The Effects of Time Dimension and Interview Period Research on Sample Loss in SIPP

Rao Xiong

3700 Walnut Street Philadephia, PA 19104 .USA

xiongrao@gse.upenn.edu

Keywords: The Survey of Income and Program Participation, SIPP, Statistics and data analysis,Linear regression model,Enterprise customer investigation and management.

Abstract: The Survey of Income and Program Participation (SIPP), is widely used in many fields of the most effective methods to analyze the data. From the history of the development of SIPP, the panel was redesigned in 1996, in order to improve the quality of information, improves the sample size and the length of the two. A change usually affect the sample loss. This paper deeply analyses the reasons of loss, to find ways to reduce sample loss, improve the survey quality possible.

Introduction

Survey of income and program participation by the United States Census Bureau designed a housing survey template, it carries on the statistics and estimates of the resident population of a region. The need is a final location survey, since 1984 has multiple versions. The survey was conducted by the interviewer management, the scope of the investigation is personal visit and telephone consultation for more than 15 years for all family members, fill in the questionnaire design. The need to provide the information is very wide, including not only the dynamic range of income, including the level of education, establish the demographic pattern and health insurance template through the investigation, to hundreds of thousands of academic journals and the economic field In the field of social science, education, health and other areas that need to use the data to provide a reference, while the survey is also widely used in various industries

Because SIPP is a longitudinal survey, the reason it is conduced is not only to focus on providing correct and comprehensive information that supports analysis in multiple programs, but also to capture changes in household and family composition over multiple periods such as the change of nation's economic well-being over time.

1996 Redesign

The first panel is 1984 Panel which started in October 1983. Originally, when SIPP started, each panel was designed to interview around 20,000 samples. The sample was divided into four subgroups and a complete interview of all four subgroups was a wave. Interviews was conducted every 4 months and households were required to answer questionnaires based on the recall of past four months. Each panel should have 32 months in total which covered 8 waves. However, some panels were terminated earlier due to insufficient budget. As shown in table 1, 1986 Panel and 1987 Panel only have 7 waves, and 1988 Panel only has 6 waves. 1992 Panel were extended to 10 waves and 1993 Panel were extended due to the effort of redesign asked by Census Bureau in 1990.

In 1996, after summarizing of first nine years of SIPP, new recommendations have been implemented. To improve the quality of estimates, the sample size has increased to a target of 37,000 households and each panel has increased to a single 4-year panel including twelve to thirteen waves instead of old 32-month panels with eight waves. But 2001 Panel only has 9 waves because of the fund issues and 2008 Panel has up to 16 waves.

The design of a longitudinal survey is usually changed to improve the quality of the data. For SIPP, the sample bases and time dimension have significantly increased after 1996 redesign. The intention of such increasing is to provide better estimate of the whole population which is the United States residents, and of the trend of changes over time. However, will the change of sample size and time dimension have negative effects on sample loss that decrease the quality of data consequently?

Time dimension

The time dimension of a longitudinal survey is constituted of the length of panel, length of each wave and the number of waves. The time dimension can be easily determined by the purpose of the survey which means the survey can be designed with the minimum length that meet the requirement of the interest demand, but the length is also seriously depended on the costs of collection (Trivellato, U., 1999).

In SIPP, each wave has constant length of four months in every panel since 1983 when it was started. The important changes are focused on the length of panel and the number of waves. Before the redesign in 1996, the length of panel is basically depended on the number of waves. In the initial design, eight waves are implemented in each panel, but there is a new panel begin in each year, so there is an overlapping of waves when the new panel begin but there are still waves from old panels processing. The overlapping panels cost lots of budget, so this may be one of the reason that it is abandon in the redesign.

The reduce of panel length is unavoidable due to the limitation of budget, but the reason to extend number of waves has not been explained specifically except by saying to improve the quality by Census of Bureau. However, the increase of length of panel may make the work of precisely tracing the households become harder and may cost more. In addition, there is no clue that whether the households are been told the lengths of panel during the interview of first wave. So, it is possible that some households will refuse or avoid the interview of future waves if they think number of wave is too many for them.

Table 1 – Number of Waves

Panel	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1996	2001	2004	2008
Waves	8	8	7	7	6	8	8	8	10	9	12	9	12	16

Sample Size

The sample size of a longitudinal survey should consider the dynamic aspects of the survey. Since the same samples will be repeatedly observed over time, the sample size should both represent the population and be cost efficient (O' Muircheartaigh, 1996). In panel survey, the target population which might be a set of resident in households is especially dynamic since the population is both increasing through births and immigration and decreasing through deaths and moving out. Also, the households as the basic units is change continually at the meantime.

While dealing the issue of changing in target sample, SIPP interviews the individuals at first wave, and the sample size will then be modified as the time passes. The following sample after the first wave will deplete by removing the number of death and moves out of households, and it will augment by adding the number of births while there are new children reach age of fifteen. Furthermore, information is collected from individuals who become relevant to the sample as they share living quarters with sample person (David, M., 1985).

Before the 1996 redesign, the target sample size is about 20,000 for each panel. Although the expectation has rarely been met with budget issues, the approximate sample size for each panel at first wave is ranging from 11,000 to 23,000, except for 1989 Panel with sample size around 4,000 since its rest of waves has been combined into 1990 Panel due to insufficient budget. With overlapping panel design, each panel lasts 32 months on average. Therefore, during each year from 1986 to 1993, there are three panels implementing at the same time which has about 60,000 samples been interviewed.

After 1996, the fixed panel length of four years has been suggested instead of 32 months. During each year, only approximate the same sample target will be interviewed for that panel. Therefore, to better represent the population, the size of target population has increased from 20,000 to 37,000, but the sample size of the first wave of 2008 Panel and 2004 Panel was as big as 52,000, and 2001 and 1996 Panel also have sample size of about 40,000. Although the starting sample size has significantly enlarged since 1996, the number of samples been interviewed each cannot reach the same size with the design of overlapping panels.

Sample Loss

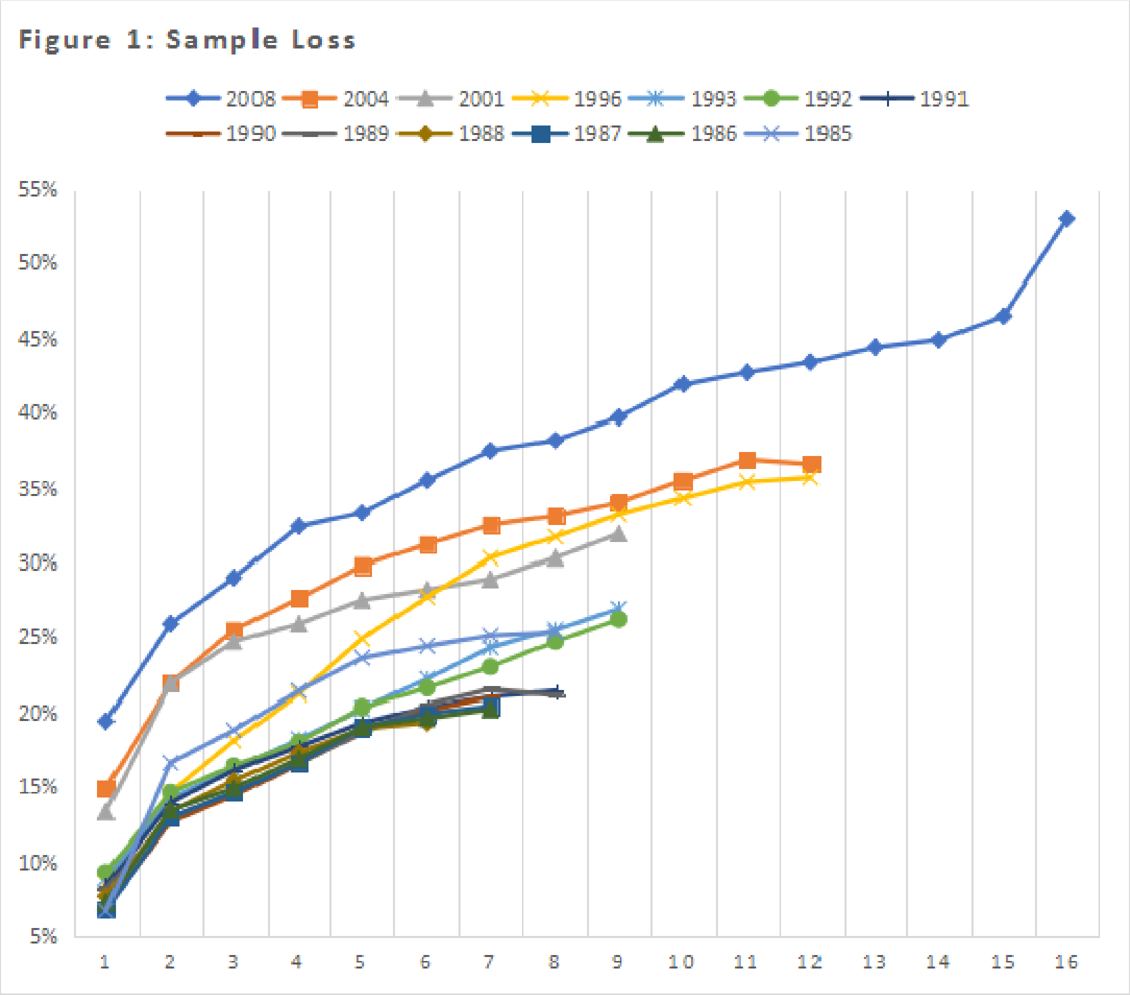
The nonresponse of longitudinal surveys, also called attrition, is different from other surveys as there is a progressive loss of sample members during the whole process of each panel which can cause serious problems. So, it is an accumulated sample loss for a panel. If the attrition is random, only the efficiency of estimates will be effect, but if the attrition is not random, statistical biases will be result without weighting or adjusting (Heckman, 1979). Therefore, the sample loss from potential target during each panel can cause serious statistical biases from the information provided.

