Bayesian SAE using Complex Survey Data
Lecture 6A: Introduction to SAE

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Motivation

Small area estimation (SAE) is an important endeavor since many agencies require estimates of health, economic indices, education and environmental measures in order to plan and allocate resources and target interventions.

SAE is an example of domain (sub-population) estimation.

"Small" here refers to the fact that we will typically base our inference on a small sample from each area (so it is not a description of geographical size).

In the limit there may some areas in which there are no data.

Small Area Estimation

Consider a study region partitioned into n disjoint and exhaustive areas, labeled by i, $i = 1, \ldots, n$.

As a concrete example, suppose we are interested in a particular condition so that the response is a binary outcome, Y_{ik}, for $k = 1, \ldots, N_i$, individuals in area i.

Based on samples that are collected in the areas[1], the aim of SAE include estimation of:

- The population totals:

$$T_i = \sum_{k=1}^{N_i} Y_{ik}.$$

- The prevalence of the condition in each area:

$$\theta_i = \frac{1}{N_i} \sum_{k=1}^{N_i} Y_{ik} = \frac{T_i}{N_i}.$$

[1] though some areas may contain no samples

Background reading on SAE

The classic text on SAE is Rao (2003), with a more recent edition (Rao and Molina, 2015); not the easiest book to read, and little material on spatial smoothing models.

An excellent review of SAE is Pfeffermann (2013).

The SAE literature distinguishes between direct estimation, in which data from the area only is used to provide the estimate in an area, and indirect estimation, in which data from other areas is used to provide the estimate.

Inference for SAE

Design based inference based on weighted estimators

Suppose we undertake a complex design and obtain outcomes y_{ik} in area i, $k \in s_i$, where s_i is the set of samples that were in area i.

Along with the outcome, there is an associated design weight w_{ik}.

Under the design-based approach to inference, it is common to use the weighted estimator of the prevalence:

$$\widehat{P}_i = \frac{\sum_{k \in s_i} w_{ik} y_{ik}}{\sum_{k \in s_i} w_{ik}}.$$

There is an associated variance, that acknowledges the design, $\widehat{V}_i$.

This variance estimate may be obtained analytically, or through resampling techniques such as the jackknife.

Asymptotically (that is, in large samples):

$$\widehat{P}_i \sim \mathrm{N}(P_i, V_i).$$

Direct Estimation

The simplest approach is to simply map the direct estimates $\widehat{P}_i$.

To assess the uncertainty, one may map the lower and upper ends of (say) a 90% confidence interval:

$$\widehat{P}_i \pm 1.645 \times \sqrt{\widehat{V}_i}.$$

If the samples in each area are large, so that $\widehat{V}_i$ is small, then this approach works well.

Hence, as usual, we would like to carry out some form of smoothing, but in the case of complex survey sampling, how should we proceed?

Design effects

The cluster design leads to a loss of information.

The so-called estimated design effect is

$$d_i = \frac{\widehat{V}_i}{\widehat{P}_i(1-\widehat{P}_i)/n_i},$$

and summarizes the information loss.

Define the effective sample size as

$$\widetilde{n}_i = \frac{n_i}{d_i} = \frac{\widehat{P}_i(1-\widehat{P}_i)}{\widehat{V}_i}.$$

Smoothed Direct Estimation

Let $\widehat{\theta}_i$ be the weighted estimator, then consider

$$\widehat{\theta}_i = \text{logit } \widehat{P}_i = \log\left(\frac{\widehat{P}_i}{1-\widehat{P}_i}\right),$$

which is on the whole of the real line.

"Data" Model[2]:

$$\widehat{\theta}_i \sim \text{N}(\theta_i, \widehat{V}_i),$$

where V_i, its variance, is known.

Prior Random Effects Model:

$$\theta_i = \beta_0 + \epsilon_i,$$

where the random effects $\epsilon_i \sim_{iid} \text{N}(0, \sigma^2_\epsilon)$.

[2]We are taking the data as the estimator

Smoothed Direct Estimation

This is very similar to the normal-normal model we saw in Lecture 3.

Fay and Herriot (1979) suggested this hierarchical model, in a landmark paper.

This model acknowledges the design and also smooths, and it is straightforward to add spatial random effects.

Smoothed Direct Estimation

The spatial version of the model has:

"Data" Model:

$$\widehat{\theta}_i \sim \mathsf{N}(\theta_i, \widehat{V}_i),$$

where $\widehat{V}_i$ is known variance.

Prior Model:

$$\theta_i = \beta_0 + \epsilon_i + S_i,$$

with

- $\epsilon_i \sim \mathsf{N}(0, \sigma_\epsilon^2)$.
- $S_i \sim \mathsf{ICAR}(\sigma_s^2)$.

Adding a term $\boldsymbol{x}_i^\mathsf{T}\boldsymbol{\beta}$ to the prior model allows covariate relationships to be investigated.

This model has been investigated and applied with simulated and real data in (Chen *et al.*, 2014; Mercer *et al.*, 2014) and (in a space-time setting) in Mercer *et al.* (2014, 2015) and Li *et al.* (2018).

Simple SAE Models with Simulated Data

SAE for DHS like Data

We simulate a set of data to mimic a simple DHS type design in which we stratify on 8 regions using the Kenya geography.

Data are simulated using a 2-stage cluster design.

