Appendices for "A Bayesian Approach to Combining Information from a Census, a Coverage Measurement Survey and Demographic Analysis"

Michael R. Elliot and Roderick J.A. Little, University of Michigan

February 15, 2000

Contents

1 Summary

Capture-recapture methods and demographic analysis (DA) of data on births, deaths, and migration have have been used to estimate the undercount in the US Census (Robinson et al. 1993). Since 1970, the Census has been supplemented by a coverage measurement survey (CMS), a detailed independent enumeration of households in a probability sample of Census blocks conducted immediately after the actual Census. To combine the Census and CMS data, individuals in the Census who were imputed or for whom insufficient information existed to match with the CMS ("imputations") were removed from the Census total. In addition, persons in the Census listing who did not appear in the CMS were re-checked, and if found to be "erroneously enumerated" were also

Table 1: Observed data and associated underlying population parameters from Census and Coverage Measurement Survey: y^S_{k11} and y^S_{k01} are estimated counts of individuals in and out of the Census on the basis of the CMS follow-up; $z^S_{k1.}$ is the Census count, minus imputations and an estimate of erroneous enumerations; ψ^S_{kij} is the population that would reside in the ijth cell if the CMS had been a complete census. (stratum=k; gender=S).

		Observed Data			Underlying Parameters		
		CMS			CMS		
		In	Out		In	Out	
Census	In	y^S_{k11}		$z^S_{k1.}$	ψ^S_{k11}	ψ^S_{k10}	$\psi^S_{k1.}$
	Out	y^S_{k01}			ψ^S_{k01}	ψ^S_{k00}	$\psi^S_{k0.}$
		$y^S_{k.1}$			$\psi^S_{k.1}$	$\psi^S_{k.0}$	$\psi^S_{k..}$

removed from the Census. Similarly, persons in the CMS were cross-checked against Census records within the sample Census blocks and a set of surrounding blocks and assigned to either an In-Census or Out-Census category. After inflating the counts in the sampled blocks by the inverse of their probability of selection, a 2×2 table (In-Out Census; In-Out CMS) was formed for males and females ($S = M, F$) in each of K post-strata within age-by-race groupings (see Table 1) (Hogan 1992).

Table 1 also gives the "true" but unknown population counts in the kth poststratum for gender S that would be obtained if the CMS had itself been a census and if those missed in both the Census and CMS were known. These population quantities $\psi_k = \{\psi^S_{kij} : i = 0, 1; j = 0, 1; S = M, F\}$, where i is the Census enumeration status and j the CMS enumeration status in the kth poststratum (1 if included, 0 otherwise), are considered unknown parameters. To estimate these parameters, and to correct for the apparent undercount of men relative to women suggested by demographic sex ratio data (Robinson 1995), several models have been suggested that attribute the low observed male-female ratio to an undercount of males rather than an overcount of females (Bell 1993). The estimated number of males in the population is increased to match the overall male-female ratio ($\rho = \psi^M_{...}/\psi^F_{...}$) from DA. The additional males are distributed over the post-strata using a method based on the assumptions of the model. In particular the fixed relative-risk (FRR) model assumes a constant relative risk for enumeration in the Census and CMS for males and independence for females:

$$\text{FRR: } \gamma^M_k = \frac{\psi^M_{k11}/\psi^M_{k1.}}{\psi^M_{k01}/\psi^M_{k0.}} = \gamma^M \text{ and } \theta^F_k = 1 \text{ for all } k. \tag{1}$$

Although several behavior models, including the FRR model, have been suggested that provide saturated estimates of model fit, we select this model on the basis of its plausibility, stability relative to the other models, and inclusion of the independence model (which does not force the population estimates to equal a "known" sex ratio) as a special case. To fit (1), we reparameterized the eight population counts in poststratum k, $\psi_k = \{\psi^S_{kij} : i = 0, 1; j = 0, 1; S = M, F\}$, as $\psi^*_k = (\psi_{k..}, \rho_k, \delta^M_k, \delta^F_k, \phi^M_k, \phi^F_k, \gamma^M_k, \gamma^F_k)$ where:

