Analyzing Longitudinal Data from Complex Surveys Using SUDAAN

Darryl Creel
Statistics and Epidemiology, RTI International, 312 Trotter Farm Drive, Rockville, MD, 20850

Abstract

SUDAAN: Software for the Statistical Analysis of Correlated Data (SUDAAN) can be used to analyze data from surveys with complex designs. A possible feature of a complex survey design is clustering. One way in which clustering can occur is to have the same information collected on a sampling unit at different points in time. This type of clustering creates data that may be referred to as longitudinal, panel, or repeated measures data. This paper provides an example of longitudinal data analysis using SUDAAN. The example covers the structure of the data and data set; analytic strategies and interpretation; and the implementation of the analytic strategies using SUDAAN.

Keywords: Longitudinal Data Analysis (LDA), SUDAAN.

1. Introduction

One of the major uses of longitudinal data is to analyze trends, or change, over time. The SUDAAN team at RTI International often receives questions about how to conduct longitudinal data analysis using SUDAAN. This paper provides an answer to this question. In Section 2, we discuss longitudinal data. In Section 3, we discuss various survey designs over time. In Section 4, we examine the variance of a difference of two means. In Section 5, we discuss the data structure SUDAAN requires for the data analysis. In Section 6, we examine the SUDAAN code to analyze longitudinal data and discuss a cautionary note. In Section 7, we provide an example to illustrate the possible differences that may occur when one does and does not account for the longitudinal structure of the data. Finally, in Section 8, we provide some recommendations and cautions.

2. Longitudinal Survey Data

Longitudinal data measures the same characteristics of the same sampling unit over time. For example, in a longitudinal health survey of children, measurements such height and weight may be measured each time the survey is conducted to create the child's body mass index (BMI). Some of the goals of collecting longitudinal data are to produce population estimates over time, study change over time, and/or study variables that affect change over time. Continuing the longitudinal health survey children example, researchers may be interested in the population estimates of BMI for specific subgroups over time. Researchers may also be interested in studying the change in BMI overtime and what variables are related to the change in BMI over time.

Longitudinal data from a survey with a complex survey design has the added complication of accounting for this complex survey design in the analysis. SUDAAN can account for different aspects of the complex survey design, e.g., stratification, clustering, and differential weighting, while conducting longitudinal data analysis.

3. Survey Designs over Time

There are four common design for surveys conducted over time: repeated surveys, panel survey, rotating panel survey, and split panel survey.[1]

In the repeated survey design, "similar measurements are made on samples from an equivalent population at different points of time, but without attempting to ensure that any elements are included in more than one round of data collection."[2] "Its particular strength is that at each round of data collection it routinely selects a sample of the population existing at that time."[3] "The major limitation of a repeated survey is that it does not yield data to satisfy objectives [of estimating change at the element level between two time points and other components of individual change] and [aggregate data for individuals over time]."[4]

[1] There is a detailed explanation of these designs by Greg Duncan and Graham Kalton in "Issues of Design and Analysis of Surveys Across Time," *International Statistical Review*, Vol. 55, No. 1, pp.97-117. This section is a brief summary of some of their discussion.
[2] Duncan and Kalton 100.
[3] Duncan and Kalton 101.
[4] Duncan and Kalton 101.

"A panel survey is one in which similar measurements are made on the same sample at different points in time."[5] The major advantage of a panel survey over a repeated survey is its much greater analytic potential. It enables components of individual change to be measured … and also the summation of a variable across time."[6] It "can be much more efficient than a repeated survey for measuring net change."[7] "[T]wo major potential problems with panel surveys are panel losses through nonresponse and the introduction of new elements to the population as time passes."[8]

"In a panel survey, sample elements are, in principle, kept in the panel for the duration of the survey. In a rotation panel survey, sample elements have a restricted panel life; as they leave the panel, new elements are added. … The limited membership in a rotating panel acts to reduce the problems of panel conditioning and panel loss in comparison with nonrotating panel survey, and the continual introduction of new sample helps to maintain an up-to-date sample of a changing population."[9]

"A split panel survey is a combination of a panel and a repeated or rotating panel survey, as advocated in Kish (1983, 1986)."[10]

4. Variance of a Difference of Two Means

The section focuses on the repeated cross-sectional survey, the panel survey, and the rotating panel survey. The split panel survey is not discussed in the section, but recall that it is a combination of a fixed panel and new sample elements from either a repeated or rotating panel.

