Chapter XIII
Cost model for an income and expenditure survey

Hans Pettersson
Statistics Sweden
Stockholm, Sweden

Bounthavy Sisouphanthong
National Statistics Centre
Vientiane, Lao People's Democratic Republic

Abstract

The present chapter describes the work of setting up a cost model for an expenditure and consumption survey in Lao People's Democratic Republic. It begins with a brief discussion of cost models and the problems of estimating the components in the model, and then describes the design of the Lao Expenditure and Consumption Survey 2002. A cost model, which is developed based on budget estimates for the survey, is used for calculations of optimal cluster sizes under different assumptions on rates of homogeneity in the clusters. The chapter concludes with an analysis of the efficiency of the chosen sample design compared with efficiency under optimal conditions.

Key terms: survey design, survey costs, efficiency, cost model, optimum sample size.

A. Introduction

1. The design of a multistage cluster sample involves a number of decisions. One important decision to be made is how to allocate the sample among sample stages in the best possible way. Clustering the sample generally has opposing influences on costs and variances: it reduces the costs and increases the variances. The economic design of a multistage sample requires the sampling statistician to estimate and balance these influences. For this task, he or she needs good information on the variances attributable to the different sampling stages and also information on the variable costs dependent on the sample size at each stage.

2. While variance models have been developed for many common multistage designs, the development of cost models has received less attention among statisticians. Nowadays, variances and design effects are compiled at least for the most important estimates in many surveys in developing countries. The use of cost models to design the sample is less common. Part of the problem is the scarcity of detailed information on survey costs in many national statistical institutes, which makes it difficult to prepare an accurate budget for a survey and to set up a realistic cost model.

3. In the present chapter, we briefly discuss cost models and describe how cost models are used together with variance models to find optimal sample size within primary sampling units (PSUs) in a two-stage design. We develop a cost model for an expenditure and consumption survey in the Lao People’s Democratic Republic and use the model to calculate optimal sample sizes within PSUs.

B. Cost models and cost estimates

Cost models

4. A simple cost model for a two-stage sample may be represented as

$$C = C_0 + C_1 \cdot n + C_2 \cdot n \cdot m \qquad (1)$$

where n = the number of primary sampling units (PSUs) in the sample; m = the number of secondary sampling units (SSUs) (for example, households) in the sample from each PSU; C_0 = the fixed costs of conducting the survey, independent of the number of sample PSUs and SSUs per PSU, including costs for survey planning, costs for development of the survey design, costs for preparatory work, costs for survey management, and costs for data processing, analysis and presentation of results (some of the costs for data processing are dependent on sample size and hence are not fixed costs, but this is disregarded here); = the average costs for adding a PSU to the sample, consisting of costs for travel by interviewers and supervisors between PSUs and home base or between PSUs (fuel costs, driver salaries) and interviewer salaries, including the cost of obtaining maps and other material for the PSU, the cost of establishing the survey in the local area, entailing, for example, meeting with and obtaining permission from local authorities, and the cost of listing and sampling of dwelling units/households within the PSU; = the average cost of including an extra household in the sample, including the costs for locating, contacting and interviewing a household, where the costs consist of interviewer and supervisor salaries and per diem, and also costs for travel by interviewers and supervisors within PSUs.

5. This cost model is simple compared with the more sophisticated cost models that have been developed. Hansen, Hurwitz and Madow (1953) developed a model that isolated the between-PSU travel costs, in which

$$C = C_0 + C_1 \cdot n + C_2 \cdot n \cdot m + C_3 \cdot \sqrt{n} \qquad (2)$$

The cost of adding a PSU (C_1) includes positioning travel cost (travel to the first PSU visited from the interviewer's home base and then back to the home base from the last PSU visited during the data-collection trip) but not the cost of between-PSU travel which is covered by the term $C_3 \cdot \sqrt{n}$. Models isolating both between-PSU travel and positioning travel have also been proposed (Kalsbeek, Mendoza and Budescu, 1983). Groves (1989) provides a relatively broad discussion on cost models, including various complex forms, for example, non-linear, discontinuous, step-function cost expression. However, complexity in the mathematical form of cost models often makes the search for optimality more difficult. Furthermore, lack of accurate data often hampers the use of complex models. In this chapter, the simple model (1) will be used and it is assumed that the second-stage units are households.

