Design weights and calibration

Kelly Toppin* Luca Sartore† Clifford Spiegelman‡

Abstract

The USDA's National Agricultural Statistics Service (NASS) conducts the U.S. Census of Agriculture in years ending in 2 and 7. Population estimates from the census are adjusted for under-coverage, non-response and misclassification and calibrated to known population totals. These adjustments are reflected in weights that are attached to each responding unit. Calculating these weights has been a two-part procedure. First, one calculates initial (Dual System Estimation or DSE) weights that account for under-coverage, non-response and misclassification. and in the second step, calibration is used to adjust the weights by forcing the weighted estimates obtained in the first step to match known population totals. Recently, a calibration algorithm, Integer Calibration (INCA), was developed to produce integer calibrated weights as required in NASS publications. This paper considers combining the two steps of calculating weights into one. This new algorithm is based on a regularized constrained dual system estimation methodology, which combines capture-recapture and calibration (CaRC).

Key Words: Dual System Estimation; Weights; Census of Agriculture, Calibration

1. Introduction

Weighting is an important part of generating estimates of population statistics from surveys and censuses. Weighting usually consists of two steps: (i) calculating design weights that account for sample selection, these weights are also adjusted for non-response and often for under-coverage and (ii) adjusting these weights so that the estimates are consistent with administrative data. The latter step is called calibration.

Another approach to generating population estimates uses Dual-System Estimation (DSE). DSE uses two independent samples to calculate its estimates. The units that are surveyed twice (recaptured) together with those only surveyed once provide information to compute the probability that a generic unit is sampled. The reciprocals of these probabilities are the DSE weights. Again these weights are adjusted for non-response and often for under-coverage. Calibration is often also applied to the DSE weights to ensure final estimates are consistent with administrative data.

Calibration was introduced by Lemel (1976) as a technique to improve estimates. After Deville (1988), and Deville and Särndal (1992) generalized it, calibration has received much attention in the past three decades. A natural question is: can calibration be performed simultaneously with dual system estimation and does this simultaneous methodology produce better final estimates?

Can one-step methods make better use of the available information? To achieve optimal estimates, a data-driven approach is needed that gives higher priority to the contribution of the information from the surveys rather than from the administrative data; otherwise, biased estimates will likely be produced.

*National Agricultural Statistics Service, United States Department of Agriculture, 1400 Independence Ave. SW, Washington, DC 20250, kelly.toppin@nass.usda.gov

†National Institute of Statistical Sciences, 19 T.W. Alexander Drive, P.O. Box 14006, Research Triangle Park, NC 27709-4006

‡Texas A&M University, 3135 TAMU, College Station, TX 77843-3135

Slud and Thibaudeau (2009) made the first attempt to develop a methodology to perform simultaneously calibration and non-response adjustments. They developed a generalized-ranking calibration that adjusts the weights by minimizing a multi-objective function. Successively, Slud et al. (2013) investigated this technique in presence of soft-constraints, and Shaffer et al. (2014) studied the role of penalty factors. Elkasabi et al. (2015) developed a joint calibration estimator for a dual survey system.

This approach is different since here the focus is combining DSE and calibration. To the best of the authors' knowledge, there have been no attempts to perform calibration and DSE simultaneously.

The proposed technique performs a constrained maximization of the capture-recapture likelihood. The constraints impose consistency with administrative data. A linear logistic model is used to compute the capture probabilities, and variables in the model are selected using an elastic net penalty function. The estimation process consists of a dynamic adjustment of the penalty factors, such that the survey weights are initially estimated and gradually calibrated during the optimization.

2. Methodology

This section formulates the optimization function for the simultaneous calibration and DSE optimization (CaRC). The optimization function is the sum of three distinct parts; logistic regression, penalty function and a calibration offset measure.

2.1 Notation

The following notation is used throughout the paper:

$\mathbf{x_j}$	Vector of covariates
$\mathbf{w}$	Vector of final calibrated weights
$\boldsymbol{\beta}$	Vector of parameters
$\mathbf{A}$	An $n \times p$ matrix of collected data
$\mathbf{a}_i$	The i-th row of matrix $\mathbf{A}$
$\mathbf{y}$	Vector of targets(known totals)

2.2 Logistic regression

Alho (1990) applied logistic regression in a DSE context to the problem of estimating the size of a closed population. Here his setup is followed. The units of a sample $\mathcal{S}$ collected from a finite population $\mathcal{U}$ are usually selected such that $\mathcal{S} \subset \mathcal{U}$. A special case arises when all the units of the population are in the sample. When the conditions $\mathcal{U} = \mathcal{S}$ is satisfied, the sample $\mathcal{S}$ is a census. The goal of a census is to observe every unit in the target population. However, this complete enumeration rarely occurs due to under-coverage, non-response and, possibly, misclassification. Thus, most censuses are extensive surveys that require the extension of the inferential results to the entire population. This can be achieved by applying standard capture-recapture techniques on the data collected from two finite samples $\mathcal{S}_1 \subset \mathcal{U}$ and $\mathcal{S}_2 \subset \mathcal{U}$ when the intersection $\mathcal{S}_1 \cap \mathcal{S}_2 \neq \emptyset$. Under the assumption that $\mathcal{U} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2) \neq \emptyset$, it is reasonable to model the selection probabilities using a multinomial distribution

