Additive Random Coefficient (ARC) Models for Robust Small Area Estimation

Ralph E. Folsom¹, Akhil K. Vaish¹, Avinash Singh²

¹RTI International, 3040 Cornwallis Road, RTP, 27709
²NORC at the University of Chicago, Chicago, IL 60603

Abstract

Unit or person-level ARC models with linear, logistic, and log-linear marginal mean functions are developed for small area estimation. ARC models take the form of a first order Taylor series approximation to the associated general linear mixed model. The area-level random coefficient vectors specify effects for demographic groups. Protection against nonignorable sample designs is provided by a hybrid solution that combines the marginal [probability (P) sampling plus ARC model (ξ)] distribution of the fixed regression coefficients with the MCMC simulated Bayes posterior distributions for the small area specific random coefficient vectors. Survey weighted estimating equations are employed in the solution for the fixed and random coefficients along with sample design consistent covariance matrix estimators. A generalized design effect matrix is used to stabilize the area-level covariance matrices for the random coefficients. A simulation study for the logistic ARC model contrasts the new method’s performance with a nonlinear version of You and Rao’s (2003) pseudo hierarchical Bayes solution that discounts the effect of nonignorable samples on the mean squared errors of small area estimates.

Key Words: small area estimation; generalized design effects; nonignorable survey sample design; general linear mixed model; additive random coefficient; survey weighted estimating equations

1. Introduction

The Additive Random Coefficient (ARC) models utilized here for small area estimation (SAE) are versions of the generalized liner mixed model (GLMM) with additive random coefficients. ARC models can be derived as a first order Taylor series approximation to a GLMM. Our unit-level ARC model is an extension of Singh and Verret’s (2006) aggregate level GLMARC model. Small area estimators based on unit level models have the potential to be more precise than aggregate model estimators. This potential derives from the unit model’s use of person and neighbourhood level fixed predictors and main effect type random coefficients for demographic groups. A major impediment to using unit level models for SAE has been the difficulty in fully accounting for complex nonignorable sample designs. We provide such a solution here based on the unit level ARC model where

\[ E(y_{dk} | \eta_{d}) = f(X_{dk} \beta) + [\hat{\partial} f(X_{dk} \beta)]Z_{dk} \eta_{d} \]  

(1)
for $d = 1, \cdots, m$ small areas and $k = 1, \cdots, n_d$ area-$d$ respondents. In this model $E(y \mid \eta)$ equals a nonlinear marginal mean function $f(X \beta)$ and an additive random effect $\eta$ contribution with a derivative multiplier $\partial f$.

For the logistic ARC model

$$\text{Prob}(y_{dk} = 1 \mid \eta_d) = f_{dk} + \partial f_{dk} Z_{dk} \eta_d$$

(2)

where $f_{dk} = [1 + \exp(-X_{dk} \beta)]^{-1}$ and $\partial f_{dk} = f_{dk} (1 - f_{dk})$. A typical $Z$ vector takes the form $Z_{dk} = (1, g_{dk} \cdot q_{dk} \cdot r_{dk})$ with $g$ denoting an indicator for male gender, $q$ specifying a vector of age group indicators, and $r$ containing indicator variables for race/ethnicity groups. We assume that the random effect vectors $\eta_d$ are $q$ variate i.i.d. normal with zero mean and general covariance matrix $\Sigma_\eta$.

The reason we favor the ARC model over GLMM is the significant computational advantage it has for achieving our SAE goals. Our SAE goal is to produce point estimates and mean squared errors (MSEs) that account fully for complex nonignorable sample designs. Existing solutions that account for complex sample features as regards the fixed model parameters $\beta$ and $\Sigma_\eta$ tend to discount the effect of nonignorable design on the MSEs of the random effects. The ARC model’s two key computational advantages are that:

- The marginal means of the $y_{dk}$ have the fixed model form $f_{dk}$. This makes it easier to estimate the fixed $\beta$ coefficients.
- Solutions for the random effect vectors $\eta_d$ do not require Newton-Raphson iterations.

