Bayes Linear estimation for finite population with emphasis to categorical data

Kelly Cristina M. Gonçalves *

Fernando A. S. Moura †

Helio S. Migon ‡

Abstract:

A Bayes linear approach is proposed for obtaining estimation of proportions for multiple categorical data associated with finite population units. A numerical example is provided to illustrate it. The advantage of Bayes linear estimation is that it only requires specification of the mean and variance of some model parameters rather than the full distribution. Here, the Bayes linear estimator is obtained from a general linear regression model, in which many common design-based estimators found in the literature can be obtained as particular cases. A new ratio estimator is also proposed for practical situation in which auxiliary information is available.

Key words: exchangeability, linear model, Bayesian linear prediction

*Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil, email: kelly@dme.ufrj.br

†Corresponding author address - Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil, email: fmoura@dme.ufrj.br

‡Corresponding author address - Universidade do Brasil-UFRJ, Rio de Janeiro, Brazil, email: migon@dme.ufrj.br

1 Introduction

Surveys have long been an important way of obtaining accurate information from a finite population. For instance, governments need to obtain descriptive statistics of the population for purposes of evaluating and implementing their policies. For those concerned with official statistics in the first third of the twenty century, the major issue was to establish a standard of acceptable practice. Neyman (1934) created such a framework by introducing the role of randomization methods in the sampling process. He advocated the use of the randomization distribution induced by the sampling design to evaluate the frequentist properties of alternative procedures. He also introduced the idea of stratification with optimal sample size allocation and the use of unequal selection probabilities. His work was recognized as the cornerstone of design-based sample survey theory and inspired many other authors. For example, Horvitz and Thompson (1952) proposed a general theory of unequal probability sampling and the probability-weighted estimation method, the so-called "Horvitz and Thompson's estimator".

The designed-based sample survey theory has been very appealing to official statistics agencies around the world. As pointed out by Skinner et al. (1989), the main reason is that it is essentially distribution-free. Indeed, all advances in survey sampling theory from Neyman onwards have been strongly influenced by the descriptive use of survey sampling. The consequence of this has been a lack of theoretical developments related to the analytic use of surveys, in particular for prediction purposes. In some specific situations, the designed-based approach has proved to be inefficient, providing inadequate predictors. For instance, estimation in small domains and the presence of the non-response cannot be dealt with by the designed-based approach without some implicit assumptions, which is equivalent to assuming a model. Supporters of the designed-based approach argue that model-based inference greatly depends on the model assumptions which might not be true. On the other hand, interval inference for target population parameters (usually total or means) relies on the Central Limit Theorem, which can not be applied in many practical situations, where the sample size is not large enough and/or independent assumptions of the random variables involved are not realistic.

Basu (1971) did not accept estimates of population quantities, which depend on the sampling rule, like the inclusion probabilities. He argued that this estimation procedure does not satisfy the likelihood principle, at which he was adept. Basu (1971) created the circus elephant example to show that the Horvitz-Thompson estimator could lead to inappropriate estimates and proposed an alternative estimator. The question that arises is whether it is possible to conciliate both approaches. In the superpopulation model context, Zacks (2002) shows that some designed-based estimators can be recovered by using a general regression model approach. Little, in Chapter 4 of Chambers and Skinner (2003), claims that: "careful model specification sensitive to the survey design can address the concerns with model specifications, and Bayesian statistics provide a coherent and unified treatment of descriptive and analytic survey inference". He gave some illustrative examples of how standard designed-based inference can be derived from the Bayesian perspective, using some models with non-informative prior distributions.

In the Bayesian context, another appealing proposal to conciliate the designed-based and model-based approaches was presented by O'Hagan (1985) in an unpublished report. O'Hagan (1985)'s approach is based upon the Bayes linear estimator (see Section 2 for further details), which is therefore distribution-free. This methodology is an alternative to the methods of randomization and appears midway between two extreme views: on the one hand the procedures based on randomization and on the other those based on superpopulation models. His model formulation assumes only second-order exchangeability, which in practice means the need of stating first and second moments only, describing prior knowledge about the structures present in the population. He dealt with several population structures, such as stratification and clustering, by assuming suitable hypotheses about the first and second moments and showed how some common designed-based estimators can be obtained as a particular case of his more general approach. He also pointed out that his estimates do not depend on how the sample was selected, that is, he assumed non-informative sampling. He quoted Scott (1977) and commented that informative sampling should be carried out by a full Bayesian analysis. An important reference about informative sampling dealing with hierarchical models can be found in Pfeffermann et al. (2006).

The paper is organized as follows. Section 2 generally describes the Bayes linear estimation (BLE) approach applied to a general linear regression model for finite population prediction and shows how to obtain some designed-based estimators as particular cases. In Section 3 a new ratio estimator is proposed for practical situation in which auxiliary information is available. Section 4 extends the BLE approach to multiple categorical data. Section 5 offers some conclusions and suggestions for further research. The appendix A contains some details of the approach developed in Section 2.

2 Bayes linear estimation for finite population

Consider $U = \{u_1, \ldots, u_N\}$ a finite population with N units. Let $\mathbf{y} = (y_1, \ldots, y_N)'$ be the vector with the values of interest of the units in U. The response vector $\mathbf{y}$ is partitioned into the known observed n-sample vector $\mathbf{y}_s$, and the non-observed vector $\mathbf{y}_{\bar{s}}$ of dimension $N - n$. A general problem is to predict a function of the vector $\mathbf{y}$, such as the total $T = \sum_{i=1}^{N} y_i = \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}'\mathbf{y}_{\bar{s}}$, where $\mathbf{1}_s$ and $\mathbf{1}_{\bar{s}}$ are the vectors of 1's of dimensions n and $N - n$, respectively. In Classical approach, it is usually done by assuming a parametric model for the population values y_i's and then obtaining the Empirical Best Linear Unbiased Predictor (EBLUP) for the unknown vector $\mathbf{y}_{\bar{s}}$ under this model. Usually, the mean square error of the EBLUP of T is obtained by second order approximation, as well as an unbiased estimator of it. See Valliant et al. (2000) for details.

