Covariance Estimates in Stratified and Multistage Clustered Sampling

Dhiren Ghosh
Synectics for Management Decisions, Inc.
1901 North Moore Street, Suite 900
Arlington, VA, 22099

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

Abstract

Important parameters, such as regression and correlation coefficients and partial regression and correlation coefficients, are mathematical functions of covariances. Estimates of these are derived in the case of a sample survey of a stratified clustered population where a sample of clusters is chosen from each stratum by simple random sampling, and a simple random sample of elements is chosen from these clusters. Each estimate of covariance has an intracluster component and an intercluster component and is unbiased. The associated mathematical functions involve operations of multiplication, division, and square roots, and their estimators are in general biased but consistent. Parameter estimates for alternative sampling designs can be dealt with in a similar way.

keywords: covariance estimates from clusters, partial regression coefficients, intracluster correlation

1 Introduction

Survey statisticians use many different techniques [1] - stratification, clustering, multistage designs, with various ways of determining the probability of selecting a unit - to obtain efficient estimates of parameters. Final estimates of some population parameters sometimes get lost in the welter of intermediate steps. In this paper we review the logic that leads to estimates of population parameters. We consider the case of stratified cluster sampling where the population is divided into a set of clusters, the clusters are divided into strata, a set of clusters is selected from each stratum with equal probability, and within the selected clusters samples are drawn with equal probability. The population parameters we consider are population regression coefficients, simple and partial, and population correlation coefficients - simple, partial, and multiple.

2 Definitions

Imagine a population of size N in which we are studying the variables y, x_1, x_2, ..., x_p, and for which we are interested in how y depends on $x_1, ..., x_p$. To begin with, consider the case of y versus a single variable x. The *(simple) regression coefficient* of y on x is defined to be the unique number $\beta(y,x) = \beta$ that minimizes $\Sigma(y - \mu_y - \beta(x - \mu_x))^2$ where the sum is taken over all observations in the population. Here μ_y and μ_x are the population means of y and x. It is a familiar fact that:

$$\beta(y,x) = \frac{Cov(y,x)}{Var(x)},$$

where Cov and Var refer to the population covariance and variance. Another parameter of importance is the *(simple) correlation coefficient* between y and

x, defined by:

$$\rho(y,x) = \frac{Cov(y,x)}{\sigma(y)\cdot\sigma(x)},$$

where σ is the population standard deviation. Since $\sigma(x) = \sqrt{Var(x)}$ and $Var(x) = Cov(x,x)$, it is apparent that both the regression coefficient and the correlation coefficient are simple mathematical functions of covariances.

Returning to y versus x_1, ..., x_p, we define *partial regression coefficients* of y on x_j to be the unique numbers β_1, β_2, ..., β_p that minimize $\Sigma(y - \mu_y - \Sigma_{j=1}^{p}\beta_j(x_j - \mu_{x_j}))^2$ where the outer sum is over the entire population. The column vector β of partial regression coefficient is given by:

$$\beta = (\mathbf{\Sigma})^{-\mathbf{1}}\mathbf{Y}$$

where $\mathbf{Y}$ is the p-dimensional column vector whose entries are $Cov(y,x_j)$ for j = 1, ..., p, and $\mathbf{\Sigma}$ is the p by p covariance matrix of x_1, ..., x_p, whose (j,k)-th entry is $Cov(x_j,x_k)$. The individual regression coefficients are the entries in the column vector β. Since $\mathbf{\Sigma}^{-1} = \frac{1}{det\mathbf{\Sigma}}adj\mathbf{\Sigma}$, it is evident that each partial regression coefficient is a rational function of the covariances $Cov(x_j,x_k)$ and $Cov(y,x_j)$. The *partial correlation coefficients* between y and x_1, ..., x_p are the numbers ρ_1, ..., ρ_p defined by:

$$\rho_j = \rho(y_j, \hat{x}_j)$$

where $y_j = y - \mu_y - \Sigma_{k\neq j}\beta_k^j(x_k - \mu_{x_k})$ and $\hat{x}_j = x_j - \mu_{x_j} - \Sigma_{k\neq j}\gamma_k^j(x_k - \mu_{x_k})$. Here for each fixed j, the coefficients $\{\beta_k^j\}$ and $\{\gamma_k^j\}$ are chosen to minimize the sums Σy_j^2 and $\Sigma\hat{x}_j^2$, summation being over the entire population. It can be shown that:

$$\rho_j = \frac{Cov(y,x_j) - \mathbf{X}_j^t\mathbf{\Sigma}_j^{-1}\mathbf{Y}_j}{\sqrt{(Var(x_j) - \mathbf{X}_j^t\mathbf{\Sigma}_j^{-1}\mathbf{X}_j)(Var(y) - \mathbf{Y}_j^t\mathbf{\Sigma}_j^{-1}\mathbf{Y}_j)}}.$$

Here $\mathbf{X}_j$ is the the (p - 1)-dimensional column vector with entries $Cov(x_j,x_k)$ for $k \neq j$, $\mathbf{Y}_j$ is the column vector $\mathbf{Y}$ with its j-th entry omitted, and $\mathbf{\Sigma}_j$ is the matrix obtained from $\mathbf{\Sigma}$ by omitting the j-th row and j-th column.