To be concrete, we name the variable “Tobacco Use”.

We create a design object and then obtain weighted (direct) estimates of logit in each area y_i along with a design-based variance estimate $\widehat{V}_i$ for $i = 1, \ldots, n = 8$ regions.

SAE for DHS like Data

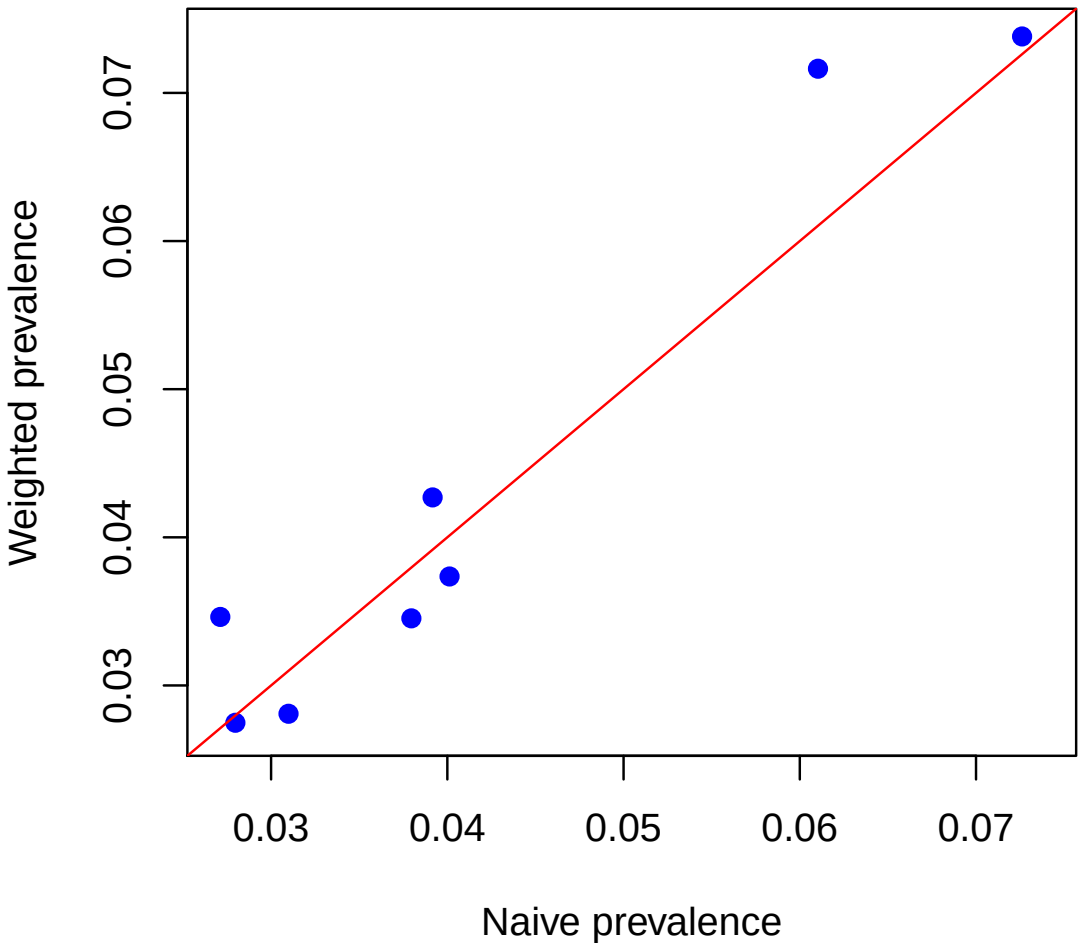

Figure 1: Weighted (direct) estimates of tobacco use versus naive proportions that ignore the design.

