Accounting for Complex Sample Designs via Finite Normal Mixture Models

Michael Elliott[1]

[1]University of Michigan School of Public Health

August 2009

Talk Outline

1. Background
   - Design-Based Inference
   - Bayesian Model-Based Inference
   - Accommodating Survey Weights in a Model
2. Finite Normal Mixture Models
   - Accommodating Sampling Weights in Mixture Models
3. Simulation
4. Discussion and Extensions

Design-Based Inference

- Randomization or "design-based" inference is standard for sample survey data.
- Treat population values $\mathbf{Y} = (Y_1, ..., Y_N)$ as *fixed*, and sampling indicators $\mathbf{I} = (I_1, ..., I_N)$ as *random*.
- Goal is to make inference about a population quantity $Q(\mathbf{Y})$.

Design-Based Inference

- Randomization or "design-based" inference is standard for sample survey data.
- Treat population values $\mathbf{Y} = (Y_1, ..., Y_N)$ as *fixed*, and sampling indicators $\mathbf{I} = (I_1, ..., I_N)$ as *random*.
- Goal is to make inference about a population quantity $Q(\mathbf{Y})$.
- Consider estimator $\hat{q}(\mathbf{y}, \mathbf{I})$ where

$$E_{\mathbf{I}|\mathbf{Y}}(\hat{q}(\mathbf{y}, \mathbf{I})) \approx Q(\mathbf{Y})$$

and variance estimator of $\hat{q}(\mathbf{y}, \mathbf{I})$ $\hat{v}(\mathbf{Y}_{inc}, \mathbf{I})$ where

$$E_{\mathbf{I}|\mathbf{Y}}(\hat{v}(\mathbf{y}, \mathbf{I})) \approx Var_{\mathbf{I}|\mathbf{Y}}(\hat{q}(\mathbf{y}, \mathbf{I}))$$

(Hansen and Hurwitz 1943; Kish 1965; Cochran 1977.)

Bayesian Survey Inference

Focus on inference about $Q(\mathbf{Y})$ based on $p(\mathbf{Y}_{nobs} \mid \mathbf{y})$:

$$p(\mathbf{Y}_{nobs} \mid \mathbf{y}) = \frac{p(\mathbf{Y})}{p(\mathbf{y})} =$$

$$\frac{\int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{p(\mathbf{y})} =$$

$$\frac{\int p(\mathbf{Y}_{nobs} \mid \mathbf{y}, \boldsymbol{\theta}) p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{p(\mathbf{y})} =$$

$$\int p(\mathbf{Y}_{nobs} \mid \mathbf{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathbf{y}) d\boldsymbol{\theta}$$

(Ericson 1969; Scott 1977; Rubin 1987).

Design Inference vs. Bayesian Inference

Randomization approach has substantial advantages.

- **Y** treated as fixed → no need for distributional assumptions.

Design Inference vs. Bayesian Inference

Randomization approach has substantial advantages.

- $\mathbf{Y}$ treated as fixed → no need for distributional assumptions.
- In scientific surveys the distribution of $\mathbf{I}$ is (largely) under the control of the investigator.

Design Inference vs. Bayesian Inference

Randomization approach has substantial advantages.

- **Y** treated as fixed → no need for distributional assumptions.
- In scientific surveys the distribution of **I** is (largely) under the control of the investigator.
- “Automatically” account for sample design in inference.

Design Inference vs. Bayesian Inference

Randomization approach does not always work well.

- Inefficient (Basu 1971)

Design Inference vs. Bayesian Inference

Randomization approach does not always work well.

- Inefficient (Basu 1971)
- Small-area estimation (Ghosh and Lahiri 1988)
- Non-response (Little and Rubin 2002)

Design Inference vs. Bayesian Inference

Randomization approach does not always work well.