1. $\psi_{k..}$, the total population count in poststratum k.
2. $\rho_k = (\psi^M_{k..})/(\psi^F_{k..})$, the sex ratio ($\psi^M_{k..} + \psi^F_{k..} = \psi_{k..}$)
3. $\delta^S_k = (\psi^S_{k1.})/(\psi^S_{k..})$, the Census undercount proportion for sex S
4. $\phi^S_k = (\psi^S_{k11})/(\psi^S_{k1.})$, the proportion of Census cases enumerated in the CMS for sex S
5. $\gamma^S_k = \frac{\psi^S_{k11}/\psi^S_{k1.}}{\psi^S_{k01}/\psi^S_{k0.}}$, the relative proportion of Census and non-Census cases enumerated in the CMS for sex S

We then selected the following independent priors for each parameter:

- $p(\psi_{k..}) \propto 1$,
- $\rho_k \sim N(\rho, \sigma^2)$ subject to the constraint that $\sum_k w_k \rho_k = \rho$ where $w_k = (\psi_{k..})/(\sum_k \psi_{k..})$ and ρ is the DA-estimated nationwide sex ratio.
- $\delta_k^S \sim BETA(a^S, b^S)$ and $\phi_k^S \sim BETA(c^S, d^S)$,
- $\gamma_k^M = \gamma^M \sim GAMMA(\alpha, \beta)$ for all k; $\gamma_k^F = 1$ for all k.

In addition, we assume that $y_{k11}^S \mid \psi_{k11}^S \sim N(\psi_{k11}^S, (\tau_{k11}^S)^2)$, $y_{k01}^S \mid \psi_{k01}^S \sim N(\psi_{k01}^S, (\tau_{k01}^S)^2)$, $z_{k1.}^S \mid \psi_{k1.}^S \sim N(\psi_{k1.}^S, (u_{k1.}^S)^2)$, where y_{k11}^S, y_{k01}^S, and $z_{k1.}^S$ are all independent.

To draw estimates of the population parameters from their joint posterior distribution, we used Gibbs sampling (Gelfand and Smith 1990; Gelman and Rubin 1992). The Gibbs sampling approach estimates the entire posterior distribution and thus allows for greater flexibility in terms of point estimation and inference.

Appendix A describes in detail the conditional distributions underlying the draws of the Gibbs sampler. Appendix B describes how initial parameter estimates of the Gibbs sampler were obtained, together with methods for assessing convergence. Appendix C shows that the assumption $y_{k11}^S \perp y_{k01}^S$ is reasonable unless the CMS sampling fraction is quite large.

A Conditional Distributions for Gibbs Sampling

The posterior distribution of $\{\psi_k^*\}$ given the DA estimate ρ and hyperparameters $a^S, b^S, c^S, d^S, \alpha, \beta, \tau_{k11}^S, \tau_{k01}^S$ and $u_{k1.}^S$ is

$$
\begin{gathered}
p(\psi_k^* \mid y_{k11}^M, y_{k11}^F, y_{k01}^M, y_{k01}^F, z_{k1.}^M, z_{k1.}^F) \propto \\
p(\gamma^M) p(\rho_k) p(\delta_k^M) p(\delta_k^F) p(\phi_k^M) p(\phi_k^F) \times \\
p(y_{k11}^M \mid \psi_{k11}^M, (\tau_{k11}^M)^2) p(y_{k01}^M \mid \psi_{k01}^M, (\tau_{k01}^M)^2) p(z_{k1.}^M \mid \psi_{k1.}^M, (u_{k1.}^M)^2) \times \\
p(y_{k11}^F \mid \psi_{k11}^F, (\tau_{k11}^F)^2) p(y_{k01}^F \mid \psi_{k01}^F, (\tau_{k01}^F)^2) p(z_{k1.}^F \mid \psi_{k1.}^F, (u_{k1.}^F)^2) \propto \\
(\gamma^M)^{\alpha-1} e^{\gamma^M/\beta} \prod_{k=1}^{K} exp(-1/2((\rho_k - \rho)/\sigma)^2) \times \\
\prod_{S=M,F} \prod_{k=1}^{K} (\frac{\Gamma(a^S + b^S)}{\Gamma(a^S)\Gamma(b^S)} (\delta_k^S)^{a^S-1} (1-\delta_k^S)^{b^S-1} \times \\
\frac{\Gamma(c^S + d^S)}{\Gamma(c^S)\Gamma(d^S)} (\phi_k^S)^{c^S-1} (1-\phi_k^S)^{d^S-1} \times \\
exp((\frac{\psi_{k11}^S - y_{k11}^S}{\tau_{k11}^S})^2 + (\frac{\psi_{k01}^S - y_{k01}^S}{\tau_{k01}^S})^2 + (\frac{\psi_{k1.}^S - z_{k1.}^S}{u_{k1.}^S})^2)).
\end{gathered} \tag{2}
$$