2. Survey Weighted Estimating Functions

To estimate the $\beta$ parameters we first considered the survey weighted pseudo-optimum estimating functions

$$S_w(\beta \mid \Sigma_\eta) = \sum_{d=1}^{m} \sum_{k=1}^{n_d} w_{dk} (\hat{\partial f}_{dk} + v_{dk}) X_{dk}' [y_{dk} - f_{dk} (\hat{\beta}) - \hat{\partial f}_{dk} (\hat{\beta}) Z_{dk}] \hat{\eta}_{wd} (\hat{\beta}, \Sigma_\eta)$$

(3)

where $v_{dk} = E_{\eta_d} \text{var}(y_{dk} \mid \eta_d)$ and $(\hat{\partial f}_{dk} + v_{dk}) = [1 - \hat{\partial f}_{dk} (Z_{dk} \Sigma_\eta Z_{dk})]^{-1}$ for the logistic ARC model. We used an optimum GEE type solution to derive the pure ARC model version of these equations and then inserted survey weights $w_{dk}$ to protect against bias from nonignorable sample designs. Equation (3) includes ARC model weighting factors involving ratios of the $\hat{\partial f}_{dk}$ derivatives and the conditional variances $v_{dk}$ of
\( y_{dk} \mid \eta_d \). For continuous data linear ARC models and count data Poisson/Exponential ARC models these ratios are constants. While the \((\partial f_{dk} + v_{dk})\) ratios are not constant for the logistic model, we chose to discard them anyway in favor of the survey weighted data fitting equations. To further reduce the computational burden, we have chosen to use the consistent but less efficient GEE or SUDAAN type estimating equations that do not include the random effect \( \hat{\eta}_{wd}(\beta, \Sigma_\eta) \) residual corrections. In the second generation of our software design, we plan to provide an option that implements the more efficient version of equation (3) with the \( \hat{\eta}_{wd}(\beta, \Sigma_\eta) \) residual corrections.

Turning to the survey weighted estimates of the random effect vector, we first define the column vectors \( \xi_{dk} = Z_{dk}' \{y_{dk} - f_{dk}(\beta)\} \) for a given \( \beta \). We then compute area-\( d \) level survey weighted total vectors \( \xi_{wd} = \sum_{k=1}^{n_d} w_{dk} \xi_{dk} \). Taking the dual expectation of \( \xi_{wd} \) first over the probability sample-\( s \) given the data-\( y \) and then over the superpopulation ARC model for \( y \) given \( \eta_d \), the resulting expected value is \( \Delta_{\Omega_d} \eta_d \) where \( \Delta_{\Omega_d} \) is the universe level matrix

\[
\Delta_{\Omega_d} = \sum_{k=1}^{N_d} \partial f_{dk}(\beta) Z_{dk}' Z_{dk}.
\]

Computing \( \Delta_{\Omega_d} \) from the universe file for a given \( \beta \) we then use \( \Delta_{\Omega_d}^{-1} \xi_{wd} \) to form \( \xi_{wd} = \Delta_{\Omega_d}^{-1} \xi_{wd} \) which is an unbiased estimate for the \( \eta_d \) realization.

To account further for complex nonignorable sample designs we employ a stratified pps with replacement cluster sample estimator for the variance-covariance matrix of \( \xi_{wd} \), say \( C_{\xi wd} \). The preferred sample designs for SAE are these where the target small areas are design strata. We assume here that the small areas are either design strata or geographic domains comprising parts of one or more design strata. For the latter case, we chose to ignore any cross area sampling covariances between \( \xi_{wd} \) and \( \xi_{wd'} \). Given the \( C_{\xi wd} \) matrices we then use a design effect matrix averaged over the \( m \) small areas to stabilize \( C_{\xi wd} \) which will often be based on relatively few clusters. With \( \bar{C}_{\xi wd} \) denoting the stabilized covariance matrix for \( \xi_{wd} \), the sampling covariance matrix for \( \xi_{wd} \) is

\[
\text{cov}_{\text{s|y}}(\xi_{wd} \mid \eta_d) = \Delta_{\Omega_d}^{-1} \bar{C}_{\xi wd} \Delta_{\Omega_d}^{-1} = \bar{C}_{\tau wd}.
\]

