysis

As mentioned above, the use of sequence analysis as a tool to generate nonresponse adjustment variables was first proposed by Kreuter and Kohler (2009). Their analysis, based on three waves of ESS call records gathered across fourteen countries, found that indicators derived from sequence analysis were predictive of the response outcome but not of key survey variables. This result led the authors to conclude that sequence analysis was not useful for generating nonresponse adjustment variables, but that it may have applications in fieldwork management.

Kreuter and Kohler (2009) focused only on call outcomes and not call times. The one measure generated through OM relied on the default substitution and cost settings. To explore the impact of different cost settings and the use of time-of-call information, we applied sequence analysis to call records while varying these algorithm inputs. This analysis is summarised below.

Data The analysis was applied to wave one call records from the Irish Longitudinal Study of Ageing (TILDA). TILDA is a prospective study of the residential population over the age of fifty, in the Republic of Ireland. As a large scale, interviewer-mediated, face-to-face household survey, the call records generated here were similar to those from the ESS where SA has been previously applied.

Methodology We repeatedly analysed the data using SA to test the sensitivity of the output to different cost settings. Nine different cost settings, based on the cross-classification of three substitution costs and three indel costs, were applied separately to sequences of call outcomes and sequences of call times. In addition three sequences simultaneously combining information on both call time and outcome were analysed. This generated a total of 21 scenarios, each of which was used to generate a distance matrix through SA. Each matrix was reduced to a categorical variable with four categories using cluster analysis. These output variables were individually included in the baseline nonresponse model[2] to quantify the added effect on bias reduction of including variables derived from SA of call records.

Results Under certain cost settings, including the variable derived from SA to the baseline nonresponse model increased the correlation between the corresponding weight and the response indicator. Including sequence derived variables to the baseline nonresponse model either did non impact or reduced the correlation between the corresponding probability weight and survey variables.

[2] The baseline nonresponse model included the following household level predictors: number of calls; proportion of noncontacts; observation on type of dwelling; observation on condition of dwelling; indicator for Dublin address.

Conclusion In this application, adjustment variables derived from SA of call records did not contribute to bias reduction beyond what could be achieved from aggregated call records combined with time-invariant interviewer observations.

6. Discussion

At this point it might be worth pausing and reminding ourselves why we are interested in SA in the first place. We want to generate summary measures from call record data suitable for nonresponse adjustment (or other purposes). Sequence analysis has the potential to summarise complex call records with a small number of household level variables. However, many summary measures are readily available and do not require specialist software, complicated algorithms or cost-setting decisions. Sequence analysis would only ever be useful if the derived variables displayed stronger correlations with the response indicator and survey variables than observed for these simpler measures. In terms of reducing nonresponse bias through post-survey adjustment, the evidence discussed above suggests that sequence derived indicators do not out-perform more straight-forward paradata variables such as interviewer observations or aggregated call records.

It should be acknowledged that the failure to find such measures is most likely not with the tool (sequence analysis) but rather with the data. There is only so much that can be inferred about a household based on the timing and outcome of interviewer visits, and it would seem that this limit is reached with simple summaries such as the number of calls and proportion of noncontacts. Sequence analysis can generate valid summary measures but in this application they do not add enough information to make the process worth while.

Is sequence analysis a useful tool for survey methodologists? There are (at least) two remaining areas where SA may have useful applications, and as Kreuter and Kohler concluded, these are in the area of fieldwork management. First, as suggested by Durrant et al (2013), SA may be a useful tool to analyse interviewer calling behaviours. Second, SA may potentially be useful for identifying special subgroups to be targeted with particular recruitment designs.

<u>Points for further discussion</u>

1. For what specific research questions could sequence analysis deliver more insight compared to traditional analytical approaches?

2. Call records do not seem promising as adjustment variables due to a low correlation with survey outcomes. Will improved paradata quality ameliorate this issue or will there always be a limit to what can be inferred about a household from call records?

References

Brzinsky-Fay, C., Kohler, U., & Luniak., M. (2006). Sequence analysis with Stata. *Stata Journal, 6*(4), 435 - 460.

Durrant, G., D'Arragio, J., & Steele, F. (2011). Using paradata to predict best times of contact, conditioning on household and interviewer influences. *Journal of the Royal Statistical Society: Series A, 174*(4), 1029 - 1049.

Durrant, G. B., D'Arrigo, J., & Steele, F. (2013). Analysing interviewer call record data by using a multilevel discrete time event history modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 251-269.

Durrant, G., Maslovskaya, O., Smith, P., & D'arrigo, J. (2013). *Using Sequence Analysis to Better Understand Interviewer Calling Behaviours: An Example from the UK Understanding Society Survey*. Paper presented at the 5th Conference of the European Survey Research Association, Ljubljana.

Halpin, B. (2008). Optimal Matching Analysis and Life Course Data: the importance of duration *Department of Sociology Working Paper Series - WP2008-01*: University of Limerick.

Kreuter, F., & Kohler, U. (2009). Analyzing Contact Sequences in Call Record Data. Potential and Limitations of Sequence Indicators for Nonresponse Adjustments in the European Social Survey. *Journal of Official Statistics, 25*(2), 203 - 226.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research, 38*(3), 389 - 419.

Pollien, a., et Joye d., "A la poursuite du répondant? Essai de typologie des séquences de contact dans les enquêtes" in *Parcours de vie et insertions sociales*, Edition Seismo, Zürich, 2011, pp. 189-212