Splitting Long Questionnaires: Some Theory and an Example

Dhiren Ghosh
Synectics for Management Decisions
1901 North Moore Street, Suite 900
Arlington, VA 22209, USA

Andrew Vogt
Department of Mathematics
Georgetown University
Washington, DC 20057-1233, USA
vogt@math.georgetown.edu

Splitting a long questionnaire into parts that are administered to subsamples is sometimes done to decrease respondent burden. However, it is desirable to choose the parts in such a way that the loss of precision is small. Our purpose here is to discuss a theoretical framework for implementing such a split that makes use of double sampling and regression estimation. We shall also describe briefly an example, namely the U. S. Consumer Expenditure Survey conducted by the Bureau of Labor Statistics.

Suppose the questionnaire consists of three sets of questions corresponding to three sets of numerical random variables, call the sets X, Y, and Z. The questions associated with the set $X \cup Y \cup Z$ are administered to n_1 units, the questions associated with the set $X \cup Z$ to n_2 other units, and the questions associated with the set $Y \cup Z$ to n_3 other units. The Z questions include basic questions related to the sampling frame as well as others that are good predictors. The X and Y questions are chosen in such a way that each variable in Y can be regressed on the variables in X and Z, and similarly each variable in X can be regressed on the variables in Y and Z.

Consider one of the Y variables, call it y. It is possible to develop an estimate of the population mean $\overline{Y}$ of y by double sampling. A regression estimate of $\overline{Y}$, denoted by $\overline{y}_{ds}$, is obtained by using the X and Z variables as auxiliary variables. In effect the mean of y on the $X \cup Y \cup Z$ subsample is enhanced by the values of the X and Z variable on the $X \cup Z$ subsample.

Separately the regular mean $\overline{y}$ is obtained from the $Y \cup Z$ subsample. Then a weighted combination $\overline{y}_c$ of $\overline{y}_{ds}$ and $\overline{y}$ is used as the final estimator. (Other choices are possible.) In fact,

$$\overline{y}_c = \frac{V}{V_{ds} + V}\overline{y}_{ds} + \frac{V_{ds}}{V_{ds} + V}\overline{y}$$

where V_{ds} and V are the respective sampling variances of the two estimators. The combined estimator $\overline{y}_c$ is the well-known optimal combination of two independent estimates. Its sampling variance is given by:

$$V_c = \frac{1}{\frac{1}{V_{ds}} + \frac{1}{V}}.$$

The sampling variances V and V_{ds} are given by:

$$V = \frac{S_y^2}{n_2} \text{ and } V_{ds} = \frac{S_y^2}{n_1}(1 - R^2)(1 + \frac{n_2}{n_1 + n_2}\frac{p}{n_1 - p - 2}) + \frac{R^2 S_y^2}{n_1 + n_2}$$

where S_y^2 is the population varaince of y and R is the multiple correlation coefficient of y with the other variates. The variance for the double sampling estimator is contained in Khan and Tripathi (1967) and was improved upon by Cochran (1977), p. 340. A term in Cochran's version with total population in the denominator is omitted. Both Khan and Tripathi and Cochran assume multivariate normality. Cochran, in addition, explicitly invokes a superpopulation model with a notion of model-unbiasedness and with variance regarded as an average sampling variance. Elsewhere we derive these estimates.

A questionnaire to which this theory has been applied is the Consumer Expenditure Survey. This survey has sections that consist of many questions, but each section can be represented by a summary variable (e.g., section 2 concerns rented living quarters and the summary variable is the sum of all rental payments made in the reference period, adjusted for business and rooms rented to others, and summed over all members of the consumer unit.) The split keeps sections intact but separates the summary variables (and hence the sections) into three different groups X, Y, and Z. Within X there should be some topical continuity to aid interviewees, but some independence is also desirable. The multiple correlation coefficient of a Y variable with the X and Z variables should be high. Likewise the same should be true when X and Y are interchanged. 1997 Interview Survey Public Use Microdata were used to propose a split into ten Z variables (sections), five X variables, and five Y variables. If the combined sample size is 1,800 (a typical monthly value) with $n_1 = n_2 = n_3 = 600$, the average standard error of a Y or X variable rises by a factor of 1.16 to 1.21.

REFERENCES

Cochran, W. G. (1977). *Sampling Techniques. Third Edition. John Wiley & Sons.*
Khan, S. and Tripathi, T. P. (1967) The Use of Multivariate Auxiliary Information in Double Sampling. *J. Ind. Stat. Assoc.*, **5**, 42-48.

RÉSUMÉ

Un questionnaire est divisé: un sous-échantillon obtient toutes les questions, deux autres obtiennent des sous-ensembles. La régression et le double prélèvement sont utilisés pour estimer des paramètres de la population. Le *Consumer Expenditure Survey* des États Unis démontre la méthode.