Global and hierarchical linear regression in two-stage sampling

Dhirendra Ghosh[1], Andrew Vogt[2]
Synectics for Management Decisions, Inc.[1]
Arlington, VA 22209
Department of Mathematics, Georgetown University[2]
Washington, DC 20057-1233

Abstract

In a two-stage sampling design where a survey collects data on two or more variables of the second stage units, we compare different forms of hierarchical linear regression, as well as global linear regression (where the first stage units, or clusters, are ignored). We consider cases where each cluster in the sample has a large sample size with or without an auxiliary cluster variable.

Keywords: clusters, auxiliary cluster variable.

Regression, univariate and multivariate, is a primary tool in the statistical analysis of data. It unearths relationships between variables and helps to predict unknown values of a variable from other known variables. The general theory of regression has been developed in the case when the observations are independent of each other. If the data are obtained from a sample survey, the underlying assumption for construction of a regression equation is usually a simple random sampling design. However, in real life sampling designs are rarely of this type. Large scale sample surveys often involve multi-stage designs. Thus the first-stage units, called primary sampling units (p. s. u.'s), may be geographical areas, or administrative, commercial, or educational institutions, and subsequent stages of sampling occur only within the sampled p. s. u.'s.

In this paper we restrict our discussion to two-stage sampling designs where the p. s. u.'s are clusters of elementary units, and these p. s. u.s are selected with either equal probability or probability proportional to size (number of elementary units). The elementary units, the second-stage units, are selected with equal probability without replacement. In developing the theory we will gloss over basic procedures that are widely applied.

1 Global approaches

The simplest approach in a two-stage design is to ignore the clustering and treat the observed data as a single aggregated random sample. An improved version of this takes the sampling design partly into account by attaching sampling weights to each individual unit. The sampling weight is the reciprocal of the probablity of selection of the particular unit [1]. Running the regression in the latter way tends to reduce the mean square error of the regression coefficients. In both of these cases, with or without sampling weight, the membership of an individual unit in a particular cluster is forgotten and plays no role in the prediction methodology.

Yet another improvement over the above is to develop consistent estimates of the regression coefficients using cluster by cluster unbiased estimates of covariances and variances and combining them at the population level.

Consider a population of size N divided into k clusters. Let μ be the population mean of a variable x. Then $\mu = \frac{\Sigma_{i=1}^{k} N_i \mu_i}{N}$ where the i-th cluster has size N_i and mean μ_i. An unbiased estimate for μ, based on a simple random sample of r clusters and a simple random sample within each of these clusters, is:

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

Here $\overline{x}_1$, ..., $\overline{x}_r$ are the sample means from the 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.

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.

Now consider the estimation of the covariance of two variables y and x under the same sampling scheme. 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:

$$\frac{k}{r} \Sigma_{i=1}^{r} \frac{N_i}{N} Cov_i(y,x) +$$

[1] We assume in such cases that each sample is ordered and for the unit selected at the k-th step the probability of its selection is the a priori probability that it would be selected at the k-th step.

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

If instead we select simple random samples from each of these clusters and calculate sample means and sample covariances in each sample [2], we obtain the following unbiased estimate of the overall covariance $Cov(y,x)$:

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

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

Now suppose that the clusters are selected, not by simple random sampling, but with probability "proportional to size", where p_i is the probability of selection of the i-th cluster. Within each cluster we shall assume that a census is conducted. It is a simple matter to replace the census results by estimates within each cluster obtained by simple random sampling as above.

First, we consider the case where the clusters are selected with replacement. Then the covariance of y and x is estimated by:

$$\frac{1}{r}\Sigma_{i=1}^{r}\frac{N_i}{N}\frac{Cov_i(y,x)}{p_i}+ \\ \frac{1}{r}\Sigma_{i=1}^{r}\frac{N_i}{N}\frac{\nu_i\mu_i}{p_i}-\frac{1}{r(r-1)}\Sigma_{i\neq j;i,j=1}^{r}\frac{N_i}{N}\frac{\nu_i}{p_i}\frac{N_j}{N}\frac{\mu_j}{p_j}. \qquad (3)$$

Although this estimator has the desirable property of being unbiased, there are cases when it has the wrong sign. For example, when y and x are the same variable, the covariance is the variance, which is intrinsically positive, and the estimator on some samples may turn out to be negative. Some statisticians prefer to use the following estimate for covariance and variance:

$$\frac{1}{r}\Sigma_{i=1}^{r}\frac{N_i}{N}\frac{Cov_i(y,x)}{p_i}+ \\ \frac{1}{r-1}\Sigma_{i=1}^{r}(\frac{N_i\nu_i}{Np_i}-\widetilde{\nu})(\frac{N_i\mu_i}{Np_i}-\widetilde{\mu}), \qquad (4)$$

where $\widetilde{\nu}=\frac{1}{r}\sum_{i=1}^{r}\frac{N_i\nu_i}{Np_i}$ and $\widetilde{\mu}=\frac{1}{r}\sum_{i=1}^{r}\frac{N_i\mu_i}{Np_i}$ are unbiased estimates of ν and μ. The estimator in (4) is always non-negative in the case of variances but has a bias equal to $\sum_{i=1}^{k}\frac{N_i}{N}\nu_i\mu_i(\frac{N_i}{Np_i}-1)$. In case $p_i=\frac{N_i}{N}$ for each cluster, the bias vanishes.