$$(v_{j10}, v_{j01}, v_{j11}, v_{j00})^\top \sim \mathsf{Mult}(1;\ \pi_{j10}, \pi_{j01}, \pi_{j11}, \pi_{j00}),$$

where π_{j10}, π_{j01}, π_{j11} and π_{j00} are the probabilities associated with the indicator variables for a generic sample unit $s_j \in \mathcal{U}$:

$$v_{j10} = \mathbb{1}_{\mathcal{S}_1 \setminus \mathcal{S}_2}(s_j), \qquad v_{j01} = \mathbb{1}_{\mathcal{S}_2 \setminus \mathcal{S}_1}(s_j),$$
$$v_{j11} = \mathbb{1}_{\mathcal{S}_1 \cap \mathcal{S}_2}(s_j), \qquad v_{j00} = 1 - \mathbb{1}_{\mathcal{S}_1 \cup \mathcal{S}_2}(s_j),$$

where the notation is

$$\mathbb{1}_{\mathcal{S}}(s) = \begin{cases} 1, & \text{if } s \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

The final formulation of the likelihood is simplified by considering only the observations in the second sample (Aberenthy et al., 2017), so that the following probability is adopted for making an inference:

$$\Pr(s_j \in \mathcal{S}_1 \cap \mathcal{S}_2 | s_j \in \mathcal{S}_2) = \pi_j^{z_j}(1 - \pi_j)^{1 - z_j}$$

where $z_j = v_{j11}$, and

$$\pi_j = \frac{\pi_{j11}}{\pi_{j01} + \pi_{j11}}.$$

To model the probabilities π_j for any $s_j \in \mathcal{S}_2$, a functional form depending on a vector of parameters $\boldsymbol{\beta} \in \mathbb{R}^q$ is assumed. The likelihood can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{s_j \in \mathcal{S}_2} \pi_j(\boldsymbol{\beta})^{z_j} \left\{1 - \pi_j(\boldsymbol{\beta})\right\}^{1 - z_j},$$

where $\pi_j(\boldsymbol{\beta})$ is a regression model usually based on the logistic function.

Once the parameters are estimated, the optimal weights w_j are computed as $\pi_j(\hat{\boldsymbol{\beta}})^{-1}$, for any $s_j \in \mathcal{S}_1$, and the total number of units in the population $\mathcal{U}$ is estimated as

$$\hat{N} = \sum_{s_j \in \mathcal{S}_1} \pi_j(\hat{\boldsymbol{\beta}})^{-1}.$$

To enforce certain boundaries on the weights, it is necessary to reformulate the probabilities such that a generic weight $w_j \in [1, u_j]$, where $u_j \geq 1$ is defined for any unit $s_j \in \mathcal{S}_1 \cup \mathcal{S}_2$. These boundaries are attained by minimizing the negative log-likelihood where the probabilities involved are formulated as

$$\pi_j(\boldsymbol{\beta}) = \frac{1 + u_j \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)}{u_j + u_j \exp\left(\mathbf{x}_j^\top \boldsymbol{\beta}\right)},$$

where $\mathbf{x}_j \in \mathbb{R}^q$ is a vector of covariates.

By substituting the probabilities in the formulation of the likelihood, one can obtain the following negative log-likelihood

$$\ell(\boldsymbol{\beta}) = \sum_{s_j \in \mathcal{S}_2} \left\{ \log\left(1 + \exp(\mathbf{x}_j^\top \boldsymbol{\beta})\right) - z_j \log\left(1 + u_j \exp(\mathbf{x}_j^\top \boldsymbol{\beta})\right) \right\}$$

2.3 Penalty function

The regression model is fitted using all the predictors with the elastic-net penalty function developed by Zou and Hastie (2005). The elastic-net penalty linearly combines the LASSO and ridge penalties (Zou and Hastie, 2005; Friedman et al., 2010),

$$(1-\alpha)\frac{1}{2}||\beta||_2^2 + \alpha||\beta||_1,$$

where the notation $\|\cdot\|_1$ represents the L^1-norm used to perform the LASSO regularization, and $\|\cdot\|_2$ denotes the L^2-norm for the ridge regularization. The factor α controls the compromise between the LASSO ($\alpha = 1$) regularization and ridge ($\alpha = 0$) regularization. The combination of these two penalties has several advantages. In particular, the elastic-net is useful in situations where there are many correlated predictor variables as is the case in the USDA's Census of Agriculture.