Bayesian approach to finite population prediction often assumes a parametric model, however it aims to find the posterior distribution of T given $\mathbf{y}_s$. Point estimates can be obtain by setting a loss function, although in many practical problems, it is often considered the posterior mean and its associated precision given by the posterior variance, i.e:

$$\begin{aligned} E(T \mid \mathbf{y}_s) &= \mathbf{1}_s'\mathbf{y}_s + \mathbf{1}_{\bar{s}}' E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \\ V(T \mid \mathbf{y}_s) &= \mathbf{1}_{\bar{s}}' V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s) \mathbf{1}_{\bar{s}}. \end{aligned} \tag{1}$$

In this article, we aim to obtain the point estimates in (1) but using Bayes linear estimation approach. This is done by proposing a hierarchical regression model for

finite population, where particular cases describing usually population structure found in practice are easily derived from it.

In many practical situations the relation between the response variable and a set of auxiliary variables can be represented by a regression model. We consider the following robust two-stages hierarchical regression model for finite population prediction purposes, specified only by their respective mean and variance-covariance matrices:

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta} &\sim [\mathbf{X}\boldsymbol{\beta}, \mathbf{V}], \\ \boldsymbol{\beta} &\sim [\mathbf{a}, \mathbf{R}], \end{aligned} \tag{2}$$

where $\mathbf{X}$ is a covariate matrix of dimension $N \times p$, with $X_i = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, N$; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is a $p \times 1$ vector of unknown parameters; $\mathbf{y}$, given $\boldsymbol{\beta}$, is a random vector with mean $\mathbf{X}\boldsymbol{\beta}$ and covariance matrix $\mathbf{V}$. It should be noted that the distributions of $\mathbf{y}$ and $\boldsymbol{\beta}$ are specified only by their respective means and variance-covariance matrices.

Since the response vector $\mathbf{y}$ is partitioned into $\mathbf{y}_s$, and $\mathbf{y}_{\bar{s}}$, $\mathbf{X}$, which is assumed to be known for all units, is analogously partitioned into $\mathbf{X}_s$ and $\mathbf{X}_{\bar{s}}$, and $\mathbf{V}$ is assumed to be partitioned into $\mathbf{V}_s$,$\mathbf{V}_{\bar{s}}$, $\mathbf{V}_{s\bar{s}}$ and $\mathbf{V}_{\bar{s}s}$. The first aim is to predict $\mathbf{y}_{\bar{s}}$ given the observed sample $\mathbf{y}_s$ and then the total T. We did this in the following steps: first, we had used a joint prior distribution that is only partially specified in terms of moments, as follows:

$$\begin{pmatrix} \mathbf{y}_{\bar{s}} \\ \mathbf{y}_s \end{pmatrix} \Bigg| \boldsymbol{\beta} \sim \left[\begin{pmatrix} \mathbf{X}_{\bar{s}}\boldsymbol{\beta} \\ \mathbf{X}_s\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \mathbf{V}_{\bar{s}} & \mathbf{V}_{\bar{s}s} \\ \mathbf{V}_{s\bar{s}} & \mathbf{V}_s \end{pmatrix} \right].$$

Therefore, applying the general result in the appendix, equation (12), the BLE of $E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ and the associated estimate of $V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta})$ are given by:

$$\begin{aligned} \hat{E}(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta}) &= \mathbf{X}_{\bar{s}}\boldsymbol{\beta} + \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}), \\ \hat{V}(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s, \boldsymbol{\beta}) &= \mathbf{V}_{\bar{s}} - \mathbf{V}_{\bar{s}s}\mathbf{V}_s^{-1}\mathbf{V}_{s\bar{s}}. \end{aligned}$$

But since $\boldsymbol{\beta}$ is unknown, we use the hierarchical Bayesian approach and find the BLE of $\boldsymbol{\beta}$ given $\mathbf{y}_s$, $\hat{\boldsymbol{\beta}}$, and its associated variance, $V(\hat{\boldsymbol{\beta}})$, as follows (details in the appendix):

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{C}\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{R}^{-1}\mathbf{a}\right), \\ V(\hat{\boldsymbol{\beta}}) &= \mathbf{C} = (\mathbf{R}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s)^{-1}. \end{aligned}$$

Applying well known properties of conditional expectations and variances, we get:

$$\begin{aligned} \hat{E}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right] &= \mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}} \\ \hat{V}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right] &= \mathbf{V}_{\bar{s}} + V\left[\mathbf{X}_{\bar{s}}\boldsymbol{\beta} \mid \mathbf{y}_s\right] = \mathbf{V}_{\bar{s}} + \mathbf{X}_{\bar{s}}\mathbf{C}\mathbf{X}'_{\bar{s}}. \end{aligned} \tag{3}$$

The general expression of BLE for the total T and its associated variance is obtained by respectively replacing $E(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s)$ and $V(\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s)$ in equations in (1) by their approximations $\hat{E}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right]$ and $\hat{V}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right]$:

$$\begin{aligned} \hat{T} &= \mathbf{1}'_s\mathbf{y}_s + \mathbf{1}'_{\bar{s}}\hat{E}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right] \\ V(\hat{T}) &= \mathbf{1}'_{\bar{s}}\hat{V}\left[\mathbf{y}_{\bar{s}} \mid \mathbf{y}_s\right]\mathbf{1}_{\bar{s}} \end{aligned} \tag{4}$$

Substituting the equations in 3 into 4, we finally have:

$$\begin{aligned} \hat{T} &= \mathbf{1}'_s\mathbf{y}_s + \mathbf{1}'_{\bar{s}}\mathbf{X}_{\bar{s}}\hat{\boldsymbol{\beta}}, \\ V(\hat{T}) &= \mathbf{1}'_{\bar{s}}\left[\mathbf{V}_{\bar{s}} + \mathbf{X}_{\bar{s}}\mathbf{C}\mathbf{X}'_{\bar{s}}\right]\mathbf{1}_{\bar{s}}. \end{aligned} \tag{5}$$

For the sake of illustration, we consider some examples discussed by O'Hagan (1985) and propose a new ratio estimator, which is one of the contributions of our work. All of them can be treated as special cases of the linear model (2).

