OPTIMALITY IN BALANCED ROTATING PANEL SURVEY DESIGN

Dhiren Ghosh, Synectics for Mngmt Decisions & Andrew Vogt, Georgetown U.
Dhiren Ghosh, SMD, 3030 Clarendon Blvd, # 305, Arlington, VA 22201

Keywords: double sampling, regression estimates, response burden

In panel surveys, consideration of response burden dictates that sample members be rotated. Rotation of subsamples over time can also serve to improve the current estimate and the estimate of change from one time period to the next. This is done by using past data as auxiliary variables.

Consider the following rotational design:

$$\begin{bmatrix} A & A & E & F & G \\ B & E & B & H & I \\ C & F & H & C & J \\ D & G & I & J & D \end{bmatrix}.$$

In this scheme every subsample is used exactly twice in a cycle of five time periods, and there is exactly one subsample common to any two times. For any epoch we fit a regression on the subsample common to that epoch and another epoch and use the other epoch's data to obtain an improved double sample estimate. The improved estimates for each subsample in a given epoch are combined to yield the final estimate for the epoch. These regression estimates can run backward and forward in time, i.e., data from subsequent years can be used to revise and improve earlier estimates.

Another such design is:

$$\begin{bmatrix} A & A & A & B & B \\ D & C & B & C & D \\ E & D & C & E & E \end{bmatrix}.$$

Here every subsample is used three times. The number of subsamples common to adjacent epochs is two, while the number common to epochs separated by one or more epochs is one. The first and last epoch also have two in common, an example of circular symmetry that permits the cycle to be repeated if desired. Neighboring epochs are typically more closely correlated and the change in the estimate from one epoch to the next is often of interest. Hence a scheme that offers greater overlap between adjacent epochs may be preferred.

Yet another design is:

$$\begin{bmatrix} A & A & B & B & C & C \\ C & G & A & G & B & G \\ D & D & E & E & F & F \\ F & H & D & H & E & H \end{bmatrix}.$$

Adjacent epochs, and the first and last, have two subsamples in common, as do epochs separated by an epoch. But notice that the first and fourth epochs, respectively second and fifth, or third and sixth, have nothing in common. This makes it possible to generate a completely independent estimate every three epochs.

The designs above are examples drawn from a general theory of combinatorial designs that has been developed in recent years (see [1], [2], [3], [4]). Work in experimental design by Fisher at Rothamsted Experimental Station and by others at Iowa State and the Indian Statistical Institute formed the background for the present mathematical theory of designs. The latter is a blend of combinatorics, projective and affine geometry, linear algebra, graph theory, Galois theory, and group theory; and has applications in coding theory as well as numerous ones in statistics.

Suppose a set consists of d objects (subsamples) and these are arranged into overlapping sets or blocks of size m, there being b blocks in total (all of them distinct) and each object being a member of exactly r of the blocks. In our original vocabulary we have b

time epochs, m subsamples per epoch, a total of d distinct subsamples, with each subsample occurring in r different time epochs. Assume for simplicity that the subsamples are all of the same size, and thus this size can be ignored. An equitable distribution of response burden leads us to impose the condition that every subsample occurs r times.

It is easily seen that $mb = rd$, and that $r < b$ and $m < d$ with all numbers being positive integers. In design theory one talks about t-designs, in which any t objects lie in a fixed number λ of blocks. Thus a 2-design with $\lambda = 3$ is one in which every pair of objects has three blocks in common. The intersection numbers of a design are the cardinalities of the intersections of distinct blocks. Thus, the first example above has intersection number 1, the second has intersection numbers 1 and 2, and the third has intersection numbers 0 and 2.

As a general rule designs are more desirable, and more practical, in which adjacent blocks - our blocks are time-ordered - have relatively large intersection numbers, while distant blocks have low intersection numbers.

Consider the following extreme design:

$$\begin{bmatrix} x_1 & x_2 & \dots & x_{d-1} & x_d \\ x_2 & x_3 & \dots & x_d & x_1 \end{bmatrix}.$$

Each subsample occurs twice, and in adjacent epochs there is a common subsample. There is also circular symmetry. But there is no other overlap. The intersection number is 1 for adjacent blocks, 0 for all others.