For probability proportional to size without replacement(with the p_i's renormalized at each stage), there are several choices of estimators for covariance, but here we shall only mention the Horvitz-Thompson estimator (again for simplicity we assume a census within each selected cluster), namely,

$$\Sigma_{i=1}^{k}\frac{N_iCov_i(y,x)}{N}\frac{I_i}{\pi_i}+ \\ \Sigma_{i=1}^{k}\frac{N_i\nu_i\mu_i}{N}\frac{I_i}{\pi_i}-\Sigma_{i,j=1}^{k}\frac{N_i\nu_iN_j\mu_j}{\pi_{ij}}, \qquad (5)$$

where all quantities except I_i and I_{ij} are treated as givens. The variable I_i takes the value 1 if the i-th cluster is in the sample of clusters, and the value 0 otherwise. Its expected value is π_i, the probability that the i-th cluster is in the sample. Likewise I_{ij} takes the value 1 if the i-th and j-th clusters are in the sample of clusters, and the value 0 otherwise. Its expected value is π_{ij}, the probability that the i-th and j-th clusters are in the sample. The only drawback of this estimator is that the quantities π_i and π_{ij} are difficult to calculate when the number of clusters selected is more than three.

The various methods just discussed yield estimates for covariances of pairs of variables on the whole population. Since the natural estimators of the regression coefficients are ratios of covariances (and variances), we are led to ratio estimates of the regression coefficients for a regression line at the population level.

If we now want to predict the value of a dependent variable from the independent variables, we use the regression equation just obtained. But notice that our prediction will not make use of a possible cluster to which the new observation might belong.

2 Hierarchical approaches

In a hierarchical approach distinct reqression equations are obtained cluster by cluster. Either we do this directly by assuming that in each selected cluster the sample size is large enough to permit development of a regression equation, or we do it indirectly by obtained a global reqression equation that includes a variable z that varies from cluster to cluster, in which case assigning z a cluster-specifc value specializes the global regression equation to a particular cluster.

Under the former option from the regression equations for each sampled cluster we can get an overall model whose regression coefficients are the averages of the coefficients obtained in each cluster. This overall model can be applied to data from unobserved clusters. The average used should be some measure of center but it may not be the arithmetic mean. Variations can be applied in which one uses one measure for one regression coefficient (for example, the intercept) and another for another regression coefficient. These variations can be tested by comparing their performance on the actual data from the clusters at hand. It is also possible to estimate some coefficients of the overall model in this way and others from the global model in the previous section. More sophisticated approaches are to be found in Raudenbush & Bryk

[2] We assume that the sample size n_i in each cluster is fixed in advance and is not a random variable.

(2001) and Goldstein (2003). Slope and intercept are assumed to be multivariate normal variables varying from cluster to cluster, and iterative procedures are used to estimate first their means, and then their variances, then to reestimate each until the estimates converge. Some of these procedures do not require approximate normality, and some are Bayesian. For a fuller discussion see Goldstein (2003).

If an auxiliary cluster variable z is available for some or all clusters, and we have estimated regression coefficients in the j-th sampled cluster of the form β_j and the value of z is z_j on that cluster, we can propose that $\beta_j = c + dz_j + e_j$ where e_j is an error term. The value of z on an unsampled cluster, or if this value is not available an average of the z_j's, may be used to estimate the regression equation for the unsampled cluster. Alternatively one can take a top-down approach and look for a global model of the form

$$y = a + bz + cx + dzx + e.$$

This can be developed directly from the data and applied to a cluster, as noted already, by specifying the z-value for that cluster.

References

Cochran, W. G. (1977), *Sampling Techniques*, Third Edition, John Wiley and Sons, New York.

Luke, Douglas (2004) *Multi-level Modeling*, Sage Publications, Inc., United Kingdom.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, Second Edition, John Wiley and Sons, London.

Raudenbush, S. W., and Bryk, A. S. (2001) *Hierarchical Linear Models: Applications and Data Analysis Methods*, Second Edition, Sage Publications, Inc., United Kingdom.

Goldstein, H. (2003) *Multilevel Statistical Models*, Third Edition, Hodder Arnold, London.