2.1 Revisiting some common survey designs

2.1.1 Simple random sampling: Full exchangeability

O'Hagan (1985) considered the simple case where the population has no relevant structure, which can be done by setting up:

$$E(y_i) = m,\ V(y_i) = v \text{ and } Cov(y_i, y_j) = c,\ i, j = 1, \ldots, N,\ \forall i \neq j. \tag{6}$$

Applying the general result established in (5) to (6) with $\boldsymbol{\beta}$ of dimension 1, $\mathbf{X} = \mathbf{1}$, $\mathbf{a} = m$, $\mathbf{R} = c$ and $\mathbf{V} = \sigma^2\mathbf{I}$, where $\sigma^2 = v - c$, we obtain the BLE of T and its respective associated variance:

$$\begin{aligned} \hat{T}_{srs} &= n\bar{y}_s + (N-n)\hat{\mu}, \\ V(\hat{T}_{srs}) &= (N-n)\sigma^2 + (N-n)^2 c\,\sigma^2(\sigma^2 + nc)^{-1}, \text{ where} \end{aligned} \tag{7}$$

$\bar{y}_s = n^{-1}\mathbf{1}_s'\mathbf{y}_s$ is the sample mean,

$\hat{\mu} = \omega\bar{y}_s + (1-\omega)m$ is the expected value of the non-observed values of $\mathbf{y}$,

$\omega = \dfrac{n\sigma^{-2}}{c^{-1} + n\sigma^{-2}}$, where $\sigma^2 = v - c$.

It should be noted that $\hat{\mu}$ is a weighted average of the prior mean m and the sample mean $\bar{y}_s$, where ω is the ratio between two population quantities. The mean m can be viewed as the investigator's prior of the true population mean $\bar{y}$. The uncertainty about y_i is split into two components: the uncertainty about the overall level of the y_i's (between variation) and the one with respect to how much each y_i may vary from that overall level (within variation). A useful measure of variability of units within the population is given by $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$. It is not difficult to show that $E(S^2) = v - c = \sigma^2$. Therefore, σ^2 can be interpreted as a prior estimate of variability within the population. We also obtain $V(\bar{y}) = c + N^{-1}\sigma^2$. In many applications, N is large and thus the constant c could be viewed as the between variation.

Letting $v \to \infty$ and keeping σ^2 fixed, that is, assuming prior ignorance, the estimates in (7) yield:

$$\hat{T}_{srs} = N\bar{y}_s \text{ and } V(\hat{T}_{srs}) = N^2\left(1 - \tfrac{n}{N}\right)\tfrac{\sigma^2}{n}.$$

These expressions are very similar to the well-known total estimate and its variance in the designed-based context for the simple random sampling case.

2.1.2 Stratification

Denote by y_{hi} the i^{th} unit, $i = 1, ..., N_h$ belonging to the h strata, $h = 1, .., H$. It is assumed that the stratum sizes, N_h, are known for all strata. The second-order exchangeability within each stratum is stated as:

$$E(y_{hi}) = m_h, Var(y_{hi}) = v_h,$$
$$cov(y_{hi}, y_{hj}) = c_h,\ i \neq j,$$
$$cov(y_{hi}, y_{lj}) = d_{hl},\ h \neq l.$$

The general model (2) is particularized to this case using that $\mathbf{X}_h = \mathbf{1}_{N_h}$ and $\mathbf{V}_h = \sigma_h^2\mathbf{I}_{N_h}$, $\forall\ h = 1, \ldots, H$, $\mathbf{a} = (m_1, \ldots, m_H)'$, $\mathbf{R}$ is an $H \times H$ - matrix with $R_{ij} = c_i$ if $i = j$

and $R_{ij} = d_{ij}$ if $i \neq j$. And the BLE of T and its measure of dispersion is obtained from (5) and these specifications.

2.1.3 Clustered population

If a population is divided into H clusters, where N_h is the number of units in the h^{th} cluster, then $N = \sum_{h=1}^{H} N_h$ is the total size of population. One way to introduce exchangeability in this case is

$$E(y_{hi}) = m_h,$$

$$cov(y_{hi}, y_{lj}) = \begin{cases} \sigma_h^2 + c_h; & h = l, i = j, \\ c_h; & h = l, i \neq j, \\ 0; & h \neq l, \end{cases}$$

for $i = 1, \ldots, N_h$, $j = 1, \ldots, N_l$, e $h, l = 1, \ldots, H$.

In the model (2) we use $\mathbf{X} = (\mathbf{1}_{N_1}, \ldots \mathbf{1}_{N_H})'$ and $\mathbf{V} = diag(\mathbf{V}_1, \ldots, \mathbf{V}_H)$ is a block diagonal matrix, with $\mathbf{V}_h = \sigma_h^2 \mathbf{I}_{N_h} + c_h \mathbf{1}_{N_h} \mathbf{1}'_{N_h}$.