- Inefficient (Basu 1971)
- Small-area estimation (Ghosh and Lahiri 1988)
- Non-response (Little and Rubin 2002)
- Lack of consistent reference distribution (Little 2004)

Design Inference vs. Bayesian Inference

Bayesian approach avoids “inferential schizophrenia” (Little 2004)

- Doesn’t rely on asymptotics
- Focus on prediction of unsampled elements

Design Inference vs. Bayesian Inference

Bayesian approach avoids "inferential schizophrenia" (Little 2004)

- Doesn't rely on asymptotics
- Focus on prediction of unsampled elements
- Requires that sampling indicator $\mathbf{I}$ need not be modeled, which in turn requires
  - (1) $p(\mathbf{I} \mid \mathbf{Y}) = p(\mathbf{I} \mid \mathbf{Y}_{obs})$ and
  - (2) $p(\mathbf{Y}_{nob} \mid \mathbf{Y}_{obs}, \mathbf{I}, \boldsymbol{\theta}) = p(\mathbf{Y}_{nob} \mid \mathbf{Y}_{obs}, \boldsymbol{\theta})$.
- (1) "ignorable" or "noninformative" sampling, satisfied in most probability samples (Rubin (1987)).
- (2) requires data model $p(\mathbf{Y} \mid \boldsymbol{\theta})$ attentive to design features and robust enough to sufficiently capture all aspects of the distribution of relevant to $Q(\mathbf{Y})$.

Accommodating Survey Weights in a Model

Stratify data by probabilities of inclusion $h = 1, ..., H$ and allow interaction between model quantities of interest and probabilities of inclusion

$$y_{ih} \sim N(\mu_h, \sigma^2)$$

$$\overline{Y} \mid \mathbf{y} \sim N(N^{-1} \sum_h \{n_h \overline{y_h} + (N_h - n_h)\hat{\bar{y}}_h\}, (1 - n/N)\sigma^2/n), \hat{\bar{y}}_h = E(\mu_h \mid \mathbf{y})$$

Accommodating Survey Weights in a Model

Stratify data by probabilities of inclusion $h = 1, ..., H$ and allow interaction between model quantities of interest and probabilities of inclusion

$$y_{ih} \sim N(\mu_h, \sigma^2)$$

$$\overline{Y} \mid \mathbf{y} \sim N(N^{-1} \sum_h \{n_h \overline{y_h} + (N_h - n_h)\hat{\overline{y}}_h\}, (1 - n/N)\sigma^2/n), \hat{\overline{y}}_h = E(\mu_h \mid \mathbf{y})$$

- Flat prior on $\mu_h \rightarrow \hat{\overline{y}}_h = \overline{y}_h$ recovers fully-weighted estimator.

Accommodating Survey Weights in a Model

Stratify data by probabilities of inclusion $h = 1, \ldots, H$ and allow interaction between model quantities of interest and probabilities of inclusion

$$y_{ih} \sim N(\mu_h, \sigma^2)$$

$$\overline{Y} \mid \mathbf{y} \sim N(N^{-1} \sum_h \{n_h \overline{y_h} + (N_h - n_h)\hat{\bar{y}}_h\}, (1 - n/N)\sigma^2/n), \hat{\bar{y}}_h = E(\mu_h \mid \mathbf{y})$$

- Flat prior on $\mu_h \rightarrow \hat{\bar{y}}_h = \bar{y}_h$ recovers fully-weighted estimator.
- Degenerate prior on μ_h at $\mu \rightarrow \hat{\bar{y}}_h = \bar{y}$ recovers unweighted estimator.

Accommodating Survey Weights in a Model

Stratify data by probabilities of inclusion $h = 1, \ldots, H$ and allow interaction between model quantities of interest and probabilities of inclusion

$$y_{ih} \sim N(\mu_h, \sigma^2)$$

$$\overline{Y} \mid \mathbf{y} \sim N(N^{-1}\sum_h \{n_h\overline{y_h} + (N_h - n_h)\hat{\bar{y}}_h\}, (1 - n/N)\sigma^2/n), \hat{\bar{y}}_h = E(\mu_h \mid \mathbf{y})$$

- Flat prior on $\mu_h \rightarrow \hat{\bar{y}}_h = \bar{y}_h$ recovers fully-weighted estimator.
- Degenerate prior on μ_h at $\mu \rightarrow \hat{\bar{y}}_h = \bar{y}$ recovers unweighted estimator.
- Assigning a proper prior $\mu_h \sim N(\mu, \tau^2)$ (Holt and Smith 1979) compromises between fully-weighted and unweighted estimator: $\hat{\bar{y}}_h = w_h \bar{y}_h + (1 - w_h)\tilde{y}$, $w_h = \frac{n_h\tau^2}{n_h\tau^2 + \sigma^2}$, $\tilde{y} = \left(\sum_h \frac{n_h}{n_h\tau^2+\sigma^2}\right)^{-1} \sum_h \frac{n_h}{n_h\tau^2+\sigma^2}\bar{y}_h$.

