BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: Census/SRD/RR-86/27

by

Kirk M. Wolter
Statistical Research Division
Bureau of the Census
Room 3203-4
Washington, D.C. 20233 U.S.A.

Recommended by: Kirk M. Wolter

Report Completed: May 2, 1986

Report issued: May 2, 1986

Revised by: Richard Griffin

Revision issued: March 11, 2008

May 2, 1986

A Combined Coverage Error Model for Individuals and Housing Units

by

Kirk M. Wolter
U. S. Bureau of the Census

1. Introduction. In Wolter (1986), basic models are presented for representing the coverage of individuals in surveys and censuses of human populations. The models are related to the capture-recapture models employed in estimating the size and density of wildlife populations, to the dual-system models employed in estimating the number of human vital events, and to the log-linear models employed in the analysis of discrete data. This paper builds on the earlier work, extending the basic models to represent the coverage of both housing units and individuals, and in the process one of the key independence assumptions specified in the basic models is relaxed. Section 2 presents the extended model, while the parameter estimators and their properties are discussed in Section 3.

2. Extended Model. We consider a given human population U, and let N denote the number of individuals in U. N is considered unknown and to be estimated. Two censuses (A and B) of U are conducted using an identical time reference, and for a variety of reasons, some individuals are missed by A or B. We will model the results of the censuses and use the model to estimate N.

Following the approach in Wolter (1986), we will single-out one of the basic coverage error models (M_t) for detailed development, however, it will be clear that the extensions developed in this article can be made for any of the basic coverage error models.

For completeness, we review briefly the basic model M_t. It is characterized by the following assumptions:

(i) (The Closure Assumption) We assume U is closed and of fixed size N.

(ii) (The Multinomial Assumption) Let ξ denote the multinomial distribution with parameters.

		List B		
		in	out	
List A	in	p_{11}	p_{12}	p_{1+}
	out	p_{21}	p_{22}	p_{2+}
		p_{+1}	p_{+2}	1 .

We assume that the joint event that the i-th individual is in List A or not and in List B or not is correctly modeled by ξ. This assumption combines (ii) and (xi) in Wolter (1986).

(iii) (Autonomous Independence) We assume that Lists A and B are created as a result of N mutually independent trials, one per individual member of U, utilizing the distribution ξ. The resulting data are

		List B		
		in	out	
List A	in	x_{11}	x_{12}	x_{1+}
	out	x_{21}	x_{22}	x_{2+}
		x_{+1}	x_{+2}	$x_{++} = N$,

where $x_{ab} = \sum_i x_{iab}$ and x_{iab} is an indicator random variable signifying whether or not the i-th individual is in cell (a,b), for a,b = 1,2,+. The count x_{22}, and thus N, is considered unknown and to be estimated on the basis of the model.

(iv) (The Matching Assumption) We assume it is possible to match correctly List B to List A, thus permitting us to observe x_{11}, x_{12}, and x_{21}.

(v) (Spurious Events Assumption) We assume that both lists are void of spurious events or that such are eliminated prior to estimation.

(vi) (The Nonresponse Assumption) We assume that sufficient identifying information is gathered about the nonrespondents in both censuses to permit an exact match from B to A.

(vii) (The Poststratification Assumption) We assume that any variable employed for poststratification is correctly recorded for all individuals on both lists.

(viii) (Causal Independence) The event of being enumerated in A is independent of the event of being enumerated in B. Thus, $p_{ab} = p_{a+} p_{+b}$ for a,b = 1,2.

Given M_t,the maximum likelihood estimator of N, also called the Petersen estimator, is given by

$$\hat{N}_t = \frac{x_{1+} x_{+1}}{x_{11}} .$$

See Wolter (1986) for a discussion of the properties of $\hat{N}_t$.

One of the main weaknesses of M_t is that the individuals in U reside in housing units (HU) and households sometimes act together in contributing to coverage error. We will improve M_t by accounting separately for the occurrence of whole HU misses and within HU misses, and in the process we will relax somewhat the assumption, (iii), of individual autonomy. This is an important improvement as evidenced by the 1970 U.S. Decennial Census, where roughly half of the total omissions of individuals were due to the omission of whole housing units, with the remaining half due to the omission of individual people within enumerated HU's. Comparable data from the 1980 Census are not available.

Let the N members of U reside in H HU's, with M_i members within the i-th HU. Now both N and H are unknown and to be estimated. The extended model, called M_{t-e}, is obtained by replacing (ii), (iii), and (viii) with (ii-e), (iii-e), and (viii-e):

(ii-e) (The Multinomial Assumption) Let ξ_1 denote the multinomial distribution with parameters

		List B		
		in	out	
List A	in	s_{11}	s_{12}	s_{1+}
	out	s_{21}	s_{22}	s_{2+}
		s_{+1}	s_{+2}	1 .

We assume that the joint event that the i-th HU is enumerated in A or not and in B or not is correctly modeled by the distribution ξ_1. Given the ξ_1 outcome for the i-th HU, we assume that the joint event that the j-th individual (j=1, ..., M_i) is enumerated in A or not and in B or not is correctly modeled by the appropriate one of the following multinomial distributions:

(ξ_{11}: given $i\varepsilon$ A and $i\varepsilon$B)

		List B		
		in	out	
List A	in	t_{11}	t_{12}	t_{1+}
	out	t_{21}	t_{22}	t_{2+}
		t_{+1}	t_{+2}	1 ;

(ξ_{12}: given iεA and i$\notin$B)

		List B		
		in	out	
List A	in	—	t_{1+}	t_{1+}
	out	—	t_{2+}	t_{2+}
		—	1	1 ;

(ξ_{21}: given i$\notin$A and iεB)

		List B		
		in	out	
List A	in	—	—	—
	out	t_{+1}	t_{+2}	1
		t_{+1}	t_{+2}	1 ;

(ξ_{22}: given i$\notin$A amd i$\notin$B)

		List B		
		in	out	
List A	in	—	—	—
	out	—	1	1
		—	1	1.

Thus, we have created a hierarchical structure for the coverage of persons; with HU coverage occuring first and person coverage occuring second, conditional up the HU coverage outcome. Note that the unconditional coverage probabilities for persons are given by

		List B		
		in	out	
List A	in	$t_{11}\ s_{11}$	$t_{12}s_{11} + t_{1+}s_{12}$	$t_{1+}s_{1+}$
	out	$t_{21}s_{11} + t_{+1}s_{21}$	$t_{22}s_{11} + t_{2+}\ s_{12} + t_{+2}s_{21} + s_{22}$	$s_{2+} + t_{2+}s_{1+}$
		$t_{+1}s_{+1}$	$s_{+2} + t_{+2}s_{+1}$	1

Let the entries in this table be denoted by p_{ab} for a,b = 1,2,+.

(iii-e) (Autonomous Independence) We assume that HU's are enumerated or not in A and B as a result of H mutually independent trials, utilizing distribution ξ_1. Conditional on the enumeration status of the i-th HU, the individuals within the HU are enumerated or not in A and B as a result of M_i mutually independent trials utilizing ξ_{11}, ξ_{12}, ξ_{21}, or ξ_{22}, as the case may be. Each of these trials corresponds to a member of the i-th HU, for i=1,...,H.

(viii-e) (Causal Independence) Regarding HU's, the event of being enumerated in A is independent of the event of being enumerated in B. That is, $s_{ab} = s_{a+}\ s_{+b}$, for a,b = 1,2. Given that the i-th HU is included in both A and B, the enumeration of an individual HU member in A is conditionally independent of the enumeration in B. That is, $t_{ab} = t_{a+}\ t_{+b}$, for a,b = 1,2. Thus, the unconditional distribution exhibits independence, with $p_{ab} = p_{a+}\ p_{+b}$, for a,b = 1,2.