More details about this and other models can be seen in Bolfarine and Zacks (1992) and the BLE of all this examples can be seen in O'Hagan (1985).

3 Auxiliary information: Ratio estimator

In many practical situations, it is possible to have information about an auxiliary variable x_i, for at least all the sample units that are correlated with the variable of interest, y_i. It is assumed that the population mean $\bar{X}$ (or population total) is also known. In practice, x_i is often the value of y_i at some previous time when a complete census was taken. In the BLE setup, we replace some hypotheses about the x's with ones about the first two moments of the rate y_i/x_i. To the best of our knowledge, the new ratio estimator proposed below is a novel contribution in sampling survey theory.

The new ratio estimator is obtained as a particular case of the model (2) and with the hypothesis of exchangeability, used in Bayes linear approach, applied to the rate y_i/x_i for all $i = 1, \ldots, N$, as described below:

$$E\left(\frac{y_i}{x_i}\right) = m,\ V\left(\frac{y_i}{x_i}\right) = v \text{ and } Cov\left(\frac{y_i}{x_i}, \frac{y_j}{x_j}\right) = c,\ i, j = 1, \ldots, N,\ \forall i \neq j. \quad (8)$$

Applying the general result established in (5) to (8) with $\mathbf{X} = (x_1, \ldots, x_N)'$, $\mathbf{a} = m$, $\mathbf{R} = c$ and $\mathbf{V} = \sigma^2 diag(x_1, \ldots, x_N)$, where $\sigma^2 = v - c$, we obtain the BLE of T as follows:

$$\hat{T}_{ra} = n\bar{y}_s + (N-n)\hat{\mu}\bar{x}_{\bar{s}}, \text{ where}$$

$$\hat{\mu} = \omega\frac{\bar{y}_s}{\bar{x}_s} + (1-\omega)m,$$
$$\omega = \frac{\sigma^{-2}n\bar{x}_s}{(c^{-1} + \sigma^{-2}n\bar{x}_s)}.$$

Letting $v \to \infty$ and $n \to \infty$, but keeping σ^2 fixed, we recover the ratio type estimator, found in the designed-based approach: $\hat{T}_{ra} = N\bar{X}(\bar{y}_s/\bar{x}_s)$.

4 Bayes linear method for categorical data

Often one may be interested in cases where the observed characteristic is whether or not the population unit possesses some attribute of interest. We can define a dichotomized variable $y_i = 1$, if the i^{th} unit has that attribute, and refer to this as a success, and $y_i = 0$, otherwise. The design-based approach uses the randomization introduced by the sampling design to justify the distribution of the binary random quantities, see for instance, Cochran (1977) for further explanations. On the other hand, there are many model-based works in the literature for predicting totals or means in the categories of interest. Malec et al. (1997) consider a logistic hierarchical model with two levels, where the clusters are the second one. They also compared the full hierarchical Bayes estimates with empirical Bayes estimates and standard methods. Moura and Migon (2002) present a logistic hierarchical model approach for small area prediction of proportions, taking into account both possible spatial and unstructured heterogeneity effects. Here again, we do not need to make any use of full model assumptions or randomization approach,

but we do need to make some assumptions about the first and the second moments of the random quantities involved.

The BLE for binary data was briefly introduced by O'Hagan (1985), but here we develop it more generally for the case where we are interested in analyzing more than one attribute in a population. The purpose is to describe the estimation of the proportion of successes with categorical data. Let y_{ij} be the variable that represents the unit i, $i = 1, \ldots, N$ in the category j, $j = 1, \ldots, k$ given by

$$y_{ij} = \begin{cases} 1, \text{ if } i\text{-th unit has } j\text{-th attribute;} \\ 0, \text{ otherwise.} \end{cases}$$

The interest is to estimate a vector $\mathbf{p} = (p_1, \ldots, p_k)'$, where $p_j = N^{-1} \sum_{i=1}^{N} y_{i\,j}$, $j = 1, \ldots, k$, is the proportion of success in the category j, given $\mathbf{y}_s$, a vector of dimension nk, defined as $\mathbf{y}_s = (y_{11}, y_{21}, \cdots, y_{n1}, \cdots, y_{1k}, y_{2k}, \cdots, y_{nk})'$. As we are dealing with situations in which for each unit it is only possible to associate a unique attribute, we have $\sum_{j=1}^{k} p_j = 1$. Thus, we only need to estimate $k-1$ parameters, since it follows that $\hat{p}_k = 1 - \sum_{j=1}^{k-1} \hat{p}_j$ and the variance estimate is also analogously obtained by this relation. Often, we do not have all the data $\mathbf{y}_s$, but some statistics, such as the sample proportion. The BLE estimator of $\mathbf{p}$ and its variance can be obtained by developing the general formula in 4 and arriving at:

$$\begin{aligned} \hat{\mathbf{p}} &= \tfrac{n\bar{\mathbf{y}}_s + (N-n)\hat{E}(\bar{\mathbf{y}}_{\bar{s}} | \bar{\mathbf{y}}_s)}{N}, \\ V(\hat{\mathbf{p}}) &= \tfrac{(N-n)^2 \hat{V}(\bar{\mathbf{y}}_{\bar{s}} | \bar{\mathbf{y}}_s)}{N^2}, \end{aligned} \tag{9}$$

where $\bar{\mathbf{y}}_s$ is a k-vector whose j-th position is given by the sample mean for category j, and analogously we have $\bar{\mathbf{y}}_{\bar{s}}$. Moreover, $\hat{E}(\bar{\mathbf{y}}_{\bar{s}} \mid \bar{\mathbf{y}}_s)$ and $\hat{V}(\bar{\mathbf{y}}_{\bar{s}} \mid \bar{\mathbf{y}}_s)$ are similar to the estimators obtained in (3), but instead of the complete vectors we consider $\bar{\mathbf{y}}_{\bar{s}}$ and $\bar{\mathbf{y}}_s$.

In the absence of any other structural information, we suppose that the units in any given category are second-order exchangeable, but we do not assume any exchangeability between units of different categories. Our prior beliefs are expressed for $i = 1, \ldots, N$, $j = 1, \ldots, k-1$, as follows:

$$\begin{aligned} m_j &= E(y_{ij}) = P(y_{ij} = 1), v_j = Var(y_{ij}) = m_j(1 - m_j), \\ cov(y_{ij}, y_{i'j}) &= m_j(m_{jj} - m_j) = c_j, \quad \forall i \neq i' \text{ and } \sigma_j^2 = v_j - c_j = m_j(1 - m_{jj}), \end{aligned}$$

where $m_{jj} = P(y_{i'j} = 1 \mid y_{ij} = 1)$, for all $i \neq i'$.

For $j \neq j'$, we define the covariance between these categories as

$$cov(y_{ij}, y_{i'j'}) = \begin{cases} m_j(m_{j'j} - m_{j'}), & \text{if } i \neq i', \\ -m_j m_{j'}, & \text{if } i = i'. \end{cases}$$

By the general notation in (2), we have $\boldsymbol{\beta}$, a vector of dimension $k-1$, $\mathbf{X}_s = \mathbf{I}_s$ and using that

$$\begin{aligned} E(\bar{\mathbf{y}}_s) &= \mathbf{X}_s\mathbf{a}, \\ Var(\bar{\mathbf{y}}_s) &= \mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s = \mathbf{Q}, \end{aligned}$$

we get $\mathbf{a} = (m_1, \ldots, m_{k-1})'$, $Q_{jj} = c_j + \sigma_j^2/n$ and $Q_{jj'} = m_j(m_{j'j} - m_{j'}) - m_j m_{j'j}/n$. Therefore, the matrix $\mathbf{R} = \{r_{jj'}\}$, $j, j' = 1, .., k-1$ with $r_{jj} = c_j$ and $r_{jj'} = m_j(m_{j'j} - m_{j'})$ and $\mathbf{V}_s = \frac{1}{n}\{v_{jj'}\}$, $j, j' = 1, .., k-1$ with $v_{jj} = \sigma_j^2$ and $v_{jj'} = -m_j m_{j'j}$. Analogously, we get $\mathbf{V}_{\bar{s}} = n/(N-n)\mathbf{V}_s$ and $\mathbf{X}_{\bar{s}} = \mathbf{I}_{\bar{s}}$.

So, the estimator is a $k-1$-vector described in (9) with the corresponding quantities defined above.

4.1 Prior elicitation

Elicitation is the process of formulating a person's knowledge and beliefs about one or more uncertain quantities into a probability distribution for those quantities. According to Garthwaite et al. (2005) it is convenient to think of the elicitation task as involving a facilitator, who helps the expert formulate the expert's knowledge in probabilistic form. In the context of eliciting a prior distribution for a Bayesian analysis, it is the expert's prior knowledge that is being elicited, but in general the objective is to express the expert's current knowledge in probabilistic form. If the expert is a statistician, or is very familiar with statistical concepts, then there may be no formal need for a facilitator, but this is rare in practice.

Garthwaite et al. (2005) presents the process in four stages: the setup stage, which prepare the elicitation, training the experts, identifying what aspects of the problem to elicit, and so on. The stage of eliciting specific summaries of the experts' distributions for those aspects, where psychologists have contributed at least as much to the methodology

as statisticians. The next stage is to fit a probability distribution to those summaries. In practice, this stage often blurs with the previous stage, in the sense that the choice of what summaries to elicit is often influenced by choice of what distributional form the facilitator intends to fit. The last stage involves assessing the adequacy of the elicitation, with the option then of returning to the second stage and eliciting more summaries from the experts.

But, for the estimators proposed in this article, prior beliefs needed to be elicited about very many quantities, but only in the form of prior means, variances and covariances. Garthwaite et al. (2005) presents an example of elicitation of engineers' prior beliefs about quantities relating to the future capital investment need of a water company. An example with full probability specification is also described and contrasted with that. This second describes an elicitation of the beliefs of hydrologists about properties of certain rocks. One of the conclusions is that the engineers were more comfortable with the concepts involved in the first example than the hydrologists.

For the BLE for categorical data presented in (9) there are a rather large number of quantities in the form of probabilities representing prior information. In this section will be presented some restrictions about the prior quantities and some alternative elicitation that may facilitate the process to an expert.

First, as m_j and $m_{jj'}$ are probabilities, and $\mathbf{R}$ and $\mathbf{V}_s$ are the covariance matrices in the model (2), the following restrictions need to be satisfied:

1. $0 < m_j < 1$ and $0 \leq m_{jj'} \leq 1$, $j, j' = 1, \ldots, k-1$;

2. $\mathbf{R}$ and $\mathbf{V}_s$ need to be positive-definite symmetric matrices.

Note in the first condition that $m_j \neq 0$ and $m_j \neq 1$, otherwise it would results in a prior variance for the units, v_j, equals to 0, what makes nonsense.

When eliciting m_j and $m_{jj'}$, $j, j' = 1, \ldots, k-1$, one way to verify if condition (2) is satisfied, is following the next steps:

(i) verify if $\mathbf{R}$ and $\mathbf{V}_s$ is symmetric, that is if $m_j m_{jj'} = m_{j'} m_{j'j}$.

(ii) Given (i), to verify if $\mathbf{R}$ and $\mathbf{V}_s$ are positive-definite matrices, just calculate the eigenvalues of $\mathbf{R}$ and $\mathbf{V}_s$. If the eigenvalues are positive, so the matrices are positive-definite.