Sample loss can result from the tracing of sample during the life of panel. There are more problems for tracing sample members when there are more panel waves because of the increased proportion of movers and the difficulty of tracking (Duncan, G., & Kalton, G., 1987). So, after the redesign in 1996, the increasing in the number of panel wave can lead to the heavier work of sample tracing which not only increases the cost but may also result in the problem of increasing sample loss. The initial sample size should also affect the level of sample loss. The bigger the sample is, more efforts should be made to reach the whole sample including the budget cost and time used.

For SIPP, the sample loss rates for each wave from 1985 Panel to 2008 Panel are calculated by adding the sample loss from four subgroups in a wave. Since the SIPP surveys also includes the member changes in adjusted sample of following waves, to simply sum up the sample loss rate of each wave is incorrect while the expansion of non-interviewed sample is impossible to be determined. Therefore, a growth factor that can estimate the non-interviewed sample expansion based on the growth in interviewed sample is used to weight the number of non-interviewed housing units (Jason M. F, 2016). The equation to calculate the sample loss is as follow:

$$Sample\ Loss = \frac{(A_1 \times GF) + A_C + D_C}{I_C + (A_1 \times GF) + A_C + D_C}$$

“ : growth factor associated with the current wave. : weighted number of interviewed households in the current wave. 1: weighted number of Type A non-interviewed households in Wave 1, : weighted number of Type A non-interviewed households in the current wave. : weighted number of Type D non-interviewed households in the current wave (Jason M. F, 2016).”



Analysis

The accumulated sample loss rate for each panel has been shown in Figure 1. Vertical axis presents the sample loss rate in percentage. Horizontal axis shows the panel waves. The trend of accumulated sample loss in each panel is drawn in deferent color. The most significant observation is that the sample loss rate is higher accompanied by increase of number of waves. It is also obvious to see that the starting points of 2001, 2004, and 1008 Panel are outstanding besides other panels which have starting point closed to each other. For panels before 1996, the sample loss rate of 1985 is higher than other panels from since the second wave. There is a possibility that since 1985 Panel is the second panel since SIPP started, the rule of tracing sample members was not well developed. With panel length of seven waves and 8 waves, the lines of sample loss rate from 1986 to 1991 Panel are very close. The sample loss rate of 1996 Panel is above all panels before redesign from wave 4, and the sample loss rate of 2001, 2004, and 2008 Panel are above all other panels before redesign at every wave. Therefore, the sample loss rate of panels after redesign is generally higher than panels before redesign. However, the difference of the sample loss rate may relate with either the number of waves, the sample size or these two factors together.

Figure 2 represents the relationship between size of eligible households of each wave which represents the adjusted target sample size of wave for panels from 1985 to 2008 and the number of households that have not been interviewed in that wave. The vertical axis shows the none interviewed households while the horizontal axis shows the eligible households. From the figure, a positive relationship can be find out which means the bigger the sizes of eligible households are, the bigger the sizes of nonresponses are. The slope of the regression line is 0.2 with p-value less than 0.05 and the R-squared of the model is 0.73. With 95 percent confidence interval, there is a significant evidence to show that target sample size is affecting the size of nonresponse.