SAE for DHS like Data

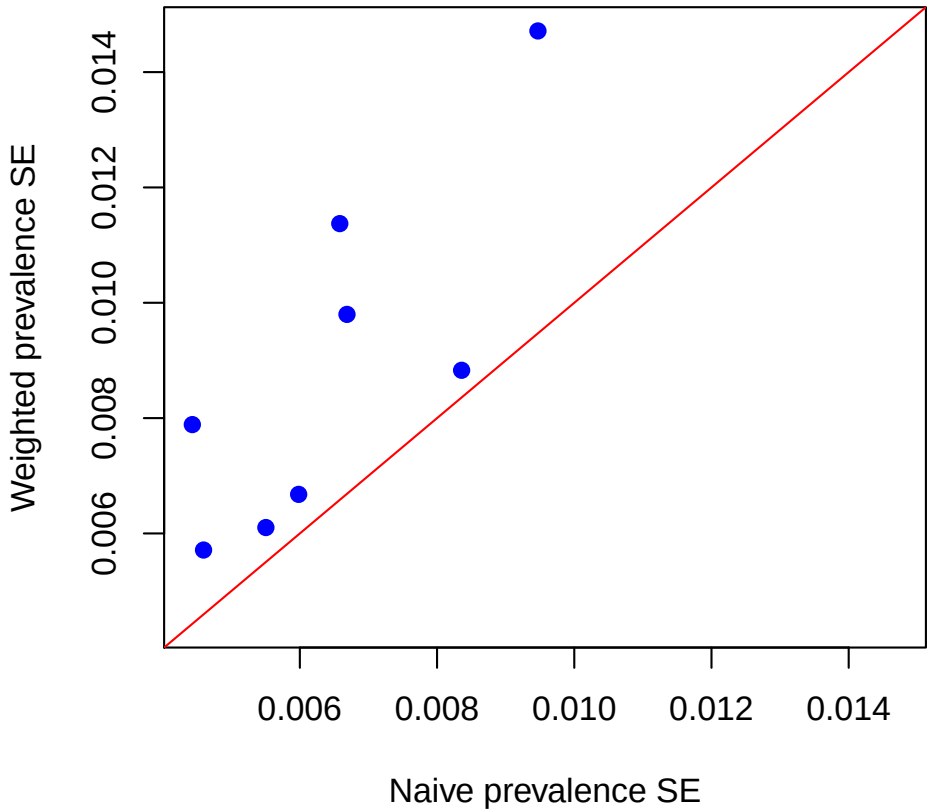

Figure 2: Standard errors of weighted estimates of tobacco use versus standard errors of naive estimates that ignore the design.

The standard errors of the weighted estimates are larger, because they acknowledge the design, in particular, the clustering.

SAE for DHS like Data

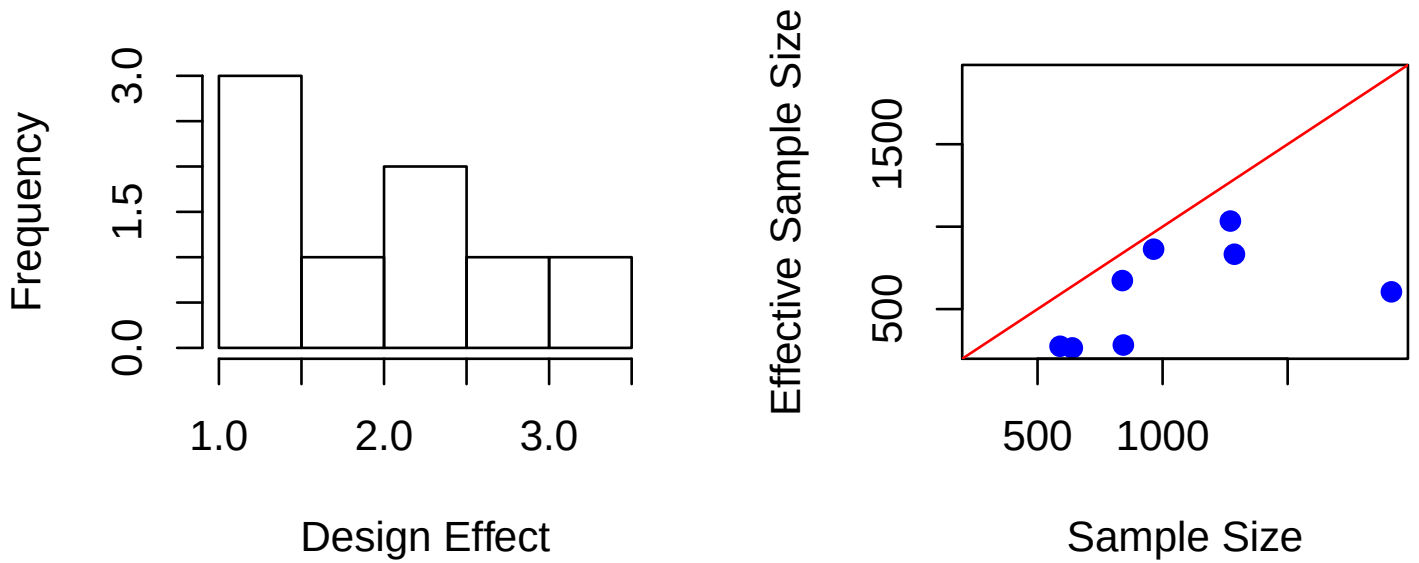

Figure 3: Histogram of design effects (left) and effective sample sizes versus sample sizes n_i.

Effective sample sizes are quite a lot smaller, due to clustering.