The eigenvalues are the roots of the characteristic polynomial. If this polynomial is of degree n, $n \leq 4$, it is possible to get analytically the roots by Bhaskara, Cardan or Ferrari formulas for example, but if $n \geq 5$ in some cases we can only get those by iterative methods. Anyway, until for matrices higher than 2×2, those restrictions based on eigenvalues will not be trivial to get analytically.

On the other hand, if an expert have difficulties in specifying some of these conditional probabilities $m_{jj'}$, the prior correlation may simplify this task. Define $\rho_{jj'}$ as the prior correlation between two different units in categories j and j', that is

$$\rho_{jj'} = corr(y_{ij}, y_{i'j'}) = \begin{cases} \frac{m_{jj}-m_j}{1-m_j}, & j = j', \\ \frac{m_j(m_{j'j}-m_{j'})}{\sqrt{m_j(1-m_j)m_{j'}(1-m_{j'})}}, & j \neq j'. \end{cases},$$

for $i, i' = 1, \ldots, n$, $i \neq i'$, $j, j' = 1, \ldots, k-1$.

Therefore, given $\rho_{jj'}$, $j, j' = 1, \ldots, k-1$, we get

$$m_{jj'} = \begin{cases} m_j + \rho_{jj}(1-m_j) & j = j', \\ \frac{m_j m_{j'} + \rho_{j'j}\sqrt{m_j(1-m_j)m_{j'}(1-m_{j'})}}{m_{j'}}, & j > j', \\ \frac{m_{j'j} m_j}{m_{j'}}, & j < j'; \end{cases} \tag{10}$$

The next theorem presents the conditions satisfied by m_j and $m_{jj'}$ to get the prior restrictions for a BLE for data with three categories.

Theorem 1. *For the BLE for three categorical data, described in (9) for $k = 3$, it is possible to elicit the prior quantities m_j and $m_{jj'}$, for $j, j' = 1, 2$, with the following steps:*

1. *elicit m_1 and m_2, such that $0 < m_j < 1$, $j = 1, 2$;*

2. *given ρ_{12}, we get m_{11}, m_{12}, m_{21} and m_{22} by (10);*

3. verify if the quantities elicited satisfy:

$$m_{11} > m_1 \text{ and } m_{22} > m_2,$$

$$m_{11}m_{22} - m_{11} - m_{22} + 1 > m_{12}m_{21},$$

$$m_{11}m_{22} - m_{11}m_2 - m_1m_{22} > m_{12}m_{21} - 2m_2m_{12}.$$

For cases with more than three categories we propose to substitute the third step by verifying if m_j and $m_{jj'}$, $j = 1, \ldots, k-1$ elicited results in $\mathbf{R}$ and $\mathbf{V}_s$ positive-definite matrices.

4.2 Prior sensitivity analysis

It is interesting to check how inferences change when we vary the prior quantities. First, it will be treated in a simpler case, the BLE for data with two categories. As a particular case of the estimator obtained in (9), we get the BLE for proportion for binary data as

$$\hat{p}_1 = \frac{n\bar{y_1} + (N-n)\hat{\mu}}{N},$$

where

$\hat{\mu} = \omega\bar{y_1} + (1-\omega)m_1$ is the expected value of the un-observed values in category 1,

$\omega = \dfrac{n{\sigma_1}^{-2}}{n{\sigma_1}^{-2} + {c_1}^{-1}}$,

and $\hat{p}_2 = 1 - \hat{p}_1$. Note that σ_1^2 and c_1 depend on $m_{11} = m_1 + \rho_{11}(1 - m_1)$. So we will find how the estimates are affected by ρ_{11}.

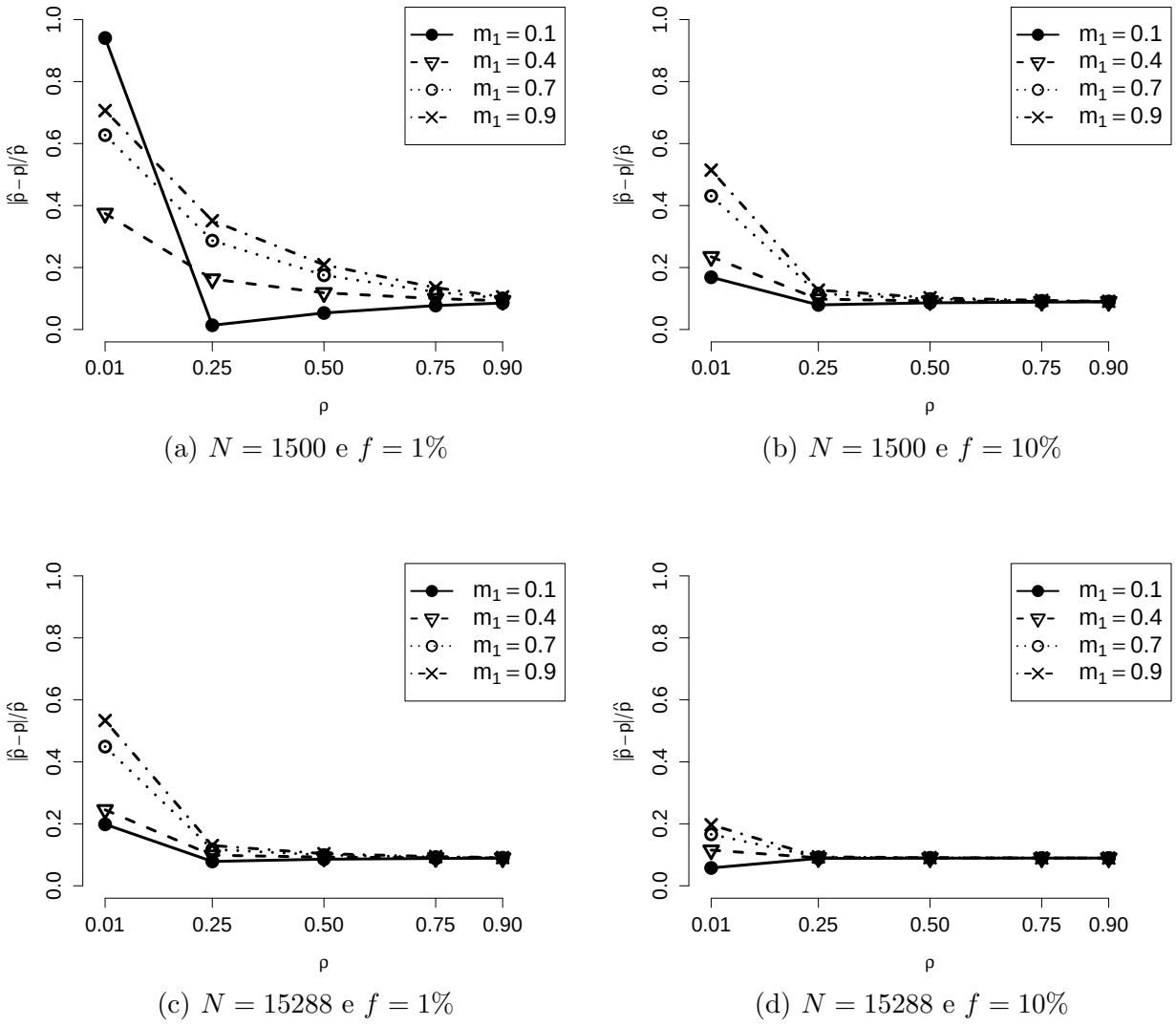
1. If $\rho_{11} \to 0$, $\omega \to 0$ and $\hat{\mu} \to m_1$. So the estimator for the un-observed values depend a lot on the prior mean.

