Bayesian SAE using Complex Survey Data
Lecture 7A: SAE

Jon Wakefield

Departments of Statistics and Biostatistics
University of Washington

Outline

Motivation

Motivating Example: Diabetes in King County

Arises out of a joint project between Laina Mercer/Jon Wakefield and Seattle and King County Public Health, which lead to the work reported in Song *et al.* (2016).

Aim we will concentrate on here is to estimate the number of 18 years or older individuals with diabetes, by health reporting areas (HRAs) in King County in 2011.

HRAs are city-based sub-county areas with a total of 48 HRAs in King County. Some of these are as are a single city, some are a group of smaller cities, and some are unincorporated areas. Larger cities such as Seattle and Bellevue include more than one HRA.

Data are based on the question, “Has a doctor, nurse, or other health professional ever told you that you had diabetes?”, in 2011.

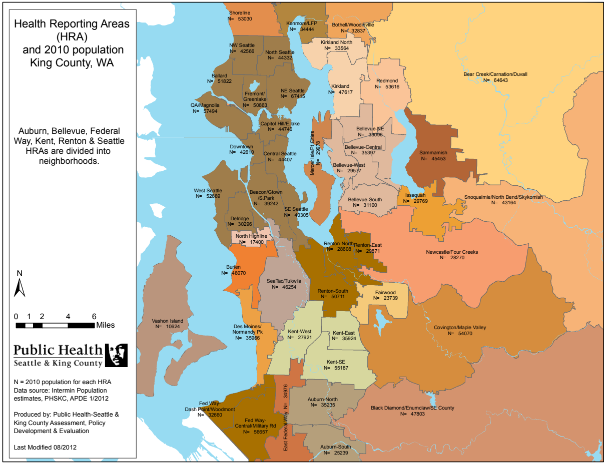


Figure 1: Health reporting areas (HRAs) in King County.

Motivating BRFSS Example

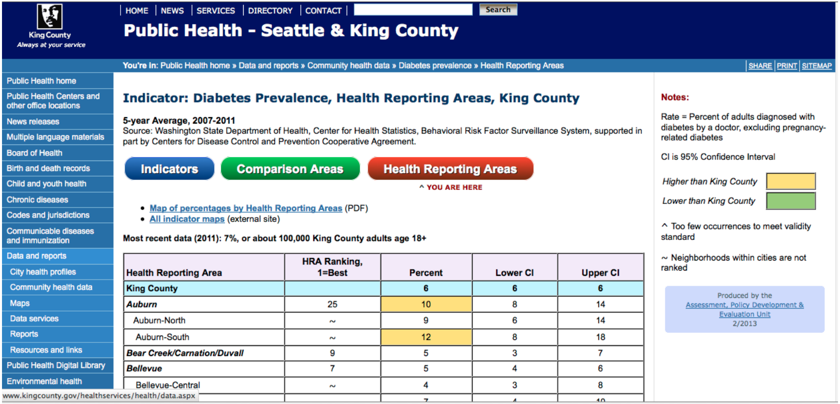
Estimates are used for a variety of purposes including summarization for the local communities and assessment of health needs.

Analysis and dissemination of place-based disparities is of great importance to allow efficient targeting of place-based interventions.

Because of its demographics, King County looks good compared to other areas in the U.S., but some of its disparities are among the largest of major metro areas.

Estimation is based on Behavioral Risk Factor Surveillance System (BRFSS) data.

The BRFSS is an annual telephone health survey conducted by the Centers for Disease Control and Prevention (CDC) that tracks health conditions and risk behaviors in the United States and its territories since 1984.



HOME | NEWS | SERVICES | DIRECTORY | CONTACT | Search

King County
Always at your service

Public Health - Seattle & King County

You're In: Public Health home » Data and reports » Community health data » Diabetes prevalence » Health Reporting Areas

SHARE | PRINT | SITEMAP

- Public Health home
- Public Health Centers and other office locations
- News releases
- Multiple language materials
- Board of Health
- Birth and death records
- Child and youth health
- Chronic diseases
- Codes and jurisdictions
- Communicable diseases and immunization
- Data and reports
  - City health profiles
  - Community health data
  - Maps
  - Data services
  - Reports
  - Resources and links
- Public Health Digital Library
- Environmental health

Indicator: Diabetes Prevalence, Health Reporting Areas, King County

5-year Average, 2007-2011
Source: Washington State Department of Health, Center for Health Statistics, Behavioral Risk Factor Surveillance System, supported in part by Centers for Disease Control and Prevention Cooperative Agreement.

Indicators | Comparison Areas | Health Reporting Areas

^ YOU ARE HERE

- Map of percentages by Health Reporting Areas (PDF)
- All indicator maps (external site)

Most recent data (2011): 7%, or about 100,000 King County adults age 18+

Health Reporting Area	HRA Ranking, 1=Best	Percent	Lower CI	Upper CI
King County		6	6	6
Auburn	25	10	8	14
Auburn-North	~	9	6	14
Auburn-South	~	12	8	18
Bear Creek/Carnation/Duvall	9	5	3	7
Bellevue	7	5	4	6
Bellevue-Central	~	4	3	8

Notes:

Rate = Percent of adults diagnosed with diabetes by a doctor, excluding pregnancy-related diabetes

CI is 95% Confidence Interval

Higher than King County

Lower than King County

^ Too few occurrences to meet validity standard

~ Neighborhoods within cities are not ranked

Produced by the Assessment, Policy Development & Evaluation Unit 2/2013

www.kingcounty.gov/healthservices/health/data.aspx

Figure 2: Public Health: Seattle and King County website.

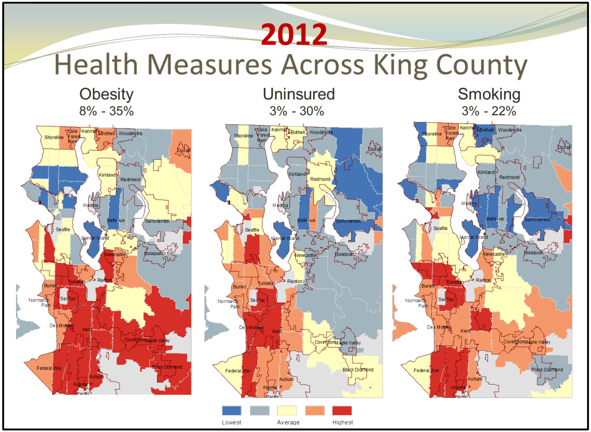


Figure 3: Summaries from Public Health: Seattle King County.

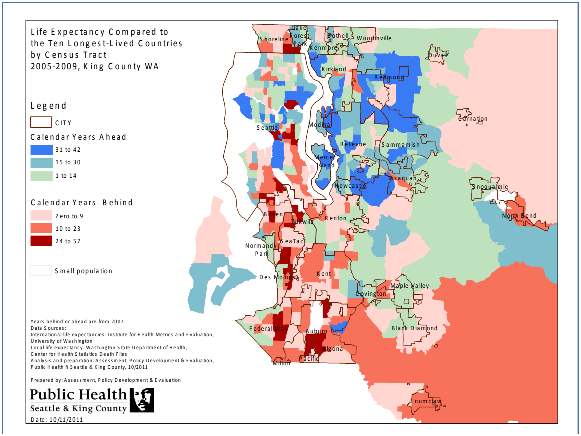


Figure 4: Summaries from Public Health: Seattle King County.

Motivating BRFSS Example

The BRFSS sampling scheme is complex: it uses a disproportionate stratified sampling scheme.

The `Sample Wt`, is calculated as the product of four terms

$$\texttt{Sample Wt} = \texttt{Strat Wt} \times \frac{1}{\texttt{No Telephones}} \times \texttt{No Adults} \times \texttt{Post Strat Wt}$$

where `Strat Wt` is the inverse probability of a “likely” or “unlikely” stratum being selected (stratification based on county and “phone likelihood”).

Table 1: Summary statistics for population data, and 2011 King County BRFSS diabetes data, across health reporting areas.

	Mean	*Std. Dev.*	*Median*	*Min*	*Max*	*Total*
Population (>18)	31,619	10,107	30,579	8,556	56,755	1,517,712
Sample Sizes	62.9	24.3	56.5	20	124	3,020
Diabetes Cases	6.3	3.1	6.3	1	15	302
Sample Weights	494.3	626.7	280.4	48.0	5,461	1,491,880

Motivating BRFSS Example

A total of $3,020$ individuals answered the diabetes question.

About 35% of the areas have sample sizes less than 50 (CDC recommended cut-off), so that the diabetes prevalence estimates are unstable in these areas.

We would like to use the totality of the data to aid in estimation in the data sparse areas.

The variability in the weights is high, from 48 to 5,461, with mean 494.

The coefficient of variation (CV) of the weights is 1.27.

Therefore, the inefficiency of using the sample weights under the assumption that unweighted mean is unbiased is about 62%, calculated as $\text{CV}^2/(\text{CV}^2 + 1)$ (Korn and Graubard, 1999).

BRFSS Sample Size by HRA

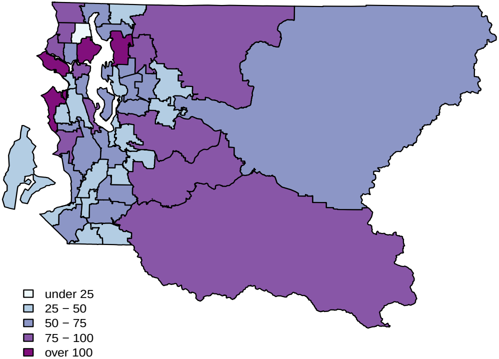


Figure 5: Sample sizes across 48 HRAs in 2011.

Observed prevalence by HRA

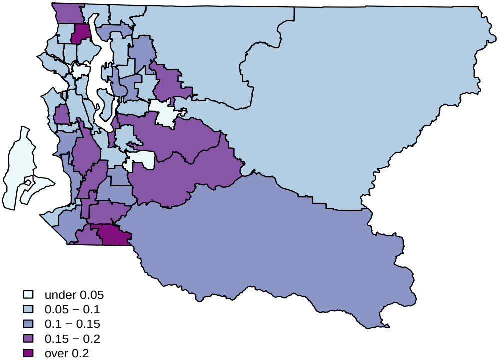


Figure 6: Diabetes prevalence by HRAs in 2011: crude proportions.

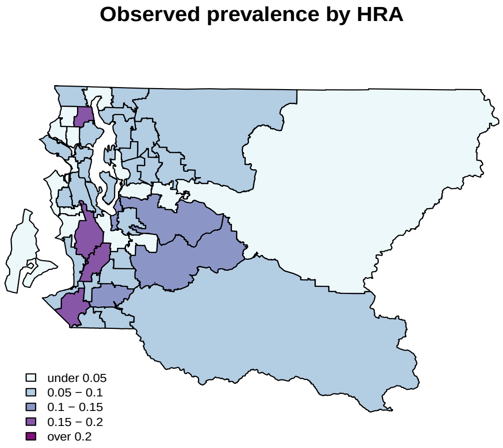


Figure 7: Diabetes prevalence by HRAs in 2011: Horvitz-Thompson weighted estimator.

Weights

More on Weighting

We have N_i individuals in area i and the indices of those selected in a sample of size n_i is denoted S_i.

The weights are often formed via

$$w_{ik} = w_{ik}^d \times w_{ik}^p \tag{1}$$

where w_{ik}^d is the design weight and w_{ik}^p is the post-stratification weight.

For the design weights

$$w_{ik}^d = \frac{1}{\pi_{ik}}$$

where π_{ik} is the probability of selection.

There may also be an additional adjustment to the weights to attempt to account for non-response.

Weighting

If N_i is not known it may be estimated by

$$\widehat{N}_i = \sum_{k \in S_i} w_{ik}^d$$

is an estimate of the total population in area i, in line with interpreting w_{ik}^d as the number of individuals that this individual represents.

Note that,

$$E[\widehat{N}_i] = \sum_{k=1}^{N_i} \mathrm{E}[I_{ik}]\pi_{ik}^{-1} = N_i,$$

so that this estimator is unbiased.

Post-stratification, as the name suggests, adjusts the weights after sampling, so that population totals in a set of stratum (e.g., age/gender) are recovered.

Post-stratification and Raking

If the post-stratification groups are indexed by j the weights are

$$w_{ik}^{p} = \frac{N_{j(k)}}{\widehat{N}_{j(k)}}$$

where $j(k)$ indicates the group to which individual k belongs, N_j are the known totals and $N_{j(k)} = \sum_{k \in S_j} w_{ik}^{d}$. This procedure adjusts the weights so that the known totals are recovered.

Previously in BRFSS in King County, post-stratification was used based only on age and gender.

Raking now used for BRFSS, adjusting for more factors: age, gender, race/ethnicity, marital status, education, owner/renter status, and cell phone/landline status).

Cannot exactly match all cross-classified tables of counts, so instead lower dimensional margins are controlled using a procedure known as iterative proportional fitting.

Modeling for Survey Data

Overview of Models For Binary Responses

- Binomial sampling model: only strictly valid if no stratified sampling and no cluster sampling.
- Direct estimates at the area level.
- Smoothed direct estimates at the area level, modeling the logit of the direct estimates of the probabilities.
- Binomial GLMM at the area level: only strictly valid if no stratified sampling and no cluster sampling.
- Binomial model for responses within each cluster with
  - strata fixed effects,
  - cluster random effects,
  - IID random effects at the area level
  - spatial random effects at the area level (via an ICAR model).
- Binomial model for responses within each cluster with
  - strata fixed effects,
  - IID cluster random effects,
  - IID household effects?
  - spatial random effects at the cluster level (via a Gaussian process model).

Smoothed Direct Estimation

We again use the model:

"Data" Model:

$$\widehat{\theta}_i \sim \mathrm{N}(\theta_i, V_i),$$

where V_i is known variance.

Prior Model:

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

with

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

BRFSS Example

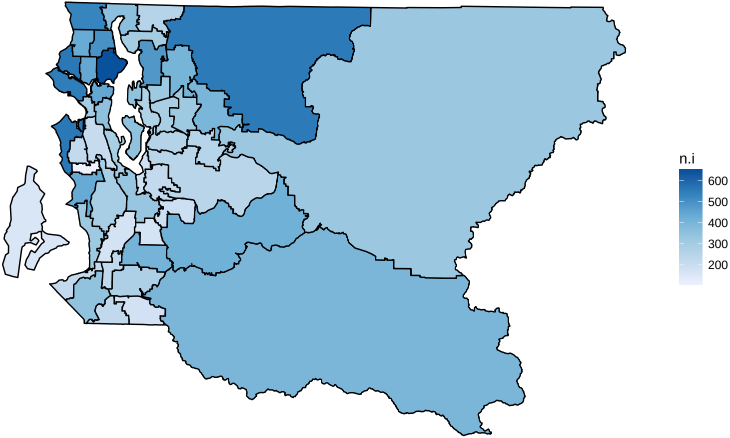


Figure 8: Sample sizes across HRAs.

BRFSS Example

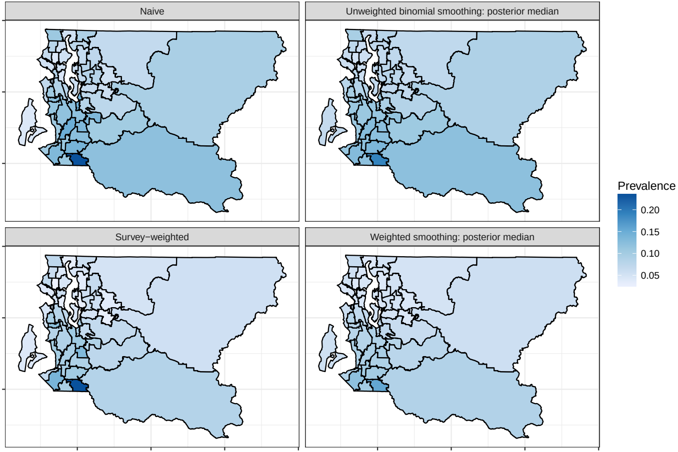


Figure 9: Diabetes prevalence estimates under different models.

