Expanded Dual System Estimation Model and its Matching Properties in Census Coverage Error Measurement

Ran Tao
Research Institutes of Statistical Sciences, National Bureau of Statistics, Beijing, CHINA, e-mail: tr@gj.stats.cn

Abstracts

U.S. Census Bureau's previous coverage measurement surveys were designed primarily to estimate net Census Coverage Errors (CCE) using Dual System Estimation Model (DSEM). National research council (2009) gave a recommendation on developing plans for measuring the components of CCE in U.S. 2010 census coverage measurement using a framework provided by Mulry and Kostