Influence of informative sampling on covariance between variables

Julia Aru[1]

[1] Statistics Estonia
e-mail: julia.aru@stat.ee

Abstract

Informative sampling can severely bias all kinds of sample estimates. In this paper we concentrate on estimation of covariance matrix. First, the independence in population is considered and conditions of it's preserving in the sample are presented. Possibilities of estimating covariance matrix analytically are illustrated on the basis of multivariate exponential family and parameters of sample distribution derived. Multivariate normal distribution is examined closer and parameters of sample distribution are derived explicitly in matrix form. Some possibilities of increasing the efficiency of estimation in general case are also proposed.

1 Introduction

Present paper is a brief summary of the author's master thesis defended at the University of Tartu in June 2008.

In case of non-ignorable or informative sampling the sampling scheme explicitly or implicitly depends on the response variable. As a result, the sample distribution of the response variable does not reflect the population distribution and does not approximate it after increasing the size of a sample either. The sample estimates are biased for the population parameters. In this paper we concentrate on the sample covariance matrix and study the effect of the informative sampling design on it.

In section 2 we present main notations and relationships, in section 3 inspect the case of independence of variables in the population, in section 4 describe the relationship between sample and populations distribution in case of multivariate exponential family, in section 5 present those relationships in matrix form for multivariate normal distribution and finally in section 6 touch on the estimation of covariance in general case and possibilities of improving it.

2 Population and sample distributions

Let $U = \{1... N\}$ define the finite population of size N. In what follows we consider single stage sampling with inclusion probabilities $\pi_i = \Pr(i \in s)$, $i = 1 \ldots N$. The vector of study variables at object i is denoted by $\mathbf{y}_i = (y_i^1, y_i^2, ..., y_i^k)'$, where k is the number of study variables. Auxiliary variables are denoted by $\mathbf{x}_i$. Let $f_p(\mathbf{y}_i \mid \mathbf{x}_i)$ be the probability density function (pdf) of study variables. Vector of parameters indexing f_p is denoted by $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_m)$.

Sample from population is denoted by s and consists of n objects from U selected according to some sample selection scheme with inclusion probabilities π_i. In practice, π_i often depends on the population values of outcome variable(s), the values of the auxiliary variables and possibly values of design variables used for the sample selection but not included in the working model under consideration.

It can be shown (Pfeffermann et al., 1999), that for random vectors $(\mathbf{u}_i, \mathbf{v}_i)$, $i \in U$, sample pdf can be expressed through the population pdf as follows:

$$f_s(\mathbf{u}_i \mid \mathbf{v}_i) = \frac{E_p(\pi_i \mid \mathbf{u}_i, \mathbf{v}_i) f_p(\mathbf{u}_i \mid \mathbf{v}_i)}{E_p(\pi_i \mid \mathbf{v}_i)} \quad .$$

Expectations E_p and E_s denote here the expectation under the population and sample distribution respectively.

In following sections we consider the case without auxiliary variable $\mathbf{x}$ for simplicity.

3 Independence in the population

Consider the case when variables y^1, .., y^k are independent in the population. Then population pdf can be rewritten as the product of marginal distributions:

$$f_p(\mathbf{y}_i) = f_p(y_i^1) \cdot f_p(y_i^2) \cdot \ldots \cdot f_p(y_i^k) .$$

If the sample selection probabilities have expectations in factorized form,

$$E_p(\pi_i \mid \mathbf{y}_i) = E_p(\pi_i \mid y_i^1) \cdot \ldots \cdot E_p(\pi_i \mid y_i^k),$$

then variables are independent in sample as well. The independence follows since

$$f_s(\mathbf{y}_i) = \frac{E_p(\pi_i \mid \mathbf{y}_i) f_p(\mathbf{y}_i)}{E_p(\pi_i)} = \frac{[E_p(\pi_i \mid y_i^{\,1}) \cdot \ldots \cdot E_p(\pi_i \mid y_i^{\,k})] \cdot [f_p(y_i^{\,1}) \cdot \ldots \cdot f_p(y_i^{\,k})]}{E_p(E_p(\pi_i \mid \mathbf{y}_i))} =$$

$$= \frac{E_p(\pi_i \mid y_i^{\,1}) f_p(y_i^{\,1})}{E_p(E_p(\pi_i \mid y_i^{\,1}))} \cdot \ldots \cdot \frac{E_p(\pi_i \mid y_i^{\,k}) f_p(y_i^{\,k})}{E_p(E_p(\pi_i \mid y_i^{\,k}))} = f_s(y_i^{\,1}) \cdot \ldots \cdot f_s(y_i^{\,k}).$$

So, even in the case of highly informative sampling, the independence between variables can be preserved if effects of different variables in the inclusion probabilities are separated from one another.

4 Population distribution from multivariate exponential family

In some cases sample covariance can be derived analytically.

Let the population distribution belong to the multivariate exponential family, i.e.

$$f_p(\mathbf{y} \mid \boldsymbol{\eta}) = h^*(\mathbf{y}) \exp\left\{ \sum_{j=1}^{m} \eta_j T_j(\mathbf{y}) - B^*(\boldsymbol{\eta}) \right\},$$

where $\boldsymbol{\eta}$ is the vector of canonical parameters, $h^*(\cdot) : \Re^k \to \Re$ and $T_j(\cdot) : \Re^k \to \Re$ are functions of $\mathbf{y}$, $B(\cdot) : \Re^m \to \Re$ is the function of $\boldsymbol{\eta}$.