Simulated DHS Example

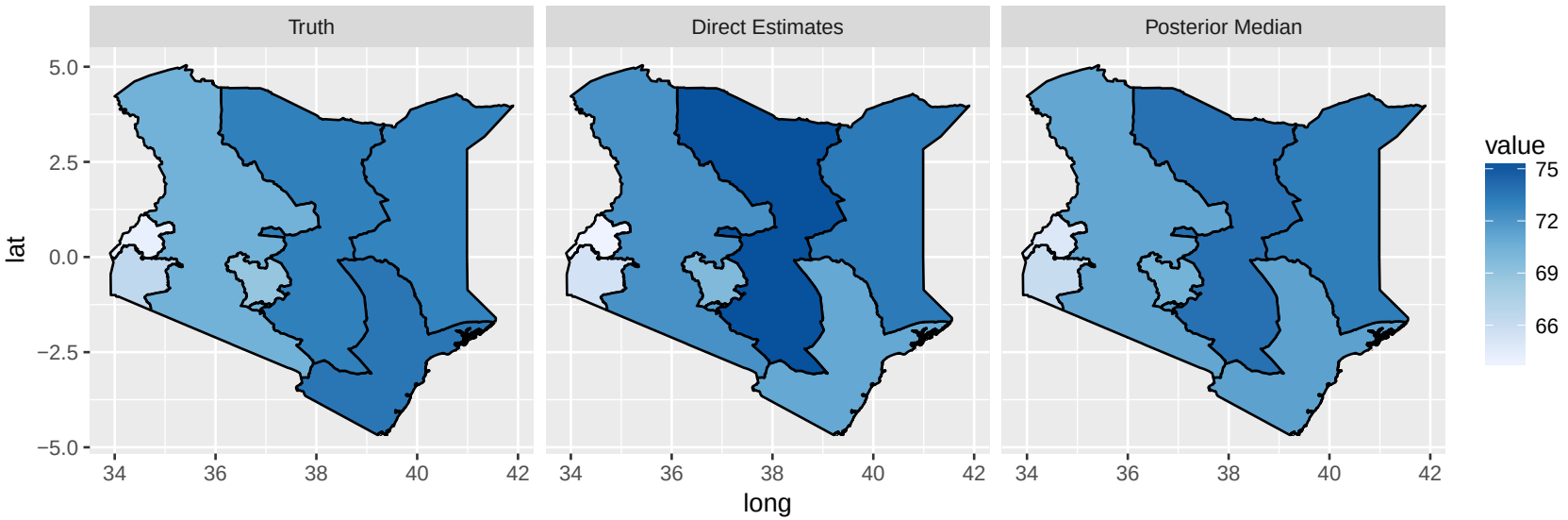

Figure 4: Maps of true proportion of tobacco use (left), the direct estimates (middle) and the posterior medians from the smoothed direct model (right).

Simulated DHS Example

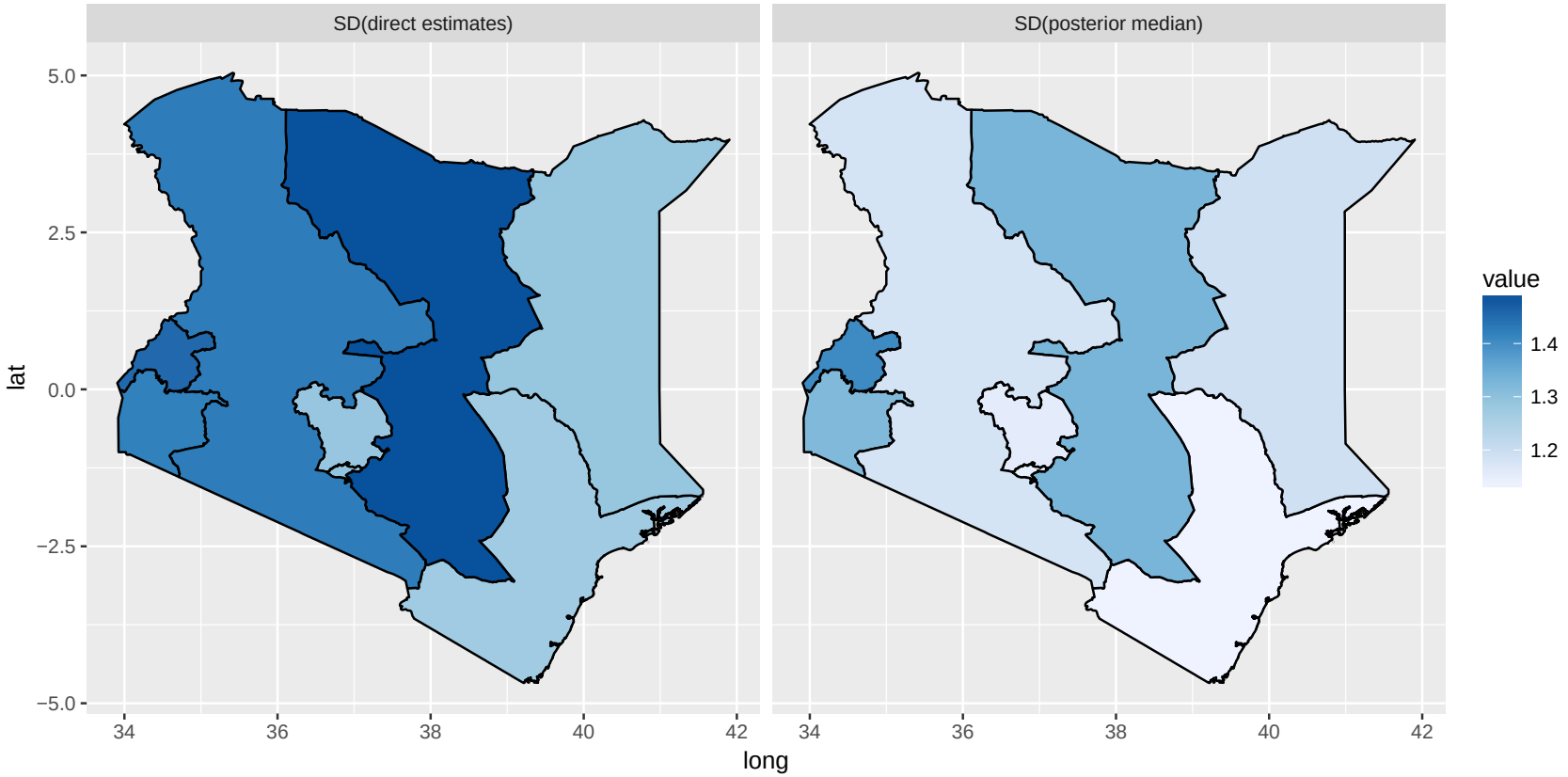

Figure 5: Maps of uncertainty from direct estimates (left), the posterior estimates from the smoothed direct model (right).