We determine the conditional distributions of each scalar parameter θ_j conditional on all data (Y=$\{y_{k11}^S, y_{k01}^S, z_{k1.}^S\}$)) and other parameters $\theta_{-\theta_j}$ by examining the posterior likelihood given by (A.1). Where possible, draws are obtained from standard distributions; otherwise draws are made via the inverse CDF method by determining the conditional PDF up to a constant. The constant of proportionality is obtained by numerical integration using 32-point Gauss-Hermite quadrature on each of 1,000 subintervals $[t_{i-1}, t_i]$ of the support or line segment where the posterior is observed to be non-trivially greater than 0. A draw X from $UNI(0,1)$ is then obtained and the midpoint of the subinterval $[t_{j-1}, t_j]$ where $min_j(\sum_1^j A_{[t_{i-1},t_i]}) > X$ is then taken as a draw of $\theta_j \mid Y, \theta_{-\theta_j}$, where $A_{[t_{i-1},t_i]}$ is the area under the normalized PDF between points t_{i-1} and t_i. We now discuss the conditional distributions required for the Gibbs cycle:

A.1 δ_k^S

$$
\begin{gathered}
p(\delta_k^M \mid Y, \theta^{(t)}_{-\delta_k^M}) \propto \\
(\delta_k^M)^{a^M-1}(1-\delta_k^M)^{b^M-1} exp\{\frac{-1}{2\sigma_1^2}(l\delta_k^M - x)^2 + \\
\frac{-1}{2\sigma_2^2}(m - m\delta_k^M - y)^2 + \frac{-1}{2\sigma_3^2}(n\delta_k^M - z)^2\} \propto \\
(\delta_k^M)^{a^M-1}(1-\delta_k^M)^{b^M-1} exp\{\frac{-(l^2/\sigma_1^2 + m^2/\sigma_2^2 + n^2/\sigma_3^2)}{2} \times \\
[\delta_k^M - \frac{(lx/\sigma_1^2 + (m-y)m/\sigma_2^2 + nz/\sigma_3^2)}{(l^2/\sigma_1^2 + m^2/\sigma_2^2 + n^2/\sigma_3^2)}]^2\}
\end{gathered}
$$

where

$$
l = \phi_k^M \frac{\rho_k}{1+\rho_k}\psi_{k..} \qquad m = \frac{\phi_k^M}{\gamma^M}\frac{\rho_k}{1+\rho_k}\psi_{k..} \qquad n = \frac{\rho_k}{1+\rho_k}\psi_{k..}
$$

and

$$
\begin{gathered}
x = y_{k11}^M \quad y = y_{k01}^M \quad z = z_{k1.}^M \\
\sigma_1^2 = (\tau_{k11}^M)^2 \quad \sigma_2^2 = (\tau_{k01}^M)^2 \quad \sigma_3^2 = (u_{k1.}^M)^2
\end{gathered}
$$

Draws from this distribution are made by the inverse CDF method.

Similar results hold for δ_k^F, with $\frac{\rho_k}{1+\rho_k}$ replaced with $\frac{1}{1+\rho_k}$

A.2 ϕ_k^S

$$
\begin{gathered}
p(\phi_k^M \mid Y, \theta^{(t)}_{-\phi_k^M}) \propto (\phi_k^M)^{c^M-1}(1-\phi_k^M)^{d^M-1} \times \\
exp\{\frac{-1}{2\sigma_1^2}(l\phi_k^M - x)^2 + \frac{-1}{2\sigma_2^2}(m\phi_k^M - y)^2\} \propto \\
(\phi_k^M)^{c^M-1}(1-\phi_k^M)^{d^M-1} exp\{\frac{-(l^2/\sigma_1^2 + m^2/\sigma_2^2)}{2} \times \\
[\phi_k^M - \frac{(lx/\sigma_1^2 + my/\sigma_2^2)}{(l^2/\sigma_1^2 + m^2/\sigma_2^2)}]^2\}
\end{gathered}
$$

where

$$
\begin{gathered}
l = \delta_k^M \frac{\rho_k}{1+\rho_k}\psi_{k..} \quad m = \frac{(1-\delta_k^M)}{\gamma^M}\frac{\rho_k}{1+\rho_k}\psi_{k..} \\
x = y_{k11}^M \quad y = y_{k01}^M \quad \sigma_1^2 = (\tau_{k11}^M)^2 \quad \sigma_2^2 = (\tau_{k11}^M)^2
\end{gathered}
$$

As with δ_k^F, similar results for ϕ_K^F hold replacing $\frac{\rho_k}{1+\rho_k}$ with $\frac{1}{1+\rho_k}$; draws are made by the inverse CDF method.