Notice that under the extended model M_{t-e}, individuals who reside in different HU's act autonomously with respect to enumeration status, but individuals within the same HU do not. Thus, we have created a more realistic condition than the original autonomy assumption in basic model M_t.

Given this extended model, if the i-th HU is enumerated by A (or B) then 0, 1, 2, ..., or M_i individuals within the HU may be enumerated. But if the i-th HU is not enumerated, then the model does not permit any of its residents to be enumerated. Thus, the model departs just slightly from real census taking outcomes, where it is possible for an individual to be enumerated while the corresponding HU is not. This occurs, e.g., in the case of apartment mixups in central city areas.

3. Estimators and Their Properties. Define indicator random variables x_{ijab}, signifying whether or not the j-th individual in the i-th HU is in cell (a,b), for a,b = 1,2,+. Define

$$m_{iab} = \sum_{j=1}^{M_i} x_{ijab},$$

i.e., the number of individuals in the i-th HU that possess enumeration status (a, b), for a,b = 1,2,+. Define indicator random variables x_{iab}, signifying whether or not the i-th HU possesses enumeration status (a,b), for a,b = 1,2,+.

- The observed data consist of counts of HU's

		List B		
		in	out	
List A	in	h_{11}	h_{12}	h_{1+}
	out	h_{21}	——	
		h_{2+}		

and counts of individuals

		List B		
		in	out	
List A	in	x_{11}	x_{12}	x_{1+}
	out	x_{21}	——	
		x_{+1}		

,

where

$$h_{ab} = \sum_{i=1}^{H} x_{iab}$$

$$x_{ab} = \sum_{i=1}^{H} \sum_{j=1}^{M_i} x_{ijab}$$

for (a,b) = (1,1), (1,2), (2,1), (1,+), (+,1).

We will consider the estimator $(\hat{N}, \hat{H})$ of (N, H), where

$$\hat{N} = \frac{x_{1+}\ x_{+1}}{x_{11}}$$

$$\hat{H} = \frac{h_{1+}\ h_{+1}}{h_{11}} .$$

$\hat{H}$ is the maximum likelihood estimator of H and $\hat{N}$ is the natural extension of the Petersen estimator to the extended model M_{t-e}.

Given standard regularity conditions, the estimation error is

$$\begin{pmatrix} \hat{N} - N \\ \hat{H} - H \end{pmatrix}$$

is asymptotically a bivariate normal random variable with mean

$$\underset{\sim}{\delta} = \begin{pmatrix} \delta_N \\ \delta_H \end{pmatrix} = \begin{pmatrix} \dfrac{p_{2+}p_{+2}}{p_{1+}p_{+1}} + \dfrac{A}{N}\ \dfrac{s_{2+}\ s_{+2}}{s_{1+}\ s_{+1}} \\ \\ \dfrac{s_{2+}\ s_{+2}}{s_{1+}\ s_{+1}} \end{pmatrix}$$

and covariance matrix

$$\underset{\sim}{\Sigma} = \begin{pmatrix} \sigma_N^2 & \sigma_{NH} \\ \text{sym} & \sigma_H^2 \end{pmatrix}$$

$$= \begin{pmatrix} N\ \dfrac{p_{2+}\ p_{+2}}{p_{1+}\ p_{+1}} + A\ \dfrac{s_{2+}\ s_{+2}}{s_{1+}\ s_{+1}} & N\ \dfrac{s_{2+}\ s_{+2}}{s_{1+}\ s_{+1}} \\ \text{sym} & H\ \dfrac{s_{2+}\ s_{+2}}{s_{1+}\ s_{+1}} \end{pmatrix},$$

where $A = \sum_{i=1}^{H} M_i\ (M_i - 1)$.