SAE for DHS like Data

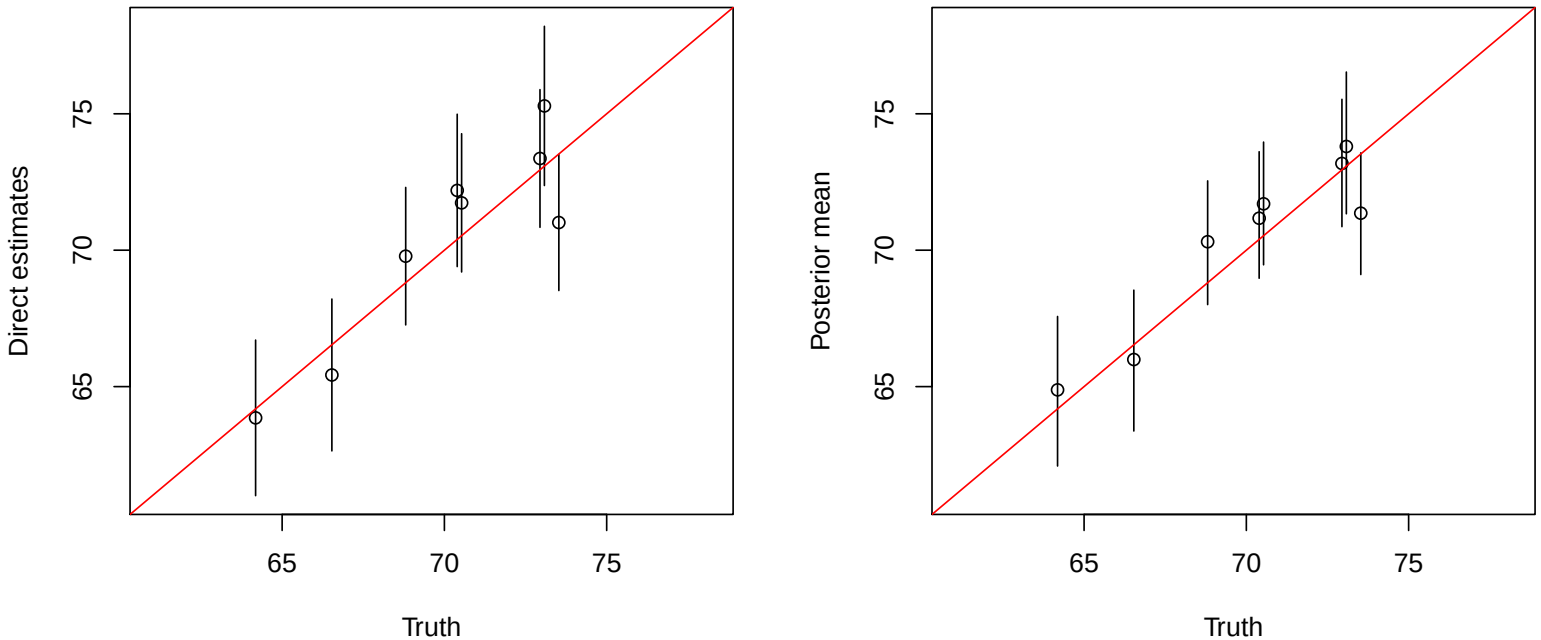

Figure 6: Weighted (direct) estimates of tobacco use with confidence intervals (left) and posterior mean estimates of tobacco use with credible intervals (right).

Discussion

Discussion

Direct smoothed estimates builds on the strengths of weighted estimates and spatial smoothing models.

In the limit the weighted estimates will dominate, which is exactly what we want!

References

Chen, C., Wakefield, J., and Lumley, T. (2014). The use of sample weights in Bayesian hierarchical models for small area estimation. *Spatial and Spatio-Temporal Epidemiology*, **11**, 33–43.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James–Stein procedure to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Li, R., Hsiao, Y., Godwin, J., B. Martin, B., Wakefield, J., and Clark, S. (2018). Changes in the spatial distribution of the under five mortality rate: small-area analysis of 122 dhs surveys in 262 subregions of 35 countries in Africa. *Submitted*.

Mercer, L., Wakefield, J., Chen, C., and Lumley, T. (2014). A comparison of spatial smoothing methods for small area estimation with sampling weights. *Spatial Statistics*, **8**, 69–85.

Mercer, L., Wakefield, J., Pantazis, A., Lutambi, A., Mosanja, H., and Clark, S. (2015). Small area estimation of childhood of childhood mortality in the absence of vital registration. *Annals of Applied Statistics*, **9**, 1889–1905.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.

Rao, J. (2003). *Small Area Estimation*. John Wiley, New York.

Rao, J. and Molina, I. (2015). *Small Area Estimation, Second Edition*. John Wiley, New York.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

Wakefield, J. (2008). Ecologic studies revisited. *Annual Review of Public Health*, **29**, 75–90.

Technical Details: Traditional SAE Approaches

Traditional SAE Approaches

Here we present some material on more standard SAE.

We assume the study region can be partitioned into $i = 1, \ldots, n$ sub-regions (domains) with N_i being the population of the domain, which may or not be known.