2. If $\rho_{11} \to 1$, $\omega \to 1$ and $\hat{\mu} \to \bar{y}_1$. So the estimator for the un-observed values do not depend on the prior.

Note that these are the only interesting cases because $0 < \rho_{11} < 1$. It happens because $\mathbf{R}$ is the prior variance of the regression parameter, in this case a scalar, so r_{11} have to be greater than zero, then $m_{11} > m_1$.

Moreover, it is trivial to see that if $n/N \to 1$, $\hat{p}_1 \to \bar{y}_1$.

To illustrate this results we created a artificial data with the true proportion $p = (0.2380, 0.7620)'$ and we fixed $\bar{y} = (0.2614, 0.7386)'$ and testes how values of m_1, N, $f = n/N$ and ρ_{11} change the estimator $\hat{p}$. In Figure 4.2 there are the two-dimensional plot of the relative bias $|\hat{p}_1 - p_1|/\hat{p}_1$ versus ρ_{11} for some particular cases. Note that, as f increases the relative bias decreases and when $\rho_{11} \to 0$ the bias increases, principally when m_1 differs a lot of the true p_1.



(a) $N = 1500$ e $f = 1\%$

(b) $N = 1500$ e $f = 10\%$

(c) $N = 15288$ e $f = 1\%$

(d) $N = 15288$ e $f = 10\%$

Figure 1: $|\hat{p}_1 - p_1|/\hat{p}_1$ for fixed $m_1 \in \{0.1, 0.4, 0.7, 0.9\}$, $N \in \{1500, 15288\}$ and $f \in \{1\%, 10\%\}$ and varying $|\rho_{11}| \in \{0.01, 0.25, 0.5, 0.75, 0.9\}$.

5 Conclusions

To elicit a complete joint prior distribution in many dimensions would be an enormous task. The Bayes linear methods only require the elicitation of prior means, variances and covariances for the parameters. This can be easier when a statistical expert is not available to conduct the elicitation. An example of a successful elicitation using this estimator is in O'Hagan (1998).

We derived the well-known designed-based estimators using the structure of the BLE applied to a general regression model approach. We extended the estimator to categorical data and concluded that even if this estimator has many quantities to elicit, it is possible to re-parameterize them or work with non-informative priors. The numerical example illustrated the behavior of the estimates as a function of the sample size and the specifications of the prior parameters.

Acknowledgments

This work is part of the master dissertation of Kelly C. M Gonçalves under the supervision of Helio Migon and Fernando Moura, in the Graduate Program of UFRJ. Kelly has a scholarship from the Office to Improve University Research (CAPES). Helio Migon and Fernando Moura receive financial support from the National Council for Scientific and Technological Research (CNPq-Brazil, BPPesq).

A Appendix: Bayes linear approach

The Bayes approach has been found to be successful in many applications, particularly when the data analysis has been improved by expert judgements. However, Goldstein and Wooff (2007) argues that as the complexity of the problem increases, our actual ability to fully specify the prior and/ or the sampling model in detail is impaired. He concludes that in such situations, there is a need to develop methods based on partial belief specification. One of this methodologies, termed Bayes linear, is fully employed in this article and is briefly described in this appendix.

Let $\mathbf{y}_s$ be the vector with observations and $\boldsymbol{\theta}$ the parameter to be estimated. For each value of $\boldsymbol{\theta}$ and each possible estimate $\mathbf{d}$, belonging to the parametric space $\boldsymbol{\Theta}$, we associate a quadratic loss function $L(\boldsymbol{\theta}, \mathbf{d}) = (\boldsymbol{\theta} - \mathbf{d})'(\boldsymbol{\theta} - \mathbf{d}) = tr(\boldsymbol{\theta} - \mathbf{d})(\boldsymbol{\theta} - \mathbf{d})'$. The main interest is to find the value of $\mathbf{d}$ that minimizes $r(\mathbf{d}) = E[L(\boldsymbol{\theta}, \mathbf{d})|\mathbf{y}_s]$.