BRFSS Example

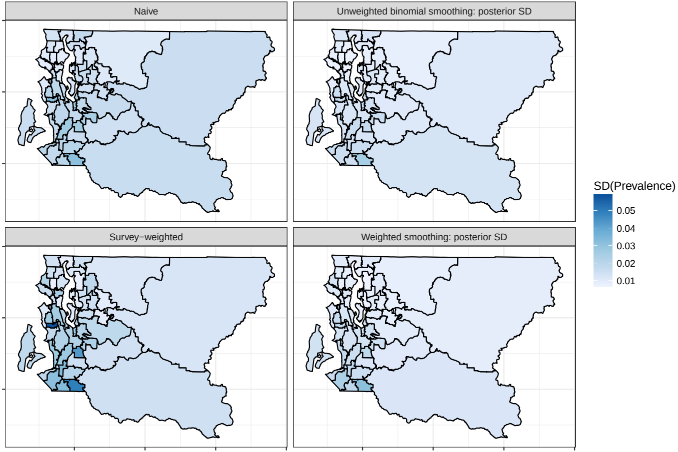


Figure 10: Comparison of uncertainty estimates under different models.

BRFSS Example

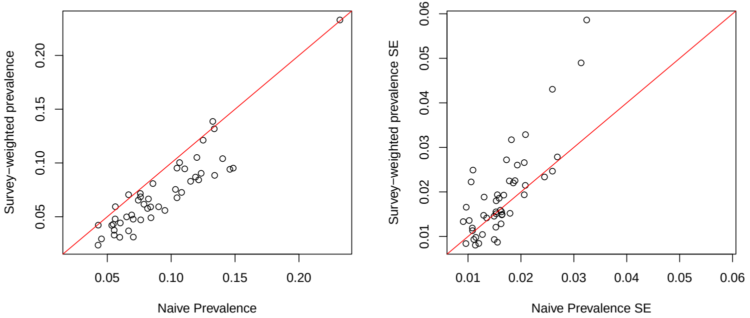


Figure 11: Diabetes prevalence uncertainty estimates under different models.

BRFSS Example

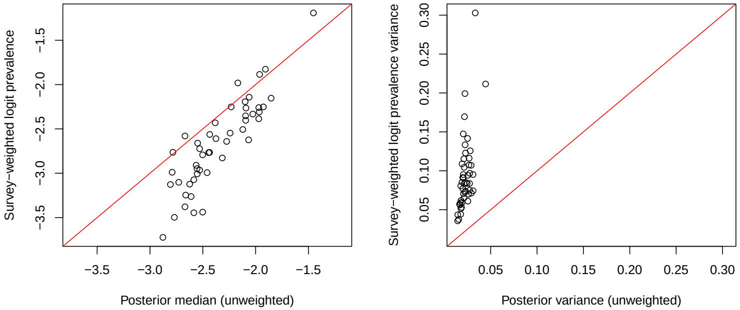


Figure 12: Diabetes prevalence uncertainty estimates under different models.

Discussion

Phone list strata not known for all population in BRFSS, so model-based more difficult.

References

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics – Theory and Methods*, **35**, 439–460.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279–292.

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.

Congdon, P. and Lloyd, P. (2010). Estimating small area diabetes prevalence in the US using the behavioral risk factor surveillance system. *Journal of Data Science*, **8**, 235–252.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153–164.

Ghitza, Y. and Gelman, A. (2013). Deep interactions with mrp: Election turnout and voting patterns among small electoral subgroups. *American Journal of Political Science*, **57**, 762–776.

Korn, E. and Graubard, B. (1999). *Analysis of Health Surveys*. John Wiley and Sons, New York.

Longford, N. (1996). Model-based variance estimation in surveys with stratified clustered design. *Australian Journal of Statistics*, **38**, 333–352.

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.

Pereira, L. N. and Coelho, P. (2010). Small area estimation of mean price of habitation transaction using timeseries and cross-sectional area-level models. *Journal of Applied Statistics*, **37**, 651–666.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, **60**, 23–40.

Potthoff, R., Woodbury, M., and Manton, K. (1992). ”Equivalent sample size” and ”equivalent degrees of freedom” refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, **87**, 383–396.

Pratesi, M. and Salvati, N. (2008). Small area estimation: the EBLUP estimator based on spatially correlated random area effects. *Statistical Methods and Applications*, **17**, 113–141.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society, Series A*, **169**, 805–827.

Raghunathan, T., Xie, D., Schenker, N., Parsons, V., Davis, W., Dood, K., and Feuer, E. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factos and screening. *Journal of the American Statistical Association*, **102**, 474–486.

Singh, B., Shukla, G., and Kundu, D. (2005). Spatio-temporal models in small-area estimation. *Survey Methodology*, **31**, 183–195.

Skinner, C. (1989). Domain means, regression and multivariate analysis. In C. Skinner, D. Holt, and T. Smith, editors, *Analysis of Complex Surveys*, pages 59–87. Wiley, Chichester.

Song, L., Mercer, L., Wakefield, J., Laurent, A., and Solet, D. (2016). Peer reviewed: Using small-area estimation to calculate the prevalence of smoking by subcounty geographic areas in king county, washington, behavioral risk factor surveillance system, 2009–2013. *Preventing chronic disease*, **13**.

Zheng, H. and Little, R. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, **19**, 99–17.

Zheng, H. and Little, R. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, **21**, 1–20.

Technical Appendix: R Packages

Appendix: R Packages

- `sae`, by Molina and Marhuenda
  - area-levels (Fay-Herriott (FH), FH with spatial correlation, FH with spatio-temporal correlation) and unit-level models (BHF)
  - estimators: direct Horvitz-Thompson under general sampling designs, post-stratified synthetic estimator and composite estimator
  - fitting and estimation (frequentist) methods: FH, ML, REML, bootstrap
- `rsae`, by Schoch
  - area-levels and unit-level models
  - fitting and estimation (frequentist) methods: ML, Huber-type M-estimation
- `JoSae`, by Breidenbach
  - unit-level models
  - estimators: EBLUP (BHF1988) and GREG (Sarndal 1984)
- `SUMMER` by Martin, Zhang, Wakefield, Clark, Mercer
  - U5MR models using method of Mercer *et al.* (2015).

R Packages

- `hbsae`, by Boonstra
  - area-levels and unit-level models
  - fitting and estimation (frequentist and Bayesian) methods: REML, HB (based on MCMC)
- `mme`, by Lopez-Vizcaino et. al.
  - area-levels multinomial models (area random effects and time random effects)
  - fitting and estimation (frequentist) methods: analytical (PQL and REML) and bootstrap
- `saery`, by Esteban et al.
  - area-level model Rao-Yu 1994
  - fitting and estimation (frequentist) methods: REML
- `sae2`, by Fay and Diallo
  - time series area-level models, Rao-Yu 1994 and extensions
  - fitting and estimation (frequentist) methods: ML and REML

R Packages

- `BayesSAE`, by Shi and Zhang
  - area-levels models: FH and extensions (You-Chapman, spatial models and more)
  - fitting and estimation (Bayesian) methods: HB (based on MCMC)
- `saeSim`, by Warnholz and Schmid
  - useful tools to simulate data for sae studies
- `small area`, by Nandy
  - area-level model (FH)
  - fitting and estimation (frequentist) methods: FH, Prasad and Rao, REML

Note that only `hbsae` and `BayesSAE` use Bayesian methods for the estimation, both use MCMC.