Finally we introduce the *multiple correlation coefficient* of y on x_1,...,x_p defined by:

$$R = \rho(y,\hat{y}),$$

where $\hat{y} = \mu_y + \Sigma_{j=1}^{p}\beta_j x_j$. It can be shown (see [2], p. 267) that

$$R = \sqrt{\frac{\mathbf{Y^t\Sigma Y}}{Var(y)}}.$$

3 Methods

The parameters of interest, i.e., the coefficients β, ρ, β_j, ρ_j, and R, are obtained, as we have seen, by applying simple mathematical functions to covariances. Estimates for these same parameters can be obtained by applying the same mathematical functions to estimates of the pairwise covariances (and variances) among the variables y, x_1, ..., x_p.

In the case of simple random sampling, unbiased estimates of covariances are readily available. For example, an unbiased estimate of $Cov(y,x)$ is $\frac{N-1}{N}s_{yx}$ where:

$$s_{yx} = \frac{\Sigma_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n-1},$$

x_1, ..., x_n, and y_i,...,y_n are the values of x and y on a random sample of size n, and $\overline{x}$ and $\overline{y}$ are their sample means.

Consider now the case when the population is divided up into clusters, and the survey statistician uses a sampling design that involves selecting clusters and taking simple random samples from the selected clusters. Aggregating the samples into a single sample and computing s_{yx} as above is not appropriate for developing unbiased estimates. Instead we take into account how the clusters that are sampled are selected.

Suppose we wished to estimate the grand mean of a variable x, namely $\mu = \Sigma_{i=1}^{k}N_i\mu_i$ where there are k clusters and the i-th cluster has size N_i and mean μ_i. Then an unbiased estimate for μ, based on a random sample of r clusters, is:

$$\overline{x}_{cl} = \frac{\frac{1}{r}\Sigma_{i=1}^{r}N_i\overline{x}_i}{\frac{1}{k}N}. \qquad (1)$$

Here $\overline{x}_1$, ..., $\overline{x}_r$ are sample means from simple random samples drawn from clusters 1, ..., r, the numbering being chosen so that the random set of clusters actually selected are numbered 1,...,r out of 1,..., k, and $N = \Sigma_{i=1}^k N_i$.

The sample means $\overline{x}_i$ are unbiased estimates of the cluster means μ_i, and the numerator in the above expression is an unbiased estimate for the average total of the x-values in a cluster. Dividing by N/k, the average cluster size, one arrives at an unbiased estimate for μ. This unbiased estimate is sometimes bypassed in favor of a ratio estimate, in which (1) is replaced by $(\Sigma_{i=1}^r N_i\overline{x}_i/\Sigma_{i=1}^r N_i)$. The ratio estimate is biased but if the cluster means μ_i and cluster sizes N_i are correlated as i varies, then this estimate has smaller mean square error than $\overline{x}_{cl}$. A choice between an unbiased estimate and a biased ratio estimate must also be made for covariances, namely the covariance of x_i and x_j and of y and x_i. However, these may be related to the cluster sizes in diverse ways. We work with unbiased estimates only.

Now consider the estimation of the covariance of two variables y and x. In terms of the individual clusters, the covariance can be expressed as:

$$Cov(y,x) = \Sigma_{i=1}^k \frac{N_i}{N} Cov_i(y,x) + \Sigma_{i=1}^k \frac{N_i}{N}(\nu_i - \nu)(\mu_i - \mu)$$

where $Cov_i(y,x)$ is the covariance of y and x within the i-th cluster only, and ν_1, ...,ν_k, ν are the cluster means and population mean of y. The first summation on the right measures intra-cluster covariance, while the second measures inter-cluster covariance.

If we take a census of r clusters randomly selected from the k clusters, then an unbiased estimate of the covariance of y and x is:

$$\frac{k}{r}\Sigma_{i=1}^r \frac{N_i}{N} Cov_i(y,x) + \frac{k}{r}\Sigma_{i=1}^r (\frac{N_i}{N} - (\frac{N_i}{N})^2)\nu_i\mu_i - \frac{k(k-1)}{r(r-1)}\Sigma_{i\neq j;i,j=1}^r \frac{N_i}{N}\nu_i\frac{N_j}{N}\mu_j.$$

Instead of a census of the selected clusters, if we select independent samples from each of these clusters and calculate sample means and sample covariances in each sample, we obtain (after a computation) the following unbiased estimate of the overall covariance $Cov(y,x)$:

$$\frac{k}{r}\Sigma_{i=1}^r \frac{N_i}{N}(1 - \frac{1}{n_i} - \frac{1}{N} + \frac{N_i}{n_i N^2})s_{i,yx} + \frac{k-1}{k(r-1)}\Sigma_{i=1}^r (\frac{kN_i\overline{y}_i}{N} - \overline{y}_{cl})(\frac{kN_i\overline{x}_i}{N} - \overline{x}_{cl}). \quad (2)$$

In the above equation $\overline{y}_{cl}$ and $\overline{x}_{cl}$ are the (unbiased) estimates of the grand means ν and μ of y and x as in (1), and $s_{i,yx}$ is the sample covariance of y and x in the i-th cluster, namely,

$$s_{i,yx} = \frac{1}{n_i - 1}\Sigma_{\text{sample from cluster }i}(y - \overline{y}_i)(x - \overline{x}_i),$$

and n_i is the size of the sample drawn in this cluster.