The second terms in δ_N and σ_N^2 represent addition bias and variance associated with the extended model M_{t-e}, but not with the basic model M_t. Indeed, letting $M_i = 1$, δ_N and σ_N^2 reduce to the bias and variance of the Petersen estimator. The bias δ_H and variance σ_H^2 of $\hat{H}$, are well-known expressions for the Petersen estimator, here applied to HU's instead of individuals.

To estimate $\underset{\sim}{\Sigma}$, we suggest the natural consistent estimator

$$\underset{\sim}{\hat{\Sigma}} = \begin{pmatrix} \hat{\sigma}_N^2 & \hat{\sigma}_{NH} \\ & \hat{\sigma}_H^2 \end{pmatrix} ,$$

where

$$\hat{\sigma}_N^2 = \frac{x_{1+}x_{+1}x_{12}x_{21}}{x_{11}^3} + \frac{x_{+1}^2\, h_{21}\, h_{12}}{x_{11}^2\, h_{11}\, h_{+1}} \sum_{i=1}^{H} m_{1+i}\ (m_{1+i} - 1),$$

$$\hat{\sigma}_H^2 = \frac{h_{1+}\, h_{+1}\, h_{12}\, h_{21}}{h_{11}^3} ,$$

and

$$\hat{\sigma}_{NH} = \frac{x_{1+}x_{+1}h_{12}h_{21}}{x_{11}\, h_{11}^2} .$$

The first terms in $\hat{\sigma}_N^2$ and $\hat{\sigma}_H^2$ take the form of the well-known estimator of variance for the Petersen estimator.

Two interesting special cases of model M_{t-e} are

(a) only within HU omissions, no whole HU omissions; and
(b) only whole HU omissions, no within HU omissions.

For these special cases, we have:

(a) $s_{1+} = s_{+1} = 1$, $t_{1+} = p_{1+}$, $t_{+1} = p_{+1}$, $\hat{H} = H$, $\sigma_H^2 = 0$, $\delta_N = p_{2+}p_{+2}/(p_{1+}p_{+1})$, $\sigma_N^2 = N\, p_{2+}p_{+2}/(p_{1+}p_{+1})$. In other words this case just reverts to the basic model M_t.

(b) $t_{1+} = t_{+1} = 1$, $s_{1+} = p_{1+}$, $s_{+1} = p_{+1}$, $\delta_N = (\sum_{i=1}^{H} M_i^2 / N)\,(p_{2+}p_{+2})/(p_{1+}p_{+1})$, $\sigma_N^2 = (\sum_{i=1}^{H} M_i^2)\,(p_{2+}p_{+2}) / (p_{1+}p_{+1})$, and the moments of $\hat{H}$ remain unchanged. his case results in considerable loss in efficiency vis-à-vis the basic model M_t. In fact, the relative efficiency

$$RE = \frac{Np_{2+}p_{+2} / (p_{1+}p_{+1})}{\sum_{i=1}^{H} M_i^2 \ (p_{2+}p_{+2}) / (p_{1+}p_{+1})} = \frac{N}{\sum_{i=1}^{H} M_i^2}$$

reaches its maximum value, max RE = H/N, whenever HU's are of equal size $M_i = \overline{M} = N/H$. In the U.S., $\overline{M} \doteq 2.7$ and thus max RE $\doteq$ .37. In U.S. Bureau of the Census (1986), the following distribution of household size is presented:

Number of Persons	Frequency
1	.234
2	.315
3	.177
4	.159
5	.071
6	.028
7+	.016

For these data, which refer to calendar year 1984, we have RE = .294. The loss in efficiency associated with case (b) is enormous, and we are fortunate that no real census is likely to consist strictly of whole HU omissions.

References

U.S. Bureau of the Census (1985), Statistical Abstract of the United States: 1986 (106th edition) Washington, D.C.

Wolter, Kirk M. (1986), "Some Coverage Error Models for Census Data," Journal of the American Statistical Association, 81, .