Given this stabilized version of \( C_{\tau wd} \), the correct sample design based shrinkage matrix for predicting \( \eta_d \) is \( \gamma_{wd} = \Sigma_\eta (\Sigma_\eta + \bar{C}_{\tau wd})^{-1} \). This leads to the \( \eta_d \) predictor \( \hat{\eta}_{wd}(\beta, \Sigma_\eta) = \gamma_{wd} \xi_{wd} \) with the mean squared prediction error.
Assuming that $\eta_{wd}$ has a joint distribution over the sample and the ARC model that is $q$-variate normal with mean vector $\eta_d$ and covariance matrix consistently estimated by $C_{r_d}$, then the conditional posterior distribution of $\eta_{d} \mid \varepsilon_{wd}$ is normal with mean vector $\gamma_{wd} \varepsilon_{wd}$ and covariance matrix $\gamma_{wd} C_{r_d}$.

3. The ARC Model Small Area Estimates

Given $\beta, \Sigma_{\eta}$ and our predictor for $\eta_{d}$ we can form the ARC model small area estimates for the area-$d$ vector $\hat{T}_{\Omega d}$ of demographic domain totals. Assuming that the area level sampling fractions $(n_d + N_d)$ are all negligible, the area-$d$ demographic domain totals are predicted by

$$
\hat{T}_{\Omega d} = \sum_{k=1}^{N_d} Z'_{dk} f_{dk} (\beta) + \delta f_{dk} Z_{dk} \hat{\eta}_{wd} (\beta, \Sigma_{\eta})
$$

(6)

The vector $\bar{\Omega}_{d}$ contains the universe level domain totals of the $f_{dk}$ fixed marginal means. With $z_{wd}$ denoting the $w_{dk}$ weighted sample total of $Z_{dk} f_{dk}$ and $y_{wd}$ depicting the corresponding weighted sample total of $Z_{dk} \gamma_{dk}$ the vector of area-$d$ small area estimates has the following composite form

$$
\hat{T}_{\Omega d} = (I - G_d) \bar{\Omega}_{d} + G_d [Y_{wd} - (z_{wd} - \bar{\Omega}_{d})]
$$

(7)

with $G_d = (\Delta_{\Omega d} \Sigma_{\eta} \Delta_{\Omega d})[(\Delta_{\Omega d} \Sigma_{\eta} \Delta_{\Omega d}) + \bar{C}_{\xi_d}]^{-1}$. Note that the compositing matrix $G_d$ has a shrinkage form incorporating the $\Delta_{\Omega d}$ matrices. Note also that $[Y_{wd} - (z_{wd} - \bar{\Omega}_{d})]$ is a vector valued nonlinear survey regression estimator for the domain totals. The conditional posterior covariance matrix for $\hat{T}_{\Omega d}$ is $G_d C_{r_d}$.

4. MCMC Posterior Variance Steps

To estimate the random effects covariance matrix $\Sigma_{\eta}$ we use a hierarchical Bayes solution based on an inverse Wishart prior with $(q+2)$ degrees of freedom and prior mean matrix $\Sigma_{\eta 0}$. The details are sketched in the MCMC steps outlined here:

1. Use SUDAAN to obtain $\beta_0$ and its covariance matrix $C_{\beta 0}$.
2. Sample $t = 1, \ldots, K$ vectors $\beta_t \sim N(\beta_0, C_{\beta 0})$.
3. Access the universe file for area-$d$ and compute $z_{\Omega d}(\beta_t)$ and $\Delta_{\Omega d}(\beta_t)$.  

4726
4. Given $\Sigma_{\eta(t-1)}$ form $\hat{T}_{\Omega_d}$ and its covariance matrix $G_d \hat{C}_d$.

5. Sample $\eta_{d(t)}(iid) \sim N(\gamma_{w_d}z_{w_d}; \gamma_{w_d} \hat{C}_{r_d})$.

6. Form $A_t = \left( \sum_{d=1}^{m} \eta_{d(t)}^2 \right)$ and draw $\Sigma_{\eta(t)} \sim W^{-1}[m + q + 2, (A_t + \Sigma_{\eta(t)})]$.

7. Use the Rao-Blackwell formula to compute the covariance matrix of $\hat{T}_{\Omega_d} = \sum_{t=1}^{K} \hat{T}_{\Omega_d} + K$.