Cost estimates

6. The survey manager often has a good idea of the time required for specific survey operations based on information from previous surveys of a similar nature. Experiences from prior surveys (or from pilot surveys) could often be used for reasonable estimates of time per household required for locating and interviewing the household. In these cases, reasonable estimates of C_2 could be compiled. More problematic, usually, is the estimate of C_0, which involves the allocation of indirect costs and the costs for staff that work in several projects/activities. It is often difficult to make estimates for the time required for the administrative, professional and supervisory personnel. Usually, there are no good cost records from previous surveys indicating the costs for that kind of staff. Also, many surveys employ technical assistance (TA) provided by foreign donors. It may be difficult in many cases to separate out the time spent by TA consultants spent on a specific survey.

7. Computing a reasonable estimate of C_1 is often difficult because it involves determining the effect of additional interviewer travel when a PSU is added to the sample. The travel depends on the size of the area being covered, the number of PSUs assigned to each interviewer, and the travel pattern of the interviewers. The travel includes between-PSU travel during a data-collection trip and positioning travel.

8. There is no easy way to overcome the difficulties inherent in making good cost estimates. Accurate and rather detailed cost accounting from previous surveys or a pilot survey is very valuable. In addition to prior experience and pilots, one might also obtain the cost data needed by instituting special cost monitoring capabilities in ongoing surveys, which is done, for example, in the National Health Interview Survey in the United States of America (Kalsbeek, Botman and Massey, 1994).

C. Cost models for efficient sample design

9. Cost modelling can be used for two purposes:

- For budgetary purposes, to set up a survey budget based on the unit costs in the cost model and the planned sample sizes at different stages

- To find an efficient sample design by combining the cost model with a sampling error model

10. In this chapter, our interest is mainly in the use of cost models to find an efficient design. We assume a two-stage design with households selected from PSUs in the second stage. The problem can be stated in this way: given the cost structure represented in the cost model, how should the sample be allocated over the two sampling stages. Separate cost models are usually prepared for urban and rural strata and in some cases for other strata. In that case, the problem also includes the allocation of the sample over urban and rural (and other) strata.

11. We do not have to consider the fixed costs (C_0) when trying to work out an efficient design; the important part is the fieldwork costs: $C_1 \cdot n + C_2 \cdot n \cdot m$. The estimated fieldwork cost per interview (C_f) is found by dividing the total field costs by the number of interviews ($n \cdot m$), giving

$$C_\mathrm{f} = C_2 + C_1/m \tag{3}$$

The variance for the design can be expressed as

$$Var = V \cdot (1 + roh(m-1)) \tag{4}$$

where V is the variance under simple random sampling of households; ρ is the rate of homogeneity (Kish, 1965); see also chap. VI above); and m is the sample size within PSUs.

It is clear from (3) that the fieldwork costs per interview (C_f) could be minimized by making m as large as possible. It is equally clear from (4) that the variance increases with a larger m (and that the variance is minimized by setting m = 1). The optimum number of households, m_{opt}, is the value of m that minimizes $Var \cdot C_f$ where

$$Var \cdot C_f = V \cdot (1 + roh(m-1)) \cdot (C_2 + C_1 / m) \tag{5}$$

It has been shown (Kish, 1965) that the optimal sample size can be found by

$$m_{opt} = \sqrt{\frac{C_1}{C_2} \cdot \frac{(1-\rho)}{\rho}} \tag{6}$$

12. The first factor in equation (6), C_1/C_2, is the cost ratio between the unit costs in the first and second stages. The cost of including a new PSU in the sample (C_1) will always be higher than the cost of including a new household in a selected PSU (C_2), hence the cost ratio will always be well above 1.0. The higher the cost ratio, the more costly it is to select a new PSU compared with selecting more households in selected PSUs; consequently, we should select more households in already selected PSUs.

13. The quantity ρ measures the internal homogeneity of the PSU. When the internal homogeneity is high, it is not desirable to take a large sample of households in the PSU inasmuch as the information gain from each new household in the sample will be small (because the households are very similar). This is reflected in the second factor in (6). When ρ is high, this factor, and m_{opt}, become small (for a given cost ratio).

14. The ρ values are often derived from design effects estimated from previous surveys. The ρ 's tend to be small -- often less than 0.01 -- for many demographic variables. For many socio-economic variables, the ρ 's may be above 0.1, and in some cases, as high as 0.2 or 0.3.

15. The cost ratio has also to be worked out from experiences in previous surveys. It should be pointed out that it is not necessary to express the ratio in terms of costs. Time (in terms of required interviewer days) is often used as the unit instead of costs: the mathematics will be approximately the same (some travel costs may be overlooked). The level of the cost ratio depends on the fieldwork design. For a survey where the time spent on the interview is very short, the cost ratio may be 20-50. If, for example, the time required per PSU independently of the household interviewing is three days and the interviewer is able to cover 10 households per day the cost ratio (calculated as the time ratio T_1/T_2) will be 30 (T_1=3 days and T_2=0.1 days). In surveys with very long interviews, the cost ratio may be below 10.

16. The mathematics employed in the calculations may give the impression that a precise and clear-cut answer can be obtained to the question how many households to select from each PSU. That is almost never the case, however, owing to several factors, namely:

- The cost model is a rather crude approximation of the reality. Simplification is needed to make the cost model manageable (as discussed in sect. B).

- The estimates of costs and ρ 's are subject to uncertainty.

- The optimum applies to one survey variable out of many. If the important survey variables in the survey have different levels of ρ, then there will be no single optimal cluster size but rather a number of different ones.

17. The calculations will provide rather crude indications of what the optimum sample size is for different values of ρ. This information can be used to decide on a sample size within PSUs that suits all the important survey variables reasonably well. In respect of the final decision, there may also be other factors to consider, often related to practical constraints on the fieldwork.

D. Case study: the Lao Expenditure and Consumption Survey 2002

18. The National Statistics Centre (NSC) of the Lao People's Democratic Republic has conducted two expenditure and consumption surveys in the last decade. The first Lao Expenditure and Consumption Survey (LECS-1) was conducted in 1992-1993; the second (LECS-2) in 1997-1998; and the third (LECS-3) in 2002-2003. The present section describes LECS-3.

19. Data from the surveys are used for a number of purposes, the most important being to produce national estimates of household consumption and production for the national accounts. This includes estimating production in household agricultural activities and business activities.

Sample design for LECS-3

20. The sample consisted of 8,100 households selected through a two-stage sample design. Villages served as primary sampling units (PSU). The villages were stratified on 18 provinces and within provinces on urban/rural sector. The rural villages were further stratified on villages "with access to road" and "with no access to road". The total first-stage sample consisted of 540 villages. The sample was allocated to provinces proportionally to the square root of the population size according to population census. The PSUs were selected with a systematic probability proportional to size (PPS) procedure in each province.

21. The households in the selected villages were listed prior to the survey. Fifteen households were selected with systematic sampling in each village, giving a sample of 8,100 households. The decision to select 15 households per village was primarily based on practical considerations. In section E, we compare the efficiency of the 15 household samples with optimum sample sizes under different assumptions on rates of homogeneity.

Data collection in LECS-3

22. Data were collected by the means of (a) a household questionnaire; (b) a village questionnaire; and (c) a price collection form. The last two questionnaires mainly served as instruments with which to collect supplemental information for the household survey.

23. A large part of the household questionnaire remained the same as in previous surveys, except for some modifications in questions that had not worked well in the previous survey. Data on expenditure and consumption were collected for a whole month based on daily recording of all transactions. At the end of the month, the household was asked about purchases of durable goods during the preceding 12 months. During the month, each member of the household should

have recorded the time use during a 24-hour period. The rice consumption of each member of the household was measured for one “yesterday” to get a more precise measure of intake at each meal for each person.

24. The village questionnaire, which was administered to the head of the village, covered such items as roads and transport, water, electricity, health facilities, local markets, schools, etc. The price collection form was used by the interviewers to collect data on local prices of 121 commodities.

Fieldwork

25. The measurement of daily consumption through a diary kept by the household put a heavy burden not only on the households but also on the field interviewers. Many households, especially in the rural areas, needed frequent support in the task of keeping the diary. In order to secure an acceptable quality in the data, it had been deemed necessary to keep the interviewers in the village for the whole month rather than have them travel to the villages for repeated interviews and follow-up. This decision was also supported by the fact that many villages, especially in the mountainous areas, were difficult to access (access to some villages required travel by foot for several days).

26. In the previous surveys, teams of two interviewers in each village had carried out the fieldwork. For LECS-3, a single-interviewer design was considered. However, in the final analysis, factors related to interviewers security and well-being weighed in favour of having two interviewers in the village. The interviewers made several visits to the selected households during the four-week period. The interviewers also worked with the village leaders to complete the village questionnaire and to update the village registers. During the month, the interviewers also collected data on prices at the local market.

27. The field staff consisted of 180 interviewers organized in 90 two-member teams. Thirty-six supervisors from the provincial statistical offices and 10 central supervisors from the head office supervised the teams.

E. Cost model for the fieldwork in the 2002 Lao Expenditure and Consumption Survey (LECS-3)

Cost estimates

28. LECS-3 was, to a large extent, similar to the two previous LECS surveys. Experiences in respect of the time required for the fieldwork in the two previous surveys were therefore used for estimating the fieldwork costs in LECS-3.

29. Table XIII.1 contains estimates of required time for fieldwork in the villages for LECS-3. Separate estimates have been made for urban and rural areas.

Table XIII.1. Estimated time for fieldwork in a village

	Field travel	Introducing survey, listing and selecting households in villages, collecting village information	Household interview work
	No of days/ village	*No of days/ village*	*No of days/ village*
Urban (100 villages)			
Province supervisors	1.5	0.5	3
Interviewers (teams of 2)	3	7	47
Rural (440 villages)			
Province supervisors	3	0.5	3
Interviewers (teams of 2)	6	7	47

30. Table XIII.2 contains estimated costs for the fieldwork calculated on the basis of the time estimates in table XIII.1. The costs include travel costs (usually by car or bus) and field allowances (per diem) for the working time in the field. The staff working with the survey was without exception permanent staff of the NSC assigned to the survey as part of their ordinary duties. The cost items therefore do not include ordinary salaries.

Table XIII.2. Estimated costs for LECS-3 (US dollars per diem)

	Field travel costs (per diem for travel time and estimated travel costs)	Introducing survey, listing and selecting households in villages, collecting village information	Household interview work
	A	***B***	***C***
Urban (100 villages)			
Province supervisors	1 540	450	2 710
Interviewers (teams of 2)	2 490	5 060	33 970
Rural (440 villages)			
Province supervisors	15 850	1 990	11 950
Interviewers (teams of 2)	25 560	22 260	149 460
Total	45 440	29 760	198 090

Cost model

31. Columns ***A*** and ***B*** in the table XIII.2 present costs related to the selection and preparation of the villages for the survey. The sum of the items in these columns divided by the number of villages constitutes the average cost (C_1) in United States dollars of including a village in the

survey: for urban areas: C_1 = (1,540+2,490+450+5060)/100 = 95; and for rural areas: C_1 = (15,850+25,560+1,990+22,260)/440 = 149. All travel is considered as between-village travel; all the travel costs are therefore included in C_1.

32. Column ***C*** in table XIII.2 presents survey costs related to the interviews of the households. The main item is interviewer time. The sum of the items in this column divided by the number of households constitutes the average cost (C_2), in United States dollars, of including a household in the survey: for urban areas: C_2 = (2,710+33,970)/(100.15) = 24; and for rural areas: C_2 = (11,950+149,460)/(440.15) = 24. When inserting the estimated values for C_1 and C_2, the cost function becomes

$$\text{Urban: } C_{fieldwork} = 95 \cdot n + 24 \cdot n \cdot m \qquad (7)$$

$$\text{Rural: } C_{fieldwork} = 149 \cdot n + 24 \cdot n \cdot m \qquad (8)$$

33. The fact that the personnel costs did not include permanent staff salaries results in an underestimate of C_1 and C_2, and consequently an underestimate of $C_{fieldwork}$. Most important for the optimization of the design, however, is the cost ratio C_1/C_2. We could expect the cost ratio to be only slightly affected by the omission of salaries, as the omission will have rather similar effects on C_1 and C_2.

34. The cost ratio between the first- and second-stage samples is C_1/C_2 = 95/24 = 3.9 for urban areas and 149/24 = 6.1 for rural areas. These cost ratios are rather low, reflecting the fact that the survey required considerable time for interview and follow-up per household over the month when the interviewer-supported diary method was used. LECS-3 was an unusual survey in that respect.

Optimum sample size within villages

35. In the previous LECSs, the two interviewers had had a workload of 20 households in each village. For LECS-3, the sample size was reduced to 15 households. The reduction in workload from 20 to 15 households stemmed from the fact that the household interviews were considerably longer in LECS-3 as compared with the previous surveys. Also, LECS-3 contained a price questionnaire that had not been included in the previous surveys.

36. How efficient was the design with two interviewers in the village covering a sample of 15 households? The cost model, along with a variance model, could be used for an assessment of the relative efficiency of the 15 household samples.

37. In table XIII.3, the optimal value of m is presented for different values of ρ. The relative efficiency of our design is shown in rows three and four. It is computed as the ratio between the minimum of $Var.C_f$ (see (5)) and the actual value of $Var.C_f$ for a given ρ and a sample size of 15. The efficiency is reasonably high for ρ values up to 0.10; it is rather low and tends to deteriorate for ρ values equal to 0.2 and above.

Table XIII.3. Optimal sample sizes in villages (m_{opt}) and relative efficiency of the actual design (m=15) for different values of ρ

	ρ =0.01	ρ =0.05	ρ =0.10	ρ =0.15	ρ =0.2	ρ =0.25
m_{opt}, urban	20	9	6	5	4	4
m_{opt}, rural	24	11	8	6	5	4
Relative efficiency (percentage) urban	99	94	82	73	66	61
Relative efficiency (percentage, rural	96	98	89	81	75	70

38. Calculations of ρ in the previous LECS had shown that there were clear urban/rural differentials in ρ for important LECS variables. The ρ´s in urban areas are considerably lower than the ρ´s in the rural areas. We could expect ρ to be in the range of 0.04-0.08 for many urban estimates in LECS, in which case a sample of eight to nine households would be optimal. Our design with a sample of 15 households per PSU will have a relative efficiency of 85-95 per cent. The ρ´s in rural areas are in the range 0.11-0.20, in which case a sample of five to seven households would be optimal. Our sample will have a relative efficiency of 75-88 per cent. There is some uncertainty, especially concerning the ρ´s we can expect in respect of important variables in LECS-3. Still, we can safely conclude that our sample of 15 households is above the optimum.

39. What are the practical implications of these results for the future LECS surveys? The efficiency losses are small in the urban areas; we may therefore decide to stay with the 15 households alternative. We would like to reduce the sample per PSU in rural areas. However, the present fieldwork set-up where the interviewers have to stay in the PSU for a full month makes it difficult to reduce the workload considerably. This means that the interviewers will not be fully occupied during the month. It may be possible to give the interviewers other tasks with which to fill the working time, for example, conducting community surveys in the area during the month. Whether that is a viable option has to be discussed.

F. Concluding remarks

40. A cost model for the fieldwork in LECS-3 has been developed and analysed. It shows that the cost ratio, C_1/C_2, for the survey was rather low. The main reason is the time-consuming interviewer-supported diary method that was used for LECS-3 where the interviewers stayed in the village for a whole month and gave the households all the assistance needed for the diary-keeping. In that respect, LECS-3 was a rather unusual survey compared with other household income and expenditure surveys where the interview time per household was usually lower.

41. Calculations of optimum sample sizes within PSUs show that the present sample size of 15 households is above the optimum, especially in rural areas. However, practical constraints may make it difficult to reduce the sample size.

42. It should be pointed out that the cost model is only a crude approximation of the reality; whole complexity cannot be completely captured by any simple model. More complex models could be built including, for example, various step-function cost expressions. However, complexity in the mathematical form of cost models will often make it more difficult to determine optimality.

References

Groves, R. M. (1989). *Survey Error and Survey Costs.* New York: John Wiley and Sons.

Hansen, M.H., W.N. Hurwitz and W.G. Madow (1953). *Sample Survey Methods and Theory,* vol. I. New York: John Wiley and Sons.

Kalsbeek, W., O.M. Mendoza and D.V. Budescu (1983). Cost models for optimum allocation in multi-stage sampling. *Survey Methodology*, vol. 9, No. 2, pp. 154-177.

Kalsbeek W.D., S.L. Botman and J.T. Massey (1994). Cost efficiency and the number of allowable call attempts in the National Health Interview Survey. *Journal of Official Statistics,* vol. 10, No. 2, pp. 133-153.

Kish, L. (1965). *Survey Sampling.* New York: John Wiley and Sons.