A.3 $\psi_{k..}$

$$p(\psi_{k..} \mid Y, \theta^{(t)}_{-\psi_{k..}}) \propto$$

$$exp\{\sum_{S=M,F}[\frac{-1}{2(\sigma_1^S)^2}(l^S\psi_{k..} - x^S)^2 + \frac{-1}{2(\sigma_2^S)^2}(m^S\psi_{k..} - y^S)^2 +$$

$$\frac{-1}{2(\sigma_3^S)^2}(n^S\psi_{k..} - z^S)^2]\}$$

$$\Rightarrow \psi_{k..} \mid Y, \theta^{(t)}_{-\psi_{k..}} \sim N(\eta_k, \xi_k^2)$$

with

$$\eta_k = \frac{\sum_{S=M,F}(l^S x^S/(\sigma_1^S)^2 + m^S y^S/(\sigma_2^S)^2 + n^S z^S/(\sigma_3^S)^2)}{((l^S)^2/(\sigma_1^S)^2 + (m^S)^2/(\sigma_2^S)^2 + (n^S)^2/(\sigma_3^S)^2)}$$

$$\xi_k^2 = \left(\sum_{S=M,F}\left[(l^S)^2/(\sigma_1^S)^2 + (m^S)^2/(\sigma_2^S)^2 + (n^S)^2/(\sigma_3^S)^2\right]\right)^{-1}$$

where

$$l^S = \delta_k^S \phi_k^S \frac{\rho_k^{I(S=M)}}{1+\rho_k} \qquad m^S = \frac{(1-\delta_k^S)\phi_k^S}{\gamma^S}\frac{\rho_k^{I(S=M)}}{1+\rho_k}$$

$$n^S = \delta_k^S \frac{\rho_k^{I(S=M)}}{1+\rho_k}$$

and

$$x^S = y^S_{k11} \quad y^S = y^S_{k01} \quad z^S = z^S_{k1.}$$

$$(\sigma_1^S)^2 = (\tau^S_{k11})^2 \quad (\sigma_2^S)^2 = (\tau^S_{k11})^2 \quad (\sigma_3^S)^2 = (u^S_{k1.})^2$$

Draws are made directly from the normal distribution.

A.4 γ^M

$$p(\gamma^M \mid Y, \theta^{(t)}_{-\gamma^M}) \propto$$

$$(\gamma^M)^{\alpha-1} exp\left\{-\gamma^M/\beta) + \sum_k \left[\frac{-1}{2(\tau^M_{k01})^2}(l_k/\gamma^M - y^M_{k01})^2\right]\right\}$$

where

$$l_k = \phi_k^M(1-\delta_k^M)\frac{\rho_k}{1+\rho_k}\phi_{k..}$$

Draws are made via the inverse CDF method.

A.5 ρ_k

Rather than draw ρ_k directly, we draw $M_k = \frac{\rho_k}{1+\rho_k}$ (the fraction of males), where $p(\rho_k) \sim N(\rho, \sigma^2) \Rightarrow p(M_k) \dot\sim N(\frac{\rho}{1+\rho}, \frac{\sigma^2}{(1+\rho)^4})$.

$$p(M_k \mid Y, \theta^{(t)}_{-M_k}) \propto$$

$$exp\{\frac{-1}{2(\sigma_1^F)^2}(l^F(1-M_k) - x^F)^2 + \frac{-1}{2(\sigma_2^F)^2}(m^F(1-M_k) - y^F)^2 +$$

$$\frac{-1}{2(\sigma_3^F)^2}(n^F(1-M_k) - z^F)^2 + \frac{-1}{2(\sigma_1^M)^2}(l^M M_k - x^M)^2 +$$

$$\frac{-1}{2(\sigma_2^M)^2}(m^M M_k - y^M)^2 + \frac{-1}{2(\sigma_3^M)^2}(n^M M_k - z^M)^2 +$$

$$\frac{-1}{2\sigma^2/(1+\rho)^4}(M_k - \frac{\rho}{1+\rho})^2\}$$

$$\Rightarrow M_k \mid Y, \theta^{(t)}_{-M_k} \sim N(\eta, \xi_k^2)$$

with

$$\eta_k = (l^F(l^F - x^F)/(\sigma_1^F)^2 + m^F(m^F - y^F)/(\sigma_2^F)^2 + n^F(n^F - z^F)/(\sigma_3^F)^2 +$$

$$l^M x^M/(\sigma_1^M)^2 + m^M y^M/(\sigma_2^M)^2 + n^M z^M/(\sigma_3^M)^2 + \frac{(1+\rho)^4}{\sigma^2})/$$

$$(\sum_{S=M,F}[(l^S)^2/(\sigma_1^S)^2 + (m^S)^2/(\sigma_2^S)^2 + (n^S)^2/(\sigma_3^S)^2] + \frac{\rho(1+\rho)^3}{\sigma^2})$$

$$\xi_k^2 =$$

$$\left(\sum_{S=M,F}[(l^S)^2/(\sigma_1^S)^2 + (m^S)^2/(\sigma_2^S)^2 + (n^S)^2/(\sigma_3^S)^2] + \frac{\rho(1+\rho)^3}{\sigma^2}\right)^{-1}$$

where

$$l^S = \delta_k^S \phi_k^S \psi_{k..} \qquad m^S = \frac{(1-\delta_k^S)\phi_k^S \psi_{k..}}{\gamma^S} \qquad n^S = \delta_k^S \psi_{k..}$$

and

$$x^S = y_{k11}^S \quad y^S = y_{k01}^S \quad z^S = z_{k1.}^S$$

$$(\sigma_1^S)^2 = (\tau_{k11}^S)^2 \quad (\sigma_2^S)^2 = (\tau_{k11}^S)^2 \quad (\sigma_3^S)^2 = (u_{k1.}^S)^2$$

Since the constraint $\sum_k \psi_{k..}^M / \sum_k \psi_{k..}^F = \rho$ implies that the draws must be make subject to the constraint $\sum_k w_k M_k = \overline{M} = \rho/(1+\rho)$ where $w_k = \psi_{k..}/\sum_k \psi_{k..}$, draws are made from $N_K(\eta, \xi)$ conditional on $\overline{M} = \rho/(1+\rho)$ where $\eta = (\eta_1, \ldots, \eta_K)'$ and $\xi = \mathrm{diag}(\xi_1, \ldots, \xi_K)$. This is accomplished via the SWEEP operator (Goodnight 1979; Little and Rubin 1987, p.112-119), which shows that draws from this conditional multivariate normal are equivalent to draws from $N_{K-1}(\eta^*, \xi^*)$, where $\eta_k^* = \eta_k - \frac{w_k \xi_k}{\sum_k w_k^2 \xi_k}\overline{\eta}$ and $\xi_{kj}^* = \xi_k I\{k=j\} - \frac{w_k \xi_k w_j \xi_j}{\sum_k w_k^2 \xi_k}$ where $I\{\cdot\}$ is the indicator function. After drawing $M_1, \ldots, M_{K-1}$ from $N_{K-1}(\eta^*, \xi^*)$, M_K is given by $(\overline{M} - \sum_1^{K-1} w_k M_k)/w_K$.

A.6 a^S, b^S, c^S, d^S

$$p(a^S, b^S \mid Y, \theta^{(t)}_{-a^S,b^S} \propto \frac{\Gamma(a^S + b^S)}{\Gamma(a^S) + \Gamma(b^S)} \prod_{k=1}^{K} (\delta_k^S)^{a^S-1}(1-\delta_k^S)^{b^S-1} =$$

$$\frac{\Gamma(a^S + b^S)}{\Gamma(a^S) + \Gamma(b^S)} e^{(a^S-1)\sum_k log\delta_k^S} e^{(b^S-1)\sum_k log(1-\delta_k^S)}$$

Draws are made from the joint distribution of a^S, b^S by determining the joint distribution of a^S, b^S over a grid, summing across the grid to approximate the marginal distribution of a^S, drawing a value of a^S via the inverse CDF method, and then drawing a value of b^S via the inverse CDF method from the distribution of b^S conditional on the drawn value of a^S.

A similar approach is used to draw c^S and d^S.

B Initial estimates and convergence of the Gibbs chain

To obtain initial estimates, we fit an alternative and somewhat simpler model that assumes $y^S_{k11} \sim N(\psi^S_{k11}, (\tau^S_{k11})^2)$, $z^S_{k1.} \sim N(\psi^S_{k1.}, (u^S_{k1.})^2)$, and $\psi^S_{k11} = \phi^S_k \psi^S_{k1.}$, where $\phi^S_k \sim BETA(c^S, d^S)$. We assume $\psi^S_{k01} = y^S_{k01}$, that is that ψ^S_{k01} has a degenerate prior distribution where $\psi^S_{k01} = y^S_{k01}$ with probability 1. Then, to a first-order approximation, $r^S_k \mid \phi^S_k \sim N(\phi^S_k, w^S_k)$ where $\frac{y^S_{k11}}{z^S_{k1.}} = r^S_k$ and $w^S_k = \frac{(\tau^S_{k11})^2}{{z^S_{k1.}}^2} + (u^S_{k1.})^2 \left(\frac{(y^S_{k11})^2}{(z^S_{k1.})^4}\right)$, the ratio estimator of the variance of r^S_k. To obtain the posterior distribution of ϕ^S_kwe note that, up to a constant,

$$f(\phi^S_k \mid r^S_k) \propto f(\phi^S_k) f(r^S_k \mid \phi^S_k) =$$

$$(\phi^S_k)^{(c^S-1)}(1-\phi^S_k)^{(d^S-1)} exp(-\frac{1}{2w^S_k}(r^S_k - \phi^S_k)^2)$$

Empirical Bayes estimates of c^S and d^S can be obtained by integrating over ϕ_k:

$$r_k \approx N(\frac{c^S}{c^S + d^S}, E(w_k) + \frac{c^S d^S}{(c^S+d^S)^2(c^S+d^S+1)})$$

Evaluating $E(w_k)$ is a non-trivial task; however, when $\psi_{k11} >> \tau_{k11}$ and $\psi_{k1.} >> u_{k1.}$, as is the case here, $E(w_k) \doteq w_k$. Algebra then gives the empirical Bayes estimators

$$\widehat{c^S} = \frac{\overline{(r^S)}^2(1-\overline{r^S})}{S^2(r^S) - \overline{w^S}} - \overline{r^S} \qquad \widehat{d^S} = \widehat{c^S}(\overline{(r^S)}^{-1} - 1)$$

The estimator of the posterior mean $\phi^S_k = E(\phi^S_k \mid r^S_k)$ can then be obtained through numerical integration. The "In Census-In CMS" cell (ψ^S_{k11}) is then estimated by $\widehat{\phi^S_k} z^S_{k1.}$, and the "In Census-Out CMS" cell (ψ^S_{k10}) by $(1-\widehat{\phi^S_k})z^S_{k1.}$. The "Out CMS-Out Census" (ψ^S_{k00}) cell is estimated under the ICR model (assumption of independence of capture-recapture) with no adjustment for sex ratio ρ. The estimates of ψ^S_{kij} are then transformed using (7) to provide initial estimates of δ^S_k, ϕ^S_k, γ^M; the initial estimate of $\psi_{k..}$ is then drawn conditional on the other parameters as described in Section A.3. Initial estimates of the hyperparameters a^S and b^S can be then made as for c^S and d^S

above. (In the application in Section 4, the empirical Bayes estimators of c^M and c^F were negative and lead to improper posterior distributions. Thus initial estimates of c^S and d^S were estimated by crude method-of-moment estimators
$(\overline{\hat{\phi}^S} = \frac{c^S}{c^S+d^S}, \widehat{Var}(\hat{\phi}^S) = \frac{c^s d^S}{(c^S+d^S)^2(c^s+d^s+1)} \Rightarrow c^S = (\frac{\overline{\hat{\phi}^S}(1-\overline{\hat{\phi}^S})}{Var(\hat{\phi}^S)} - 1)\overline{\hat{\phi}^S}, d^S = (\frac{\overline{\hat{\phi}^S}(1-\overline{\hat{\phi}^S})}{Var(\hat{\phi}^S)} - 1)(1 - \overline{\hat{\phi}^S})).)$

To formally test that convergence has occurred, we use the method suggested in Gelman et al. (1995, p.331-332). A series of J different startpoints of the Gibbs sampler are utilized, and a subset of n iterations at which it is believed convergence has been achieved is selected and the between- and within-sequence posterior variances and for each (scalar) parameter calculated. The estimated ratio of the marginal posterior variance for all sequences to the mean within-sequence posterior variance, $\widehat{R}$, then provides a check of the degree of convergence for the parameters. In particular, Gelman et al. suggest that $\sqrt{\widehat{R}} \approx 1.2$ or less indicates acceptable convergence has been obtained for most purposes.