A survey is carried out and n_i is the sample size in domain i; if the survey was not designed to fix the sample size n_i for domain i then it is a random variable with respect to the randomization distribution and we need to consider ratio estimation.

Unless specified otherwise, we assume that n_i is random.

Let U_i and S_i, $i = 1, \ldots, n$, be the index sets for the units of the population and the sample respectively in domain i with $U = U_1 \cup \cdots \cup U_n$ and $S = S_1 \cup \cdots \cup S_n$.

Notation

The population mean in domain i is

$$\overline{y}_{U_i} = \frac{\sum_{k \in U_i} y_k}{N_i}.$$

We define

$$\begin{aligned} z_k &= \begin{cases} 1 & \text{if } k \in U_i \\ 0 & \text{if } k \notin U_i \end{cases} \\ u_k = y_k z_k &= \begin{cases} y_k & \text{if } k \in U_i \\ 0 & \text{if } k \notin U_i \end{cases} \end{aligned}$$

so that z_k is just an indicator of whether unit k lies within domain i and u_k is the value of y for such units.

z_k is random when n_i is not fixed in advance.

Notation

Then

- $t_u = \sum_{k=1}^{N} u_k$ is the population total in domain i.
- $\overline{u}_U = \frac{t_u}{N}$ is the population total in domain i divided by N (so this is not the domain mean).
- $t_z = \sum_{k=1}^{N} z_k = N_i$ is the population size in domain i.
- $\overline{z}_U = \frac{N_i}{N}$ is the fraction of the total population in domain i.

Direct estimation without auxiliary information

A direct estimator is one in which response data y from the domain only are used.

An indirect estimator uses responses from other domains.

The population average in domain i can be written as a ratio of totals:

$$\begin{aligned} \overline{y}_{U_i} &= \frac{t_u}{t_z} = B \\ &= \frac{\overline{u}_U}{\overline{z}_U} = \frac{\sum_{k \in U} u_k / N}{N_i / N} = \frac{\sum_{k \in U_i} y_k}{N_i}. \end{aligned}$$

Direct domain estimation

We have, under SRS,

$$\begin{aligned} \overline{u}_S &= \frac{\sum_{k \in S} u_k}{n} \\ \overline{z}_S &= \frac{n_i}{n}. \end{aligned}$$

The ratio domain direct estimator is:

$$\begin{aligned} \widehat{\overline{y}}_i &= \frac{\widehat{t}_u}{\widehat{t}_z} = \widehat{B} \\ &= \frac{\overline{u}_S}{\overline{z}_S} = \frac{\sum_{k \in S} u_k / n}{n_i / n} = \frac{\sum_{k \in S_i} y_k}{n_i}. \end{aligned}$$

This ratio estimator is biased (since n_i is assumed random) but the bias goes to zero with increasing n_i.

This is because $\mathrm{E}[n_i/n] = \mathrm{E}[N_i/N]$ and so both numerator and denominator are unbiased.

Direct domain estimation

Since the ratio estimator is

$$\widehat{\overline{y}}_i = \widehat{B} = \frac{\widehat{t}_u}{\widehat{t}_z},$$

the variance estimator (we emphasize, under SRs) is,

$$\begin{aligned} \widehat{\text{var}}(\widehat{\overline{y}}_i) &= \left(1 - \frac{n}{N}\right) \frac{n}{n_i^2} \frac{(n_i - 1)s_{yi}^2}{n - 1} \\ &\approx \left(1 - \frac{n}{N}\right) \frac{s_{yi}^2}{n_i} \end{aligned} \tag{1}$$

where

$$s_{yi}^2 = \frac{\sum_{k \in S_i} (y_{ik} - \widehat{\overline{y}}_i)^2}{n_i - 1}.$$

An asymptotic 95% confidence interval is $\widehat{\overline{y}}_i \pm 1.96 \times \text{s.e.}(\widehat{\overline{y}}_i)$.

This confidence interval has a randomization interpretation, so 95% of the intervals we construct from samples S will contain the fixed (but unknown) $\overline{y}_U$.

Direct domain estimation

For general (i.e. not SRS) sampling we refer to Särndal et al (1992, Section 10.3).

When N_i is unknown, the domain mean estimator is

$$\widehat{\bar{y}}_i = \frac{1}{\widehat{N}_i} \sum_{k \in S_i} w_k y_k,$$

where $w_k = 1/\pi_k$ and

$$\widehat{N}_i = \sum_{k \in S_i} w_k.$$

Direct domain estimation

To estimate the domain total when N_i is unknown

$$\widehat{t}_{yi} = \sum_{k \in S_i} w_k y_k .$$

To estimate the domain total when N_i is known

$$\widetilde{t}_{yi} = N_i \times \widehat{\overline{y}}_i = \frac{N_i}{\widehat{N}_i} \sum_{k \in S_i} w_k y_k .$$

Variance estimators are given in Särndal et al (1992, Section 10.3).

Direct domain GREG with study auxiliary information

The problem with using the direct ratio estimators is that the variance may be large in areas with low n_i, as in (1).

When auxiliary variable is available, this may be used to define a new estimator; suppose we have a single variable x for which the total is known, across all domains, t_x, and we have a HT estimator $\widehat{t}_x$.

In general, GREG with multiple x values and a linear regression model may be utilized; we describe some special cases.