The repeated cross-sectional survey design uses the same survey design each year but samples a different group of members each year. This approach is conceptually straight forward, samples from the current population, and avoids the complexity of a panel survey, fixed or rotating. However, it is difficult to tell if the differences are simply due to the different samples or are a true difference in the outcome variable. Also, when analyzing the difference of two means between years the repeated cross-sectional survey design is not the most efficient survey design. Because of the independent samples, the variance of the difference of two means is relatively large compared to other methods. Using simplifying assumptions that the variances of the means are equal for the two time periods, $S_1^2 = S_2^2 = S^2$, and that the sample sizes are equal for the two time periods, $n_1 = n_2 = n$, the variance for the difference of two means, where m_1 is the mean for time period one and m_2 is the mean for time period two, for repeated cross-sectional surveys is

$$\operatorname{var}(m_2 - m_1) = \tfrac{2}{n} S^2 .$$

Contrast this with the most efficient survey design to measure differences between time periods which is the fixed panel survey design, i.e., a single sample on which data is collected at different points in time. The efficiency of the fixed panel survey depends on the correlation between the outcome variable at two time periods, ρ_{12}. Using the same assumptions that were used for the variance of a difference of two means for repeated cross-sectional surveys, the variance of a difference of two means for the fixed panel survey is

$$\operatorname{var}(m_2 - m_1) = \tfrac{2}{n} S^2 (1 - \rho_{12}) .$$

Comparing the variance of the difference of two means for repeated cross-sectional survey and a fixed panel survey, the variance of the difference of two means for the fixed panel survey has a smaller variance by the factor $(1 - \rho_{12})$. Consequently, the higher the correlation between the two time periods is the smaller the variance of the difference of two means.

Although the fixed panel survey is the most efficient at measuring differences between years, it is not without its limitations. Generally, a fixed panel survey has three limitations: the panel is selected at one point in time, panel attrition, and panel conditioning. If the population is changing, then selecting the sample once and not every year may cause the sample to become less and less representative of the population and bias the survey estimates. Panel attrition can arise because of the added response burden for panel members to provide data every year. Panel conditioning means that panel member's responses change in some way because they are part of the panel.

A rotating panel survey design can mitigate the problems associated with the fixed panel survey

[5] Duncan and Kalton 101.
[6] Duncan and Kalton 102.
[7] Duncan and Kalton 102.
[8] Duncan and Kalton 103.
[9] Duncan and Kalton 103.
[10] Duncan and Kalton 104.

without losing all of the benefits of the reduction in the variance of the differences. In a rotating panel survey design, panel members are only retained in the panel for a set period of time and new panel members are brought into the panel. This mitigates the panel attrition and panel conditioning which is a concern for a fixed panel. Also, because of the rotation in of new groups into the panel at each time period, the panel is not static and is updated with new panel members from the current population. This will account for any changes in the population over the course of the life of the survey. Because there is not complete overlap, there will be some loss in the efficiency of the rotating panel that is proportional to the size of the panel that does not overlap from one time period to the next. That is, the formula for the variance of the difference of two means has an added term that represents the amount of overlap, λ,

$$\mathrm{var}(m_2 - m_1) = \tfrac{2}{n} S^2 (1 - \lambda \rho_{12}) .$$

With λ = ½, the variances will only benefit by half of the correlation of the outcome variable between the two time periods. If λ = 1, i.e., there is complete overlap, then the variance is equal to the fixed panel variance. If λ = 0, i.e., there is no overlap, then the variance is equal to the repeated cross-sectional survey variance.

5. Structure of the Data Sets

Let us assume that there are two data sets. One data set is from 2004 and the other is from 2005. The two data sets do have some overlap. That is, there are some primary sampling units (PSU) that are on both data sets. Also, each of the data sets has a common set of analytic variables that are not shown in the following tables. Table 1 shows the stratum, PSU, and year for the 2004 data set.

Table 1: 2004 Data Set Showing the Stratum, Primary Sampling Unit, and Year

Stratum	PSU	Year
1	***1***	***2004***
1	***2***	***2004***
1	***3***	***2004***
1	***4***	***2004***

Table 2 shows the same information for the 2005 data set. Note that the data in Table 1 are italicized and bolded; the data in Table 2 are not. This distinction is carried through in the other tables.

Table 2: 2005 Data Set Showing the Stratum, Primary Sampling Unit, and Year

Stratum	PSU	Year
1	2	2005
1	3	2005
1	4	2005
1	5	2005

In order to perform the longitudinal data analysis, SUDAAN requires that the two separate data sets be combined into one data set. The combined data set is shown in Table 3.

Table 3: Combined 2004 and 2005 Data Set Showing the Stratum, Primary Sampling Unit, and Year

Stratum	PSU	Year
1	***1***	***2004***
1	***2***	***2004***
1	***3***	***2004***
1	***4***	***2004***
1	2	2005
1	3	2005
1	4	2005
1	5	2005

SUDAAN also requires that the data set is sorted by the variables on the nest statement. The nest statement that will be used in our first set of example code contains year, stratum, and PSU. The data set sorted by these variables in shown in Table 3. This sorting used year as a stratification variable. Consequently, the results using this data set are similar to results from a repeated cross-sectional survey. That is, there is no benefit for the correlation between responses over the two time periods.