We used 5 startpoints for the parameters: those obtained through the simplified model described above, and those estimates with ϕ_k^S and γ^M changed to a 2×2 design where ϕ_k^S=0.75 or 0.99 for all S, k and γ^M=0.75 or 2.5. The 201st for 1000th iterations were selected. Examination of all parameters estimated under our model for the 12 African-American 30-49 strata shows that $1.00 < \sqrt{\widehat{R}} < 1.07$.

The autocorrelation of the Gibbs sequence is ignored in the above approximation, which is reasonable if n is large (Gelman, et al. 1995, p. 330). Although a moderate degree of autocorrelation is present in the data, our posterior estimates are made from 4,000 draws from five independent chains, diminishing this effect.

C Correlation between In-Census-In CMS and Out-Census-In CMS counts

We make the simplifying assumption that the total number of persons sampled in the CMS of a given age-race-gender category in stratum k conditional on the true age-race-gender population $\psi_{k..}^F$ is $w_{k.1}^S \sim BIN(\psi_{k..}^S, \pi_k^S)$ where π_k is the probability of selection into the CMS. Then the number of persons in the CMS sample who were also observed in the Census conditional on the CMS sample size is $w_{k11}^S \mid w_{k.1}^S \sim BIN(w_{k.1}^S, p_k^S)$. Making the further somewhat simplified assumption that $y_{k11}^S = w_{k11}^S/\pi_k^S$, we have

$$Cov(y_{k11}^S, y_{k01}^S) = (1/\pi_k^{S\,2})Cov(w_{k11}^S, w_{k01}^S) =$$
$$(1/(\pi_k^S)^2)(E[-w_{k.1}^S p_k^S(1-p_k^S)] + Cov[w_{k.1}^S p_k^S, w_{k.1}^S(1-p_k^S)]) =$$
$$-\psi_{k..}^S p_k^S(1-p_k^S)$$

Similarly, $Var(y_{k11}) = \frac{\psi_{k..}^S p_k^S}{\pi_k^S}[1 - p_k^S\pi_k^S]$ and $Var(y_{k01}) = \frac{\psi_{k..}^S(1-p_k^S)}{\pi_k^S}[1-(1-p_k^S)\pi_k^S]$. Thus $Corr(y_{k11}^S, y_{k01}^S) = -\pi_k^S/(1-p_k^S\pi_k^S)(1-(1-p_k^S)\pi_k^S)$, which implies $Corr(y_{k11}^S, y_{k01}^S) \to 0$ as $\pi_k^S \to 0$.

References

Bell, W. (1993), "Using Information from Demographic Analysis in Post-Enumeration Survey Estimation," *Journal of the American Statistical Association*, 88:1106-1118.

Gelfand, A.E., and Smith, A.M.F. (1990). "Sampling-based approaches to calculating marginal densities," *Journal of the American Statistical Association*, 85:389-409.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis.* New York:Chapman and Hall.

Gelman, A., and Rubin, D.B. (1992). "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7:457-472.

Goodnight, J.H. (1979). "A tutorial on the SWEEP operator," *American Statistician*, 33:149-158.

Hogan, H. (1992). "The 1990 Post-Enumeration Survey: Operations and New Estimates," in Proceedings of the American Statistical Association, Survey Research Methods Section.

Little R.J.A., and Rubin, D.B. (1987). *Statistical Analysis with Missing Data.* New York:Wiley.

Robinson, J.G. (1993). "Estimation of population coverage in the 1990 United States Census based on demographic analysis," *Journal of the American Statistical Association*, 88:1061-1071.

Robinson, J.G. (1995). "Coverage measurement and evaluation in the 2000 Census: what is the role of demographic analysis," Paper presented at the Census Advisory Committee of Professional Associations meeting.