Suppose that the joint distribution of $\boldsymbol{\theta}$ and $\mathbf{y}_s$ is partially specified by only their first two moments:

$$\begin{pmatrix} \boldsymbol{\theta} \\ \mathbf{y}_s \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{a} \\ \mathbf{f} \end{pmatrix}, \begin{pmatrix} \mathbf{R} & \mathbf{AQ} \\ \mathbf{QA}' & \mathbf{Q} \end{pmatrix} \right], \tag{11}$$

where $\mathbf{a}$ and $\mathbf{f}$ respectively denote mean vectors and $\mathbf{R}$, $\mathbf{AQ}$, $\mathbf{QA}'$ and $\mathbf{Q}$ the covariance matrix elements of $\boldsymbol{\theta}$ and $\mathbf{y}_s$.

The BLE of $\boldsymbol{\theta}$ is the value of $\mathbf{d}$ that minimizes the expected value of this quadratic function within the class of all linear estimates of the form $\mathbf{d} = \mathbf{d}(\mathbf{y}_s) = \mathbf{h} + \mathbf{H}\mathbf{y}_s$, for some vector $\mathbf{h}$ and matrix $\mathbf{H}$. Thus, the BLE of $\boldsymbol{\theta}$, $\hat{\mathbf{d}}$, and its associated risk matrix, $V(\hat{\mathbf{d}})$, are respectively given by:

$$\begin{aligned} \hat{\mathbf{d}} &= \mathbf{a} + \mathbf{A}\left(\mathbf{y}_s - \mathbf{f}\right), \\ V(\hat{\mathbf{d}}) &= \mathbf{R} - \mathbf{AQA}'. \end{aligned} \tag{12}$$

Now, if we come back to the model (2), the first step is to adapt the structure (11) and use the results in (12) to obtain the BLE of $\boldsymbol{\beta}$ and its measure of dispersion, respectively given by:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \mathbf{a} + \mathbf{R}\mathbf{X}_s'\left(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s\right)^{-1}\left(\mathbf{y}_s - \mathbf{X}_s\mathbf{a}\right), \\ V(\hat{\boldsymbol{\beta}}) &= \mathbf{C} = \mathbf{R} - \mathbf{R}\mathbf{X}_s'\left(\mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s\right)^{-1}\mathbf{X}_s\mathbf{R}. \end{aligned} \tag{13}$$

From the following equations: $\mathbf{C}^{-1} = \mathbf{R}^{-1} + \mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s$ and $\mathbf{A} = \mathbf{R}\mathbf{X}_s'\mathbf{Q}^{-1} = \mathbf{C}\mathbf{X}_s'\mathbf{V}_s^{-1}$, where $\mathbf{Q} = \mathbf{X}_s\mathbf{R}\mathbf{X}_s' + \mathbf{V}_s$, we rewrite (13) as:

$$\hat{\boldsymbol{\beta}} = \mathbf{C}\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s + \mathbf{R}^{-1}\mathbf{a}\right).$$

It should be noted that if we place a vague prior distribution on $\boldsymbol{\beta}$, taking $\mathbf{R}^{-1} \to 0$, we obtain the minimum least square estimator of $\boldsymbol{\beta}$, given by $\hat{\boldsymbol{\beta}}_{LS} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$.

References

Basu, D., 1971. An essay on the logical foundations of survey sampling, Part 1 (with discussion). In: Godambe, Sprott (Eds.), Foundations of Statistical Inference. Holt, Reinhart and Wilnston, Toronto, pp. 203–242.

Bolfarine, H., Zacks, S., 1992. Prediction theory for finite populations. Springer-Verlag New York:.

Chambers, R., Skinner, C., 2003. Analysis of survey data. Vol. 338. John Wiley & Sons Inc.

Cochran, W., 1977. Sampling Techniques. John Wiley and Sons.

Garthwaite, P., Kadane, J., O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association 100 (470), 680–701.

Goldstein, M., Wooff, D., 2007. Bayes Linear Statistics: Theory and Methods. Durhan University, UK: Wiley series in probability and statistics.

Horvitz, D., Thompson, D., 1952. A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47 (260), 663–685.

Malec, D., Sedransk, J., Moriarity, C. L., LeClere, F. B., 1997. Small area inference for binary variables in national health interview survey. Journal of the American Statistical Association 92, 815–826.

Moura, F., Migon, H., 2002. Bayesian spatial models for small area estimation of proportions. Statistical Modelling 2 (3), 183–201.

Neyman, J., 1934. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. Journal of the Royal Statistical Society 97 (4), 558–625.

O'Hagan, A., 1985. Bayes linear estimators for finite populations. Tech. Rep. 58, Department of Statistics - University of Warwick.

O'Hagan, A., 1998. Eliciting expert beliefs in substantial practical applications. The Statistician 47 (1), 21–35.

Pfeffermann, D., Moura, F. A. S., Silva, P. L. N., 2006. Multi-level modelling under informative sampling. Biometrika 93 (4), 943.

Scott, A. J., 1977. Large-sample posterior distributions for finite populations. Annals of Mathematical Statistics 42, 1113–17.

Skinner, C., Holt, D., Smith, T., 1989. Analysis of complex surveys. John Wiley & Sons.

Valliant, R., Dorfman, A., Royall, R., 2000. Finite population sampling and inference: a prediction approach. Wiley New York.

Zacks, S., 2002. In the Footsteps of Basu: The Predictive Modelling Approach to Sampling from Finite Population. Sankhyā: The Indian Journal of Statistics, Series A 64, 532–544.