5. Nonignorable Sample Simulation

For our nonignorable sample simulation we generated area-$d$ ($d=100$ areas) populations of $N_d = 6,000$ binary observations using a $N(0,1)$ latent variable and unequal fractions of the conditional means $\mu_{dk}$ as the tail probabilities. Specifically, with $e_{dk} \sim N(0,1)$

$$y_{dk} = 1 \text{ if } e_{dk} \geq \Phi^{-1}[1 - 0.75 \mu_{dk}] \text{ or } e_{dk} \leq \Phi^{-1}[0.25 \mu_{dk}]$$

$$= 0 \text{ otherwise}$$

where $\mu_{dk} = \text{Prob}(y_{dk} = 1 | \zeta_d)$ and $\Phi(\cdot)$ is the $N(0,1)$ cumulative distribution function. We specified the $\mu_{dk}$ using a logistic mixed model

$$\log[\mu_{dk} + (1 - \mu_{dk})] = X_{dk} \alpha + Z_{dk} \zeta_d$$

with $X_{dk} = (1, a_{dk} \cdot r_{dk}, x_{dk})$ and $Z_{dk} = (1, a_{dk} \cdot r_{dk})$ where $a_{dk}$ and $r_{dk}$ denote 1/0 indicators for binary age and race groups; the continuous covariate was generated using $x_{dk} = u_d + \varepsilon_{dk}$ with $u_d \sim N(0,1)$ and $\varepsilon_{dk} \sim N(0,1)$; and $\zeta_d \sim N(\Omega, \Sigma_{\zeta})$.

We set the fixed coefficients $\alpha$ and $\Sigma_{\zeta}$ to generate a wide range of small area domain proportions and then drew 160 sample records from each area-$d$ with 120 drawn via srs from stratum $\Omega_{d+}$ where the latent variables $e_{dk} \geq 0$ and 40 from stratum $\Omega_{d-}$ where $e_{dk} < 0$. We generated such 200 populations with the same fixed predictors $X_{dk}$ and domain indicators $Z_{dk}$ but new random effects $\zeta_d$ and new latent variables $e_{dk}$.

As a competitor for the ARC model we implemented a nonlinear version of You and Rao’s (2003) pseudo-hierarchical Bayes (PHB) solution. The PHB solution assumes the sample is ignorable for posterior variance estimation.

6. Simulation Results

To present the results we focus on one of the two age groups where the sample sizes per area were ~80. Results for the other demographic domains were similar. For comparison the small area totals for age group 1 were converted to percentages. For both the PHB and ARC solutions there were 100 area percentage estimates for each of 200 populations yielding a total of 20,000 estimates for each method. To examine the performance as a
function of the true finite population percentages we formed 20 groups of 1000 area by population combinations using the true finite population percents to rank the 20,000 combinations. Figure 1 graphs the average bias calculations for the 20 groups. Both methods have fairly linear bias plots that exhibit some over shrinkage on the low and high ends. The ARC model fares a bit better in this regard.

**Figure 1. Bias: ARC vs. PHB**

![Graph showing bias vs. population mean for ARC and PHB methods.](image)

Figure 2 shows the true and the estimated root mean squared errors (RMSEs) by group. The two parabolic curves are the true RMSEs. The two linear curves are the estimated RMSEs. Both methods underestimates the true RMSE on the ends but the ARC estimate tracks the true values better in the mid-range.

**Figure 2. RMSE (%): ARC vs. PHB**

![Graph showing RMSE vs. population mean for ARC and PHB methods.](image)
These MSE results are reflected in the interval coverage probabilities plotted in Figure 3. Both coverages drop off considerably in the outer groups with the PHB drops more pronounced. In the mid-range the ARC method stays closer to the desired coverage level.

Figure 3. 95% Coverage Probability: ARC vs. PHB

7. Future Developments

For a more challenging simulation we plan to select nonignorable cluster samples. We will also consider the more efficient estimating equations for $\beta$ coefficients that are conditioned on the estimated random effect vectors.

References


Contact Information

Ralph Folsom: ref@rti.org
Akhil Vaish: avaish@rti.org
Avinash Singh: singh-avi@norc.org