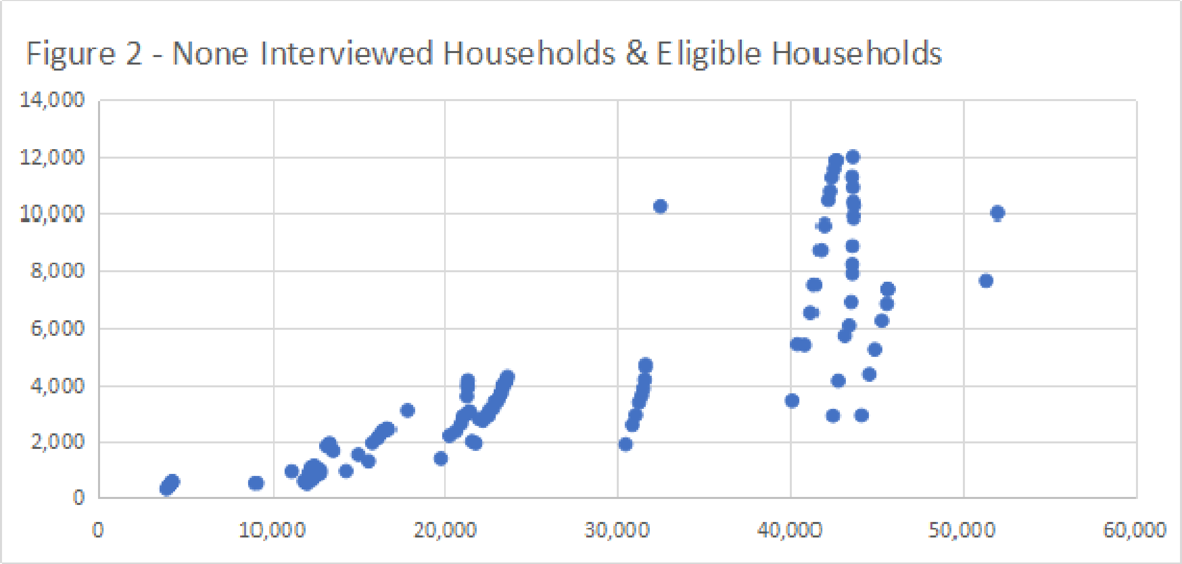


Figure 2 - None Interviewed Households & Eligible Households

Table 2

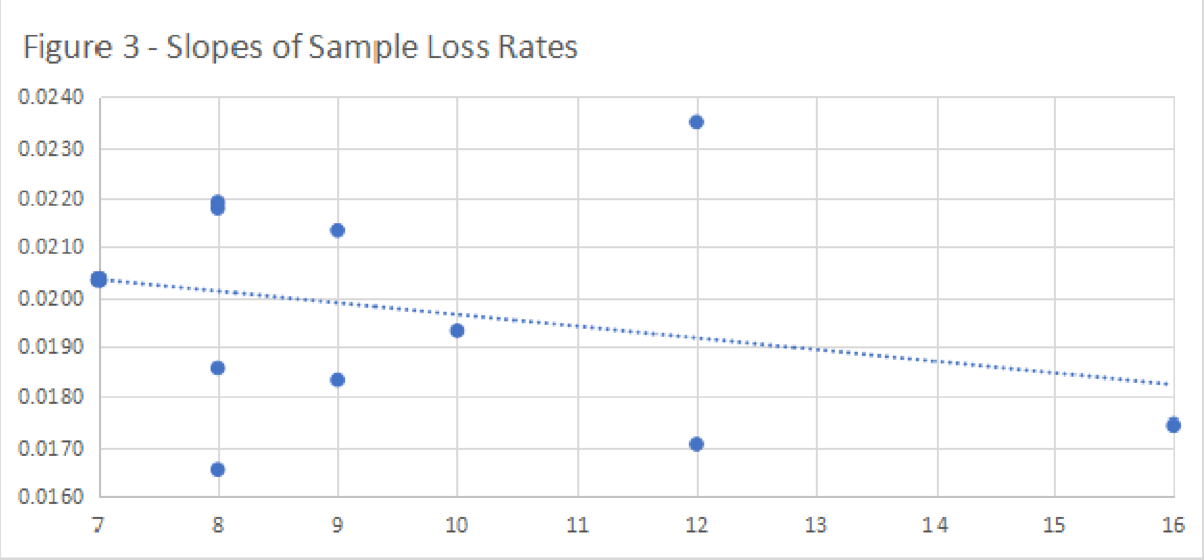
Panel	2008	2004	2001	1996	1993	1992	1991	1990	1989	1988	1987	1986	1985
Intercept	23.18%	19.08%	16.72%	11.06%	8.89%	9.67%	9.68%	7.14%	8.64%	7.67%	7.71%	7.71%	10.39%
Slope	0.0174	0.0171	0.0183	0.0235	0.0213	0.0193	0.0165	0.0218	0.0186	0.0214	0.0204	0.0204	0.0219
Eligible Hhlds	52,031	51,363	40,489	40,188	21,806	21,588	15,626	19,766	3,861	12,517	12,517	12,425	14,306
Waves	16	12	9	12	9	10	8	8	8	6	7	7	8

To better analyze the sample loss rate, linear regression model has been fitted to each panel and the intercepts and slopes are shown in Table 2. The table also states the eligible households of the first wave of each panel and the number of waves again for purpose of comparison.