Let $Z_i = \sin^{-1}\sqrt{\widehat{P}_i}$ represent the variance stabilizing transformation of P_i.

Then define the likelihood as

$$Z_i|\lambda_i \sim \mathrm{N}\left(\lambda_i, \frac{1}{4\widetilde{m}_i}\right),$$

where $\lambda_i = \sin^{-1}\sqrt{P_i}$.

Note that $0 \le \lambda_i \le \pi/2$ which is not ideal.

An obvious second stage model is

$$\lambda_i|\alpha, \beta, \tau^2 \sim_{ind} \mathrm{N}(\alpha + \beta x_i, \tau^2).$$

A full Bayes approach would add a third stage with priors for α, β, τ^2.

We could also add spatial effects to this model at the second stage.

Implementation for this model is awkward because of the restricted range for λ_i.

An alternative formulation for binary outcomes is due to Chen *et al.* (2014); Mercer *et al.* (2014).

Define the effective sample size as before and the effective number of responders as

$$\widetilde{y}_i = \widetilde{m}_i \times \widehat{P}_i.$$

Likelihood is $\widetilde{y}_i | P_i \sim \text{Binomial}(\widetilde{m}_i, P_i)$.

The usual hierarchical models can then be applied at the second stage.

An obvious choice is

$$\log\left(\frac{P_i}{1-P_i}\right) = \alpha + \boldsymbol{x}_i\boldsymbol{\beta} + V_i + U_i.$$

Hierarchical modeling: notes

Inference may be carried out via likelihood or Bayes, with the latter placing priors on $\boldsymbol{\beta}, \sigma_{\varepsilon}^2, \sigma_{\epsilon}^2$.

If a likelihood approach is taken, the random effect estimates $\widehat{\varepsilon}_i$, are obtained as best linear unbiased predictors (BLUPs).

If there are no data in particular areas we can still make predictions, if we assume the model holds for all areas.

Note: can add area level covariates to model.

Hierarchical Modeling of Survey Sample Data

A Horvitz-Thompson weighted estimator of the log-likelihood for binary data is

$$\sum_{i=1}^{n} \sum_{k=1}^{m_i} w_{ik} \left\{ y_{ik} \log P_i + (1 - y_{ik}) \log(1 - P_i) \right\}. \tag{2}$$

(Binder, 1983) where y_{ik} is the binary outcome on person k in area i, with associated weight w_{ik}.

Method known as pseudo-likelihood.

Pseudo-likelihood (Skinner, 1989; Pfeffermann *et al.*, 1998) has been used within a hierarchical modeling framework with the scaling of the weights being a major issue (Potthoff *et al.*, 1992; Longford, 1996; Asparouhov, 2006; Rabe-Hesketh and Skrondal, 2006).

Congdon and Lloyd (2010) use a weighted likelihood to analyze BRFSS data and introduce residual spatial random effects at the state level.

Further References

Although there is a huge literature on small area estimation the spatial smoothing of survey data with complex weights is not routinely carried out.

In terms of spatial smoothing techniques, a number of authors allow for spatial correlation between areas, see for example Singh *et al.* (2005), Pratesi and Salvati (2008) and Pereira and Coelho (2010).

These models are subject to bias, however, since they do not adjust for the sampling scheme.

We shortly describe a relatively new approach based on the concept of “effective sample size” and “effective number of cases”.

A related Bayesian model has recently been suggested by Ghitza and Gelman (2013), while a quite different approach, based on a penalized spline model, is described in Zheng and Little (2003) and Zheng and Little (2005).

Hierarchical Modeling of Survey Sample Data

We describe an approach described in Raghunathan *et al.* (2007).

These authors consider the estimation of a population proportion across areas i, P_i. Let

$$\widehat{P}_i = \frac{1}{N_i} \sum_{k=1}^{m_i} w_{ik} Y_{ik}$$

represent the weighted prevalence estimate in area i with associated sample size m_i and design-based variance estimate v_i, for example (**??**).

Let P_i represent the true population proportion in area i.

If a SRS were taken the variance would be $P_i(1 - P_i)/m_i$.