A ratio estimator is

$$\widehat{t}_i^{\text{dir,rat1}} = \widehat{t}_i \times \frac{t_x}{\widehat{t}_x}, \tag{2}$$

where $\widehat{t}_i$ is the usual HT estimator.

This is a direct domain estimator because y values only from the domain are used, though the total x, t_x from all domains are used.

Direct domain GREG with study auxiliary information

This estimator is approximately unbiased, if the overall sample size n is large (because $\widehat{t}_x \rightarrow t_x$) , and design consistency occurs as the domain sample size n_i increases.

See Rao and Molina (2015, Section 2.4.2) describe GREG estimators, and give a number of special cases including (2).

Notice that the same adjustment is made to every area.

Example: Smoking by county in Washington State

As an example suppose we wish to estimate the number of current smokers across the 39 counties of Washington State, based on a survey.

For each individual in the survey, information is collected on $y_k, k \in S$, a binary indicator of current smoking status, along with x_k, the income and the basic demographics (age and gender).

Suppose we know the total income of residents in Washington State, t_x; then (2) may be directly applied with $\widehat{t}_x = \sum_{k \in s} w_k x_k$ being the estimated total income from the sample.

Direct domain GREG with study auxiliary information

Suppose now we have auxiliary information on the population sample sizes across (usually demographic) groups g, $g = 1, \ldots, G$; this is a special case of the direct estimator.

A post-stratified estimator (Rao and Molina 2015, Section 2.4.2) is

$$\widehat{t}_i^{\,\text{dir,ps1}} = \sum_{g=1}^{G} \frac{N_{.g}}{\widehat{N}_{.g}} \sum_{k \in S_{ig}} w_k y_k = \sum_{g=1}^{G} \frac{N_{.g}}{\widehat{N}_{.g}} \widehat{t}_{ig}. \tag{3}$$

where

- S_{ig} is the set of samples falling in post-stratification group g of domain i,
- $\widehat{t}_{ig}$ is the estimate of the total for y in domain i and group g (note that $\widehat{t}_i = \sum_g \widehat{t}_{ig}$), and
- $\widehat{N}_{.g} = \sum_{k \in S_{.g}} w_k$.

This estimator is approximately unbiased (and is design consistent) but the variance can be large since the adjustments between domain i and the whole region may be large.

Example: Smoking by county in Washington State

Suppose we know the population totals for Washington State by 18 age bands and gender, $N_{.g}, g = 1, \ldots, G = 36$.

In county i, to use (3), we would estimate:

- the total number of smokers by stratum g, $\widehat{t}_{ig} = \sum_{k \in s_{ig}} w_k y_k$, this may have high variability as $|s_{ig}| = n_{ig}$ may be small,
- the population total by group g, across the state,is estimated by $\widehat{N}_{.g} = \sum_{k \in s_{.g}} w_k$.

Direct domain GREG with domain auxiliary information

To reduce the bias we may use domain-specific auxiliary information, as described in (Rao and Molina 2015, Section 2.4.3).

A ratio estimator is

$$\widehat{t}_i^{\,\text{dir,rat2}} = \widehat{t}_i \times \frac{t_{xi}}{\widehat{t}_{xi}}, \tag{4}$$

where $\widehat{t}_i$ is the HT estimator, and the second adjustment term is now area (domain) specific.

This gives an area-specific adjustment.

This is a direct domain estimator since it uses y (and x values) only from the domain.

Direct domain GREG with domain auxiliary information

A post-stratified estimator is

$$\widehat{t}_i^{\,\text{dir,ps2}} = \sum_{g=1}^{G} \frac{N_{ig}}{\widehat{N}_{ig}} \widehat{t}_{ig}, \tag{5}$$

where $\widehat{N}_{ig} = \sum_{k \in s_{ig}} w_k$.

The adjustment term,

$$\frac{N_{ig}}{\widehat{N}_{ig}},$$

is now area-specific, and uses stratified population totals in area i.

Example: Smoking by county in Washington State

For (4), suppose we know the income totals by county (from the census, for example), t_{xi}, and we then estimate $\widehat{t}_{xi} = \sum_{k \in s_i} w_k x_k$.

As another (post-stratified) example, for (5), suppose we know the population totals for Washington State by 18 age bands and gender and by domain (county), $N_{ig}, g = 1, \ldots, G = 36$.

In area i we would then estimate:

- the total number of smokers by stratum g, $\widehat{t}_{ig} = \sum_{k \in s_{ig}} w_k y_k$,
- the population total by group g, in area i, $\widehat{N}_{ig} = \sum_{k \in s_{ig}} w_k$.

Technical Details: Indirect Domain Estimation

Synthetic estimation

Now we consider indirect estimators, and begin with synthetic estimation, as described in (Rao and Molina 2015, Section 3.2).

The simplest synthetic estimator of a domain mean for area i does not use auxiliary information and is

$$\widehat{\overline{y}}_i^{\text{syn,basic1}} = \frac{\widehat{t}_y}{\widehat{N}}, \tag{6}$$

which is the mean over the complete study region.

Large bias will result in domains within which the means deviate from the overall mean, i.e. in which it is not true that $\overline{y}_{U_i} \approx \overline{y}_U$.

The variance of the estimator will be very small, however.