The intercept of the regression line represents the estimated sample loss rate of each panel when it started. The intercept is a little different than the sample loss rate of the first wave as it is affected by the value of later waves. In the ANOVA test with type II sums of squares, after removing the effect from target sample size, the effect of number of waves on the intercept of sample loss rate has p-value bigger than 0.05. Since it is reasonable that length of panel does not affect the proportion of samples can be accessed, this result shows that there is significant evidence that people will refuse to participate in the survey at the beginning if they know the number of waves they will be interviewed.

Therefore, the rate of sample loss tends to be bigger at the beginning of the panel if the initial size of target sample is larger. While the purpose of increasing the target sample size in redesign in 1996 was to improve the quality of data, this directly increases sample loss rates on average. Since the interviewed households are less in each year comparing with the overlapping panel design and the sample loss rates are higher, the redesigned survey does not work better than initial one in the aspect of representing the population.

Since the sample loss rates are accumulated from previous waves, the value of rate is usually increasing while the number of wave increases. If there is an effect from the number of waves in each panel on the sample loss rate, such as it is more difficult to reach target samples and households are refused to be interviewed when the length of panel increases, the sample loss rate will increase faster with higher number of waves. Therefore, the relationship between the slopes of each regression line of sample loss rate and lengths of panels has been measured. In Figure 3, vertical axis scales the slopes of change in sample loss rates for each panel and the horizontal axis scales the numbers of waves. Because the p-value of the relationship is bigger than 0.05 and the R-squared of the regression model is 0.09, the length of panel have no influence on the sample loss in SIPP surveys. Though the decision of increasing the length of panel since 1996 Panel may increase the costs of tracing, it improves the measurement of changes over time and has no negative influence on the sample loss.



Conclusions

While both the sample size and time dimension of survey affect the sample loss which can result in biases and reducing the quality of survey, the sample loss of SIPP is only affected by the sample size. Comparing with the panels before 1996, the redesign covers a bigger sample size each panel with a simple four-hear panel to improve the quality. However, the quality of information provided is be threatened by the overall increasing sample losses. For example, the 2008 Panel has target sample size of 52,000, but in the sixteenth wave, the sample loss rate has accumulated to 53%, so that more than half of the sample member has not been interviewed. When the initial sample size doubled, the sample loss rate also doubled, which seriously reduce the quantity and quality of information collected. On the contrary, SIPP is successful on maintain and control the sample loss while the length of each panel increases. But with a higher starting point of sample loss, the increasing speed of the accumulation should also be reduced further when time dimension of each panel become longer. Therefore, in the future improvement, it is better for Census Bureau to find out some way to reduce the sample loss of SIPP when they want to increase the sample size.

Reference:

[1] David, M. (1985). Introduction: The Design and Development of SIPP. Journal Of Economic & Social Measurement, 13(3/4), 215-224.

[2] Duncan, G., & Kalton, G. (1987). Issues of Design and Analysis of Surveys across Time. International Statistical Review / Revue Internationale De Statistique, 55(1), 97-117. doi:10.2307/1403273

[3] Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. Econometrica 46: 931–961.

[4] Heather L., Rachel S., & Lynne S. (1999). Strategies for Reducing Nonresponse in a Longitudinal Panel Survey. Journal of Of®cial Statistics, Vol. 15, No. 2, pp. 269±282.

[5] Jason, M. F. (2016). Sample Loss Rates For SIPP 1985 Through SIPP 2008 Panels. Retrieved from https://www2.census.gov/programs-surveys/sipp/tech-documentation

[6] Laurie, Heather; Scott, Lynne. (Jun 1999). Journal of Official Statistics; Stockholm15.2: 269.

[7] O'Muircheartaigh, C. (1996). Measurement errors in panel surveys: implications for survey design and for survey instruments. In: Società Italiana di Statistica, Atti della XXXVIII Riunione Scientifica. Rimini: Maggioli Editore, Vol. 1, pp. 219–230.

[8] Trivellato, U. Quality & Quantity (1999) 33: 339. doi:10.1023/A:1004657006031

[9] U.S. Bureau of the Census. (2016). Survey of Income and Program Participation: SIPP Introduction and History. Retrieved from https://www.census.gov/sipp

Author:

Rao Xiong, statistics and measurement technology professional of University of Pennsylvania, Graduate Students.