Accommodating Survey Weights in a Model

Elliott and Little (2000) extend to consider $\boldsymbol{\mu} \sim N(f(h,\beta),\Sigma)$:

- Adding structure to the mean and variance increases robustness of estimation of population mean when stratum mean strongly associated with probability of selection, though efficiency gains over design-based estimator of $\overline{Y}$ when stratum means are weakly associated with probability of selection are reduced.
- A Bayesian smoothing spline estimator of the mean is quite robust but can still can yield efficiency gains.

Accommodating Survey Weights in a Model

Developing models to accommodate survey weights for more complex population quantities such as population regression parameters more challenging.

Elliott (2007) extends weight stratum models to linear and generalized linear regression models by allowing for interactions between weight strata and regression parameters.

- Efficiency gains are possible over design-based regression estimators
- Proliferation of parameters can make practical implementation difficult, if number of covariates large and sample size modest.

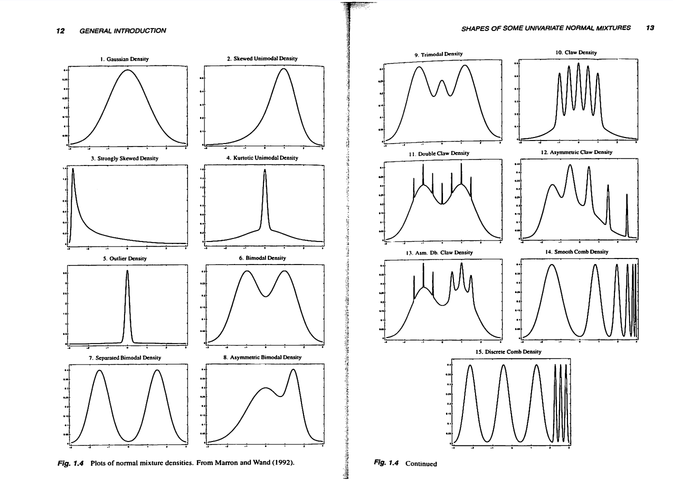
Finite Normal Mixture Models

A simple finite normal mixture model without covariates:

$$Y_i \mid C_i = c, \mu_c, \sigma_c^2 \sim N(\mu_c, \sigma_c^2),\ C = 1, \ldots, K$$

$$C_i = c \mid \pi_1, \ldots, \pi_K \sim MULTI(1; \pi_1, \ldots, \pi_K)$$

Finite Normal Mixture Models



12 GENERAL INTRODUCTION

1. Gaussian Density

2. Skewed Unimodal Density

3. Strongly Skewed Density

4. Kurtotic Unimodal Density

5. Outlier Density

6. Bimodal Density

7. Separated Bimodal Density

8. Asymmetric Bimodal Density

Fig. 1.4 Plots of normal mixture densities. From Marron and Wand (1992).

SHAPES OF SOME UNIVARIATE NORMAL MIXTURES 13

9. Trimodal Density

10. Claw Density

11. Double Claw Density

12. Asymmetric Claw Density

13. Asm. Db. Claw Density

14. Smooth Comb Density

15. Discrete Comb Density

Fig. 1.4 Continued

Application to Complex Sample Data

- Maintain robustness of design-based approach?
  - Use of models that include a large number of classes to model highly non-normal data.
- Increased efficiency of model-based approach?
  - If the data suggest that a small number (or single) class of normal data is sufficient.

Normal Regression Mixture Model for Complex Sample Design Data

$$Y_i \mid \mathbf{x}_i, C_i = c, \boldsymbol{\beta}_c, \sigma_c^2 \sim N(\mathbf{x}_i' \boldsymbol{\beta}_c, \sigma_c^2),\ C = 1, ..., K$$

$$C_i = c \mid \boldsymbol{\alpha}, \boldsymbol{\gamma}, \pi_i \sim MULTI(1; p_1, ..., p_K), \eta_{ij} = \Phi(\gamma_j - f(\pi_i, \boldsymbol{\alpha})) \text{ for}$$

$$\eta_{ij} = \sum_{k=1}^{j} p_k, j = 1, ..., K-1$$

where $\gamma_1 = 0$ to avoiding aliasing with the α parameters.