Synthetic estimation

One possibility is to consider a larger region r that contains i rather than the complete study region and use

$$\widehat{\overline{y}}_i^{\text{syn,basic2}} = \frac{\widehat{t}_y(r)}{\widehat{N}(r)}, \tag{7}$$

which is approximately design unbiased if $\overline{y}_{U_i} \approx \overline{y}_{U(r)}$.

These estimators are not design consistent, though the MSE may be relatively small, if the regional sample size is large.

This is a very basic form of spatial smoothing.

Example: Smoking by county in Washington State

For (6), we would estimate the total number of smokers in Washington State, $\widehat{t}_y = \sum_{k \in s} w_k y_k$, and the total population size $\widehat{N} = \sum_{k \in s} w_k$.

For (7), we could split Washington State into (say) contiguous regions based on predictors of smoking.

For example, we could group together contiguous urban and rural counties (this categorization could be based on population density, or percent of farmland,...).

We would estimate the total number of smokers in region r, $\widehat{t}_y(r) = \sum_{k \in s(r)} w_k y_k$, where $s(r)$ is the set of indices of samples in region r and the total population size $\widehat{N}(r) = \sum_{k \in s(r)} w_k$.

Synthetic estimation

With auxiliary information consisting of known totals in domain i, $\boldsymbol{x}_i$, the synthetic estimator is

$$\widehat{t}_i^{\,\text{syn,reg}} = \boldsymbol{x}_i^{\mathsf{T}} \widehat{\boldsymbol{B}}, \tag{8}$$

where

$$\widehat{\boldsymbol{B}} = \left(\sum_{i=1}^{n} \sum_{k \in s_i} w_{ik} \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{x}_{ik} \right)^{-1} \sum_{i=1}^{n} \sum_{k \in s_i} w_{ik} \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{y}_{ik}, \tag{9}$$

is the WLS estimator over all of the units who provide responses, and w_{ik} are the design weights.

This estimator is not design unbiased.

Synthetic estimation

The design bias of $\widehat{t}_i^{\,\text{syn,reg}}$ is approximately $\boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{B} - t_i$, where

$$\boldsymbol{B} = \left(\sum_{i=1}^{n}\sum_{k\in U_i}\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{x}_{ik}\right)^{-1}\sum_{i=1}^{n}\sum_{k\in U_i}\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{y}_{ik}, \tag{10}$$

is the population regression coefficient.

The bias will be small if the domain specific regression coefficient

$$\boldsymbol{B}_i = \left(\sum_{k\in U_i}\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{x}_{ik}\right)^{-1}\sum_{k\in U_i}\boldsymbol{x}_{ik}^{\mathsf{T}}\boldsymbol{y}_{ik},$$

is close to $\boldsymbol{B}$.

Synthetic estimation

A special case is the ratio estimator

$$\widehat{t}_i^{\,\text{syn,rat}} = \widehat{t}_y \times \frac{t_{xi}}{\widehat{t}_x}. \tag{11}$$

Another special case is the post-stratification estimator

$$\widehat{t}_i^{\,\text{syn,ps}} = \sum_{g=1}^{G} \frac{N_{ig}}{\widehat{N}_{.g}} \widehat{t}_{.g}. \tag{12}$$

These estimators have low variance since information from all domains is used, but the bias may be large.

Notice that, in contrast to the direct GREG estimators described previously, these forms are adjusting a global response estimate, using domain specific auxiliary information.

Example: Smoking by county in Washington State

For (8), suppose we have domain-specific totals on income $\boldsymbol{x}_i = [1, t_{xi}]^{\mathsf{T}}$, along with individual income levels in the sample; the latter are used to estimate the population regression coefficient (9).

To examine whether $\boldsymbol{B}_i$ is close to $\boldsymbol{B}$ we could calculate

$$\widehat{\boldsymbol{B}}_i = \left(\sum_{k \in s_i} w_k \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{x}_{ik} \right)^{-1} \sum_{k \in s_i} w_k \boldsymbol{x}_{ik}^{\mathsf{T}} \boldsymbol{y}_{ik},$$

and see how close these estimates are to $\widehat{\boldsymbol{B}}$ (though the former may have large uncertainty).

Example: Smoking by county in Washington State

For (11), we use the total incomes in domain i and the estimated income across the whole state $\widehat{t}_x = \sum_{k \in s} w_k x_k$, along with the estimated total smokers across the state $\widehat{t}_y = \sum_{k \in s} w_k y_k$.

For (12), we use the total population in domain i and stratum g, N_{ig} and the estimated stratum g population across the whole state $\widehat{N}_{.g} = \sum_{k \in s} w_k$, along with the estimated total stratum g population across the state $\widehat{t}_{.g} = \sum_{k \in s_g} w_k$.

Composite estimators

The direct estimator is approximately unbiased but will have large variance if n_i is small, while the synthetic estimator may have large bias but has small variance.

This suggests the composite estimator:

$$\overline{y}_i^{\text{comp}} = \phi_i \times \overline{y}_i^{\text{dir}} + (1 - \phi_i) \times \overline{y}_i^{\text{syn}}.$$

Rao and Molina (2015, Section 3.3) discusses how ϕ_i may be estimated, by attempting to minimize the MSE of $\overline{y}_i^{\text{comp}}$.

Next we will consider model-based approaches in which a formal method is used to balance using data from domain i, and the totality of data.