To summarize, we have described, in the case of an unstratified but clustered population, how to obtain unbiased estimates of the grand means ν and μ of variables y and x, and the covariance of y and x. These are the outputs. The inputs to this estimation are the values of y and x on independent random samples of sizes n_1, ..., n_r drawn from r clusters selected at random from the k clusters of the population. Other needed inputs are the cluster sizes N_1,, ..., N_r of the selected clusters and the total population size N. From the samples we compute $\overline{y}_i$, $\overline{x}_i$, and $s_{i,yx}$ for i = 1, ..., r (or else $\overline{y}_i$, $\overline{x}_i$, and $\overline{yx}_i$, using the shortcut formula for sample covariance). The estimates are then given in (1) with either x or y as the variable, and in (2).

In this way we can obtain covariance estimates for y and x_i, i = 1, ..., p, and for x_i and x_j, $1 \leq i, j \leq p$. Then they can be assembled according to the formulas of the preceding section to give estimates for all regression and correlation coefficients.

4 The Stratified Case

Suppose the population is divided into m strata, and each stratum is divided into clusters. Then the same methodology can be used to estimate the covariances of population variables and hence their regression and correlation coefficients. Indeed the methods of the previous section will give us unbiased estimates for

ν_h, μ_h, $Cov_h(y,x)$, and the like, the means and covariances within stratum number h. If the strata have sizes M_h and the population has size $M = \Sigma_{h=1}^m M_h$, the M_h playing the role of N in the last section, then the grand means ν and μ of y and x are given by:

$$\nu = \Sigma_{h=1}^m \frac{M_h}{M}\nu_h, \mu = \Sigma_{h=1}^m \frac{M_h}{M}\mu_h,$$

and the overall covariance of y and x is:

$$Cov(y,x) = \Sigma_{h=1}^m \frac{M_h}{M} Cov_h(y,x) + \Sigma_{h=1}^m \frac{M_h}{M}(\nu_h - \nu)(\mu_h - \mu).$$

The second set of terms on the right side of the last equation can be rewritten as:

$$\Sigma_{h=1}^m \frac{M_h}{M}(\nu_h - \nu)(\mu_h - \mu) = (\Sigma_{h=1}^m \frac{M_h}{M}\nu_h\mu_h) - \nu\mu = \Sigma_{h=1}^m(\frac{M_h}{M} - (\frac{M_h}{M})^2)\nu_h\mu_h - \Sigma_{h\neq h';h,h'=1}^m \frac{M_h}{M}\nu_h\frac{M_{h'}}{M}\mu_{h'}.$$

As already noted, $Cov_h(y,x)$ can be estimated in an unbiased way by the methods of the previous section. The quantities $\nu_h\mu_h$ and $\nu_h\mu_{h'}$ can be rewritten as:

$$E_h(yx) - Cov_h(y,x) \text{ and } E_h(y)E_{h'}(x),$$

and each of these expression has an unbiased estimate. In particular $E_h(yx)$ can be estimated by $\overline{yx}_{cl}$ as in Equation (1), which expresses the estimate in terms of $\overline{yx}_i$, the sample mean of yx on a sample within cluster number i from stratum number h. Since $\overline{yx}_i = \overline{y}_i\overline{x}_i - ((n_i - 1)/n_i)s_{i,yx}$, this can be reduced to the same sample estimates as before. Likewise $E_h(y)E_{h'}(x)$ for $h \neq h'$, because of independence of the strata, can be estimated without bias by $\overline{y}_{cl}\overline{x}_{cl}$, and by (1) this reduces to sample means of y and x within clusters in stratum number h. Notice that we can assume stratum number h consists of k_h clusters and from these clusters r_h clusters are chosen. Thus k and r can vary from stratum to stratum.

5 Alternative Designs

In large scale sample surveys, the clusters, or primary sampling units, are often drawn within a stratum with probability proportional to size (with or without replacement). The with-replacement case can be handled relatively easily by the methods above. The without-replacement case, when two clusters are selected from each stratum, uses Brewer's method or Murthy's method ([1], pp. 261-5). The method of Rao, Hartley, and Cochran, or other methods related to systematic sampling,([1], pp. 265-7) can be used when more than two clusters are drawn from each stratum. In all of these without-replacement cases, the approach described here can be applied, but the resulting formulas are complex.

Instead of choosing simple random samples within the clusters to estimate within-cluster means and covariances, we can also proceed to additional stages and obtain more elaborate, but still unbiased, estimates.

References

[1] W. G. Cochran, Sampling Techniques, Third Edition, John WIley & Sons, Inc., New York, 1977.

[2] C. R. Rao, Linear Statistical Inference and its Applications, Second Edition, John Wiley & Sons, Inc., New York, 1973.