- Accounts for regression model misspecification and skewness and overdispersion in the residual errors term
- Fits simple, highly efficient models when the data allow.
- $f(\pi_i, \boldsymbol{\alpha})$ could be simple parametric form (e.g., linear in π), or non-parametric (e.g., linear P-spline).

Normal Mixture Model Priors

To ensure a proper posterior, we utilize conjugate priors of the form

$$p(\boldsymbol{\beta}_c) \sim N(\boldsymbol{\beta}_0, \Sigma_0)$$

$$p(\sigma_c^2) \sim Inv - \chi^2(a, s)$$

$$p(\boldsymbol{\alpha}) \sim N(\boldsymbol{\alpha}_0, \Omega_0)$$

$$p(\gamma_j) \sim UNI(0, A)$$

By choosing relatively non-informative values for the prior parameters, we should be able to avoid influencing the results of the inference to an untoward degree.

Draws from $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}, \gamma \mid \mathbf{y})$ are obtained using a Gibbs sampling algorithm.

Normal Mixture Model Posterior Predictive Distribution

Using simulations $(\boldsymbol{\beta}^{rep}, (\sigma^2)^{rep}, \boldsymbol{\alpha}^{rep}, \gamma^{rep})$ from $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\alpha}, \gamma \mid \mathbf{y}, \mathbf{x})$, we obtain a simulation from $p(\mathbf{B} \mid \mathbf{y}, \mathbf{x})$ where $\mathbf{B}$ is the population slope $(\sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i=1}^{N} \mathbf{x}_i y_i$:

$$\hat{y}_i^{rep} = \sum_{c=1}^{K} \tilde{p}_{ic}^{rep} \mathbf{x}_i' \boldsymbol{\beta}_c^{rep}, \tilde{p}_{ic}^{rep} \sim p(p_{ic} \mid \mathbf{y})$$

$$B^{rep} = (\sum_{i=1}^{n} w_i \mathbf{x}_i \mathbf{x}_i')^{-1} \sum_{i=1}^{w} w_i \mathbf{x}_i \hat{y}_i^{rep}$$

Model Averaging

- We do not know "true" number of classes K.
- Use model averaging to obtain a single posterior distribution of the population quantity of interest:

$$p(\mathbf{B} \mid \mathbf{y}) = \sum_{K=1}^{L} p(\mathbf{B} \mid \mathbf{y}, M = K) p(M = K \mid \mathbf{y}).$$

- Use product space search method of Carlin and Chib (1995) to obtain simulations from $p(M = K \mid \mathbf{y})$.
  - Sample over M as well as the product of the parameter space across all L models.
  - Posit "pseudo-prior" $p(\boldsymbol{\theta}_K \mid M \neq K)$; then

$$p(\mathbf{y}, \boldsymbol{\theta}, M = J) = f(\mathbf{y} \mid \boldsymbol{\theta}_J, M = J) \left\{ \prod_{K=1}^{L} p(\boldsymbol{\theta}_K \mid M = J) \right\} P(M = J)$$

  - Use MCMC to obtain draws from $p(\boldsymbol{\theta}, M \mid \mathbf{y})$.

Simulation Model

$$Y_i \mid X_i, \sigma^2 \sim N(\alpha_0 + \sum_{h=1}^{10} \alpha_h (X_i - h)_+, \sigma^2),$$
$$X_i \sim UNI(0, 10),\ i = 1, \ldots, N = 20000.$$
$$P(I_i = 1 \mid H_i) = \pi_h \propto (1 + H_i) H_i$$
$$H_i = \lceil X_i \rceil$$

- Elements (Y_i, X_i) had $\approx$1/55th the selection probability when $0 \leq X_i \leq 1$ as when $9 \leq X_i \leq 10$.
- $n = 1000$ elements were sampled without replacement for each of 50 simulations.
- $\boldsymbol{\alpha}_C = (0, 0, 0, 0, .5, .5, 1, 1, 2, 2, 4)$, $\sigma^2 = 10, 1000, 100000$: bias important for σ^2 small.

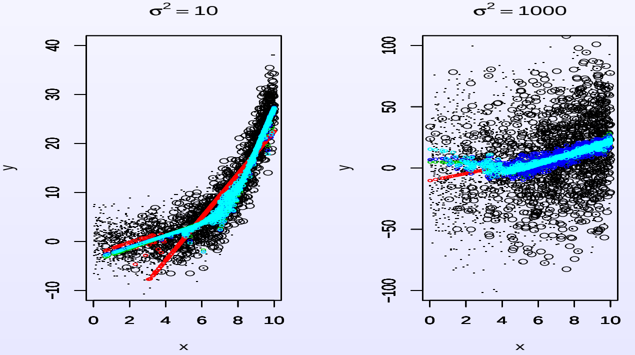
Estimation procedures

- Fully-weighted, unweighted, crude trimming weight (maximum normalized value of 3).
- 2- through 5-class mixture model; model averaged.
- Penalized probit model for the mixture class as a function of the case weights:

$$f(\pi_i, \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 w_i + \sum_{l=1}^{m} b_l (w_i - k_l)_+, \;\; b_l \overset{iid}{\sim} N(0, \tau^2)$$

  where $(x)_+ = x$ if $x > 0$ and 0 otherwise and k_l are fixed knots at the weighted deciles of the weights.
- $\beta_0 = \hat{\beta}$, $\Sigma_0 = n^{1.5} \hat{V}_\beta$, $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{V}_\beta = \hat{\sigma}^2 (X'X)^{-1}$ for $\hat{\sigma}^2 = (n-p)^{-1} \sum_i (y_i - x_i' \hat{\beta})^2$.
- $a = 1$, $s^2 = \hat{\sigma}^2$.
- $\alpha_0 = 0$, $\Omega_0 = diag(10^6)$, and $A = 10$.

Posterior predictive mean of y_i for each of the four mixture models considered (2=red, 3=green, 4=blue, 5=turquoise).



Simulation Results

	RMSE relative to FWT			True Coverage		
	Variance $\log_{10}$			Variance $\log_{10}$		
Estimator	1	3	5	1	3	5
UNWT	13.18	1.58	.55	0	8	96
FWT	1	1	1	90	94	96
TWT3	4.51	.78	.63	0	92	98
MWT2	2.56	1.59	.55	74	16	96
MWT3	1.40	.81	.68	52	92	98
MWT4	1.10	.89	.65	76	96	98
MWT5	1.15	.88	.78	70	92	98
MWT	1.17	.78	.71	70	96	98

Simulation Results

- Unweighted and trimmed estimator (TWT) behave poorly when σ^2 is small, but have better MSE properties than the fully weighted estimator and conservative coverage as variance increases.
- Fully weighted estimator is approximately design-unbiased; coverage is approximately correct when model misspecification is weak, but is somewhat anti-conservative when misspecification is clearly present.

Simulation Results

- At least 4 mixture components appear to be required.
- The model average estimator has RMSE approximately 15% greater than the fully weighted estimator when substantial model misspecification is present and 30% less when model misspecification is minor.
- Coverage is approximatly correct when model misspecification is minor but substantially below nominal when model misspecification is large.

Discussion

- If the goal is to obtain a single model average estimate, a method such as reversible jump method (Green 1995) may be preferable.
- Quantile estimation or quantile regression may yield greater dividends, since heteroscedasticity easily accounted for in mixture models.
- Valuable if partial covariate information available for entire population.

Extensions

- Generalized linear models may be accommodated by embedding normal model for the outcome in a latent variable context (e.g., probit modeling).
- t-mixture models (Lange et al. 1989; Liu and Rubin 1995) might better capture the presence of outliers with fewer mixture components.
- Recent advances in non-parametric Bayes (Dunson et al. 2007; MacEachern and Muller 1998) allow K to be data-dependent, which may achieve better robustness and efficiency.

Take-Home Message

Advances in statistical modeling during the past 10-15 years are beginning to allow development of models sufficiently robust to compete with design-based approaches.