The nest statement that will be used in our second set of example code contains stratum and PSU. The data set sorted by these variables in shown in Table 4. Sorting by stratum and PSU, and not using year, creates a data set that has year clustered within PSU. Consequently, the results using this data set are similar the panel design, although we do not have complete overlap. We still have the advantage of the variance reduction because of the overlap that we do have and the correlation between the responses.

Table 4: Combined 2004 and 2005 Data Set Showing the Stratum, Primary Sampling Unit, and Year Sorted by Stratum and PSU

Stratum	**PSU**	**Year**
1	***1***	***2004***
1	***2***	***2004***
1	2	2005
1	***3***	***2004***
1	3	2005
1	***4***	***2004***
1	4	2005
1	5	2005

6. SUDAAN Code for Longitudinal Data Analysis and Cautionary Note

6.1 SUDAAN Code

The focus of the following SUDAAN code is to calculate the contrast, and associated information, between 2007 and 2006. Often we see examples of SUDAAN code, that contain the year variable as a stratification variable as shown in the following SUDAAN code:

```
proc descript data = dataSet design = wr;
nest year stratum PSU / psulev = 3;[11]
weight aWeight;
class year / noFreqs;
var aVar.;
contrast year = ( -1 1 ) / name = "2007 – 2006
Contrast";
print nsum mean semean t_mean p_mean;
run;
```

Using the year variable as a stratification variable, does not allow us to benefit from the longitudinal structure of the data. That is, the observations for a PSU are classified across multiple years and not clustered within PSU.

One way to capture the multiple years of data collected for a PSU is not to use year as a stratification variable. The following code only includes the stratum variable as the stratification variable:

```
proc descript data = dataSet design = wr;
nest stratum PSU;
weight aWeight;
class year / noFreqs;
var aVar.;
contrast year = ( -1 1 ) / name = "2007 – 2006
Contrast";
print nsum mean semean t_mean p_mean;
run;
```

Consequently, this SUDAAN code treats the years as clustered within the PSU and allows us to take advantage of the longitudinal structure of the data.

6.2 Cautionary Note

The focus of the previous SUDAAN code is using a combined data set to produce contrasts between years. The number for the degrees of freedom (d.f.) for our simple example that SUDAAN uses is correct for this purpose; it would use 4 d.f. There is a caution when one analyzes a single year's data. Each single year data set for our simple example would have 3 d.f. and this is what SUDAAN would use for the single year data sets. For the combined data set, SUDANN would use 4 d.f. even for the single year analysis. Consequently, one should use the DDF = 3 option for the combined data set for single year analysis or use the single year data sets.

7. Example

We have included one example using simulated data so that one can see the potential impact of not taking the longitudinal data structure into account, and possibly getting smaller standard errors. The simulated data set had 500 observations in a single

[11] The psulev = 3 option on the nest statement tells SUDAAN that the third variable on the nest statement is the PSU which implies that the first two variables on the nest statement are stratification variables. A full description of the SUDAAN language can be found in the *SUDAAN Language Manual, Release 9.0.*

stratum, a $\lambda = 1$, and $\rho = 0.66$. The results of analyzing that data treating it as a repeated cross-sectional data structure and a panel data structure are presented in Table 5.

Table 5: Results of Simulated Data Set Analyzed as a Repeated Cross-Section Survey and a Panel Survey

	Repeated Cross-Section	**Panel**
Contrast Mean (CM)	0.11	0.11
SE CM	0.06	0.04
Lower Limit 95% CI CM	-0.01	0.04
Upper Limit 95% CI CM	0.23	0.18
T-test CM	1.78	3.05
P-Value T-test CM	0.0757	0.0024

Note that the estimates for the contrast mean are the same, but the standard error estimates for the contrast mean is smaller for the panel survey than for the repeated cross-section survey. This difference carries through to the confidence intervals and testing, which results in a statistically significant difference at the $\alpha = 0.05$ level for the panel but not for the repeated cross-section survey. Hence, the analytic approach has the possibility of making a difference in your interpretation of the output.

8. Recommendations

The main point is to take advantage of the longitudinal data structure and possibly smaller standard errors. One can account for the longitudinal data structure easily using SUDAAN to produce contrasts. Finally, use a data set that combines years of information for contrasts or single year analysis using the DDF option. One could also use the single year data sets for the single year analysis.

References

Duncan, Greg and Kalton, Graham (1987), "Issues of Design and Analysis of Surveys Across Time," *International Statistical Review*, Vol. 55, No. 1, pp. 97-117.

Kish, Leslie (1983), "Data Collection for Details over Space and Time," *Statistical Methods and the Improvement of Data Quality*, Ed. T. Wright, New York: Academic Press, pp. 73-84.

Kish, Leslie (1986), "Timing of Surveys for Public Policy," *The Australian Journal of Statistics*, Vol 28, pp. 1-12.

Research Triangle Institute (2004), *SUDAAN Language Manual, Release 9.0*, Research Triangle Park, NC: Research Triangle Institute.