Another extreme design is:

$$\begin{bmatrix} x_1 & x_1 & x_2 & x_3 & \dots & x_m \\ x_2 & x_{m+1} & x_{m+1} & x_{m+2} & \dots & x_{2m-1} \\ x_3 & x_{m+2} & x_{2m} & x_{2m} & \dots & x_{3m-3} \\ \dots & & & & & \\ \dots & & & & & \\ x_m & x_{2m-1} & x_{3m-3} & x_{4m-6} & \dots & x_d \end{bmatrix}.$$

The only intersection number here is 1, and any two blocks have a subsample in common, or a fraction $\frac{1}{m}$ of the sample at each epoch. The total number of subsamples is $d = {}_{m+1}C_2$. This type of design is of less interest to us for panels.

Any design offers opportunities for optimization depending on one's objectives. The formula for a double sampling regression estimate (see [5, p. 346]), with finite population correction factor omitted, is:

$$\sigma^2 \left(\frac{1-\rho^2}{kn} + \frac{\rho^2}{mn} \right)$$

where σ is the standard deviation of the variable to be estimated (say, for the latest epoch), n is the size of each subsample, m the number of subsamples per epoch, k the number of subsamples in common between the latest epoch and a second epoch, and ρ the correlation coefficient between the variables of the two epochs. When the double sampling estimate is combined with the mean for the current epoch on the $mn - kn$ unmatched elements, with optimal weights and with k chosen optimally, the variance of the resulting estimate is:

$$\frac{\sigma^2}{2mn} \left(1 + \sqrt{1-\rho^2} \right)$$

Without detailed knowledge of ρ we can still recommend designs that offer improved estimates for the different epochs. Cochran [5] shows that if the current estimate is to be optimized on the basis of overlap with one previous epoch, then the preferred ratio of k to m never exceeds 50%. On the other hand, if the changes in an estimate from one epoch to the next are of interest, then the preferred ratio exceeds 50%. Economy also dictates high overlap between adjacent epochs, as Cochran [5] further observes. Our examples above suggest that overlaps as desired can be achieved.

In general, a design of the following form has suitable properties:

$$\begin{bmatrix} x_1 & x_{m-r+1} & \dots\dots & x_{(d-1)(m-r)+1} \\ x_2 & x_{m-r+2} & \dots\dots & x_{(d-1)(m-r)+2} \\ x_3 & x_{m-r+3} & \dots\dots & x_{(d-1)(m-r)+3} \\ \dots & & & \\ \dots & & & \\ x_m & x_{2m-r} & \dots\dots & x_{dm-(d-1)r} \end{bmatrix}.$$

Here an element x_i represents a subsample of size n, but we treat each x_i as a unit for simplicity. The intersection number between adjacent blocks in this design is r provided $r < m$ and $m + (m - r) \leq d$. It can be shown that if $gcd(m-r, d) = 1$, then each element occurs exactly r times and the design is equitable. The overlapping fraction between adjacent epochs is $\frac{r}{m}$ and the frequency of appearance of a given subsample is $\frac{r}{d}$, i. e., r times in every d epochs. Given a desired overlap ratio, we choose r and m accordingly. The remaining constraints affect d, but d can still be chosen large to minimize the response burden.

BIBLIOGRAPHY

[1] D. R. Hughes and F. C. Piper, Design theory, Cambridge Univ. Press, Cambridge, 1985.

[2] E. S. Lander, Symmetric Designs: An Algebraic Approach, Cambridge University Press, Cambridge, 1983.

[3] M. S. Shrikhande and S. S. Sane, Quasi-symmetric Designs, Cambridge University Press, Cambridge, 1991.

[4] T. Beth, D. Jungnickel, and H. Lenz, Design Theory, Cambridge University Press, Cambridge, 1986.

[5] W. G. Cochran, Sampling Techniques, John Wiley and Sons, New York, Third Edition, 1977.