2.4 Calibration equations

The calibration equations are encapsulated with the following loss function:

$$F(\mathbf{y} - \mathbf{A}g(\boldsymbol{\beta})) = \sum_{k=1}^{p}\left|\frac{\mathbf{a}_k^\top g(\boldsymbol{\beta}) - y_k}{y_k}\right|$$

where $g(\boldsymbol{\beta})^\top = (\pi_1(\boldsymbol{\beta})^{-1}, \pi_2(\boldsymbol{\beta})^{-1}, \ldots, \pi_N(\boldsymbol{\beta})^{-1})$ Since the quantities $\left|\frac{\mathbf{a}_k^\top g(\boldsymbol{\beta}) - y_k}{y_k}\right|$ are the relative errors for any $k = 1, \ldots, p$, these functions ensure that the totals produced by the optimization are close to the known population totals.

The three components are combined to obtain the following objective function, which should be minimized with respect to $\boldsymbol{\beta}$:

$$\omega(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \lambda_\beta\left((1-\alpha)\frac{1}{2}||\boldsymbol{\beta}||_2^2 + \alpha||\boldsymbol{\beta}||_1\right) + \lambda_w\sum_{k=1}^{p}\left|\frac{\mathbf{a}_k^\top g(\boldsymbol{\beta}) - y_k}{y_k}\right|,$$

where $\ell(\cdot)$ is a negative log-likelihood. λ_β and λ_w are positive scalar quantities that balance the effects of the elastic net penalty and the calibration equations, respectively. This method produces calibrated weights by taking into account simultaneously variable selection, model fitting, and calibration adjustments.

3. Case study

This section works through a simple simulation to demonstrate the algorithm. The population U consisting of 10000 units is generated by simulating 100 variables such that $a_{kj} = V_k Z_k$, where $V_k \sim \mathsf{Poisson}(\gamma_k)$, with $\gamma_k \sim \mathsf{Gamma}(2, 0.05)$, and $Z_k \sim \mathsf{Bernoulli}(\omega_k)$, with $\omega_k \sim \mathsf{Unif}(0, 1)$, for any $k = 1, \ldots, 100$, and $j = 1, \ldots, 10,000$. The tuning parameters, $\lambda_{\boldsymbol{\beta}}$ and λ_w are order pairs with $\lambda_{\boldsymbol{\beta}} \in \{e^{-10}, e^{-6.25}, e^{-2.5}, e^{1.25}, e^5\}$ and $\lambda_w \in \{e^{-5}, e^{-2}, e, e^4, e^7\}$.

The optimizations at each ordered pair are evaluated and the weights are generated for models using α values equal to 0, 0.5 and 1. Table 1 shows the results of the simulation. Table 1 shows that the LASSO penalty is the best penalty when the ratio of the number of target variables to total variables is low, and the ridge penalty is the best when the ratio is high. There are two important factors of the simulation to highlight. The first is that a target is considered attained if its estimate is within $\pm 10\%$ of the target valued. The second is that the weights presented here are rounded using INCA, see Sartore and Toppin (2016).

	Ridge (α=0)		α=0.5		LASSO (α=1)	
# of targeted variables	MV	Rel Error	MV	Rel Error	MV	Rel Error
20	24	0.763	25	0.844	23	0.844
40	25	0.794	22	0.622	21	0.622
60	16	0.501	12	0.563	14	0.563
80	5	0.29	9	0.223	9	0.223
100	0	0.096	0	0.098	0	0.099

Table 1: Table showing the number of missed targets and relative error for α = 0,0.5 and 1 for various amount target variables. MV is the number of missed variables out of a total of 100 variables.

4. Summary

This is a preliminary report on this research to spark interest and generate feedback about this approach of combining the two steps of weight procedures into a single step. The plan is to explore methods of moving λ_β and λ_w from a discrete grid to a continuous search algorithm. An important question must be answered before this method can be implemented at the NASS. How should the weights generated by CaRC and the current two step method (DSE followed by calibration) be compared? It is also important to consider the fact that currently at NASS the variables used to generate the DSE weights are different from the calibration variables. The results of CaRC are encouraging and exploration of techniques to improve CaRC's methodology is the next step in this research.

References

Aberenthy, J., Sartore, L., Benecha, H., and Spiegelman, C. (2017). Estimation of capture probabilities by accounting for sample designs. In *2017 Joint Statistical Meetings paper*.

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics*, pages 623–635.

Deville, J.-C. (1988). Estimation linaire et redressement sur information auxiliaire d'enqutes par sondage.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Elkasabi, M. A., Heeringa, S. G., and Lepkowski, J. M. (2015). Joint calibration estimator for dual frame surveys. *Statistics in Transition*, 16(1):7–36.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

Lemel, Y. (1976). Une gnralisation de la mthode du quotient pour le redressement des enqutes par sondage. *Annales de l'ins*, (22/23):273–282.

Sartore, L. and Toppin, K. (2016). *inca: Integer Calibration*. R package version 0.0.2.

Shaffer, B., Cheng, Y., and Slud, E. (2014). Single-stage generalized raking application in the american housing survey. In *2014 Joint Statistical Meetings paper*.

Slud, E., Grieves, C., and Rottach, R. (2013). Single stage generalized raking weight adjustment in the current population survey. In *2013 Joint Statistical Meetings paper*.

Slud, E. V. and Thibaudeau, Y. (2009). Simultaneous calibration and nonresponse adjustment. In *2009 Joint Statistical Meetings paper*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.