If inclusion probabilities also have exponential form:

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp\left\{ \sum_{j=1}^{m} p_j T_j(\mathbf{y}) \right\},$$

then it can be shown that sample distribution belongs to the same family as population distribution but with different parameters, $\eta_j^{\,*} = \eta_j + p_j$. The following illustrates this feature on the example of multinomial distribution.

Example 1. Consider the variables $\mathbf{y} = (t, y, z)$ having multinomial distribution with parameters n and $\mathbf{p} = (p_t, p_y, p_z)$, i.e.

$$f_p(\mathbf{y} \mid n, \mathbf{p}) = \frac{n!}{t!\,y!\,z!} p_t^{\,t} p_y^{\,y} p_z^{\,z}, \quad p_t + p_y + p_z = 1, \quad t + y + z = n.$$

Multinomial distribution belongs to multivariate exponential family, since

$$f_p(\mathbf{y} \mid n, \mathbf{p}) = \frac{n!}{t!\,y!\,z!} \exp\{t \log p_t + y \log p_y + z \log p_z\}.$$

Vector of canonical parameters is thus $\boldsymbol{\eta} = (\log p_t, \log p_y, \log p_z)$, $T_i(\mathbf{y}) = y_i$, where $y_1 = t$, $y_2 = y$, $y_3 = z$.

If inclusion probabilities have the form

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp\{a \cdot t + b \cdot y + c \cdot z\},$$

then sample distribution is also multinomial with canonical parameters $\boldsymbol{\eta}^* = (\log p_t + a,\ \log p_y + b,\ \log p_z + c)$. The vector of probabilities of sample distribution is thus $\mathbf{p}^* = (p_t e^a, p_y e^b, p_z e^c)$, $c = \log(1 - p_t e^a - p_y e^b) - \log p_z$. Knowing analytical expressions of sample parameters we can compare correlations of variables *t* and *y* in population and sample. The correlation coefficients are (respectively)

$$\rho_p(t, y) = -\sqrt{\frac{p_t p_y}{(1 - p_t)(1 - p_y)}} \quad \text{and} \quad \rho_s(t, y) = -\sqrt{\frac{p_t e^a p_y e^b}{(1 - p_t e^a)(1 - p_y e^b)}}.$$

So, for example, if *a* and *b* indexing inclusion probabilities are both positive (which means that objects with large *t* and *y* prevail in the sample), then negative correlation between *t* and *y* in the sample is stronger than that in the population.

5 Multivariate normal population distribution

Although multivariate normal distribution belongs to the exponential family, it is more convenient to present the results for it in matrix form, so the parameters of sample distribution will be derived here explicitly.

The multivariate normal density function in matrix form is

$$f_p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{k/2}\sqrt{|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\right\},$$

where $\boldsymbol{\mu}$ is the vector of expectations and $\sum$ is covariance matrix.

The inclusion probabilities are again in exponential form, but we now present them in matrix form:

$$E_p(\pi \mid \mathbf{y}) = c_0 \exp(\mathbf{y}'\mathbf{A}\mathbf{y} + \mathbf{b}'\mathbf{y}).$$

It can be shown (Aru, 2008), that sample distribution is in this case again normal with vector of expectations $\boldsymbol{\lambda}$ and covariance matrix $\boldsymbol{\Omega}$:

$$\boldsymbol{\lambda} = (\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \mathbf{b}),$$
$$\boldsymbol{\Omega} = (\boldsymbol{\Sigma}^{-1} - 2\mathbf{A})^{-1}.$$

From the above expressions we can make some conclusions on the relationship between population and sample covariance matrixes in case of normal distribution:

- Sample covariance matrix is different from population covariance matrix only if matrix **A** is different from the matrix of zeros, that is if the expectation of inclusion probabilities depends on the squares and products of study variables. If **A**=0 then the mean changes but not the structure of dependencies.

- If variables are independent in the population, i.e. $\sum$ is diagonal, then independence is preserved in the sample iff **A** is also diagonal. This fact confirms previous results.

- With appropriate **A** the structure of dependencies between variables can drastically change: dependent variables can become independent, and vise versa, the sign of covariance can change.

6 Estimation in general case

In general case a population covariance can be estimated with ordinary weighted sample covariance. But using conditional expectations of weights with respect to study variables instead of ordinary sample weights can potentially decrease the variability of estimates. To calculate conditional expectation, a model should be fitted to the inclusion probabilities with study variables as auxiliary variables.

This method was tested in a simulation study based on real data from Estonian EU-SILC Survey. Conditional expectation of inclusion probabilities was calculated with several models and results show that with appropriate model decrease in mean square error of covariance estimates is ca 5-13%. This agrees with the results of earlier studies on informative sampling that incorporating modelling of inclusion probabilities into estimation procedure increases the efficiency of estimates.

References

Aru, J. (2008) Influence of informative sampling on covariance between variables. Master thesis. Manuscript at the Institute of Mathematical Statistics of University of Tartu.

Pfeffermann, D., Sverchkov, M. (1999) Parametric and semi-parametric estimation of regression models fitted to survey data. *The Indian Journal of Statistics. Special Issue on Sample Surveys*, Volume **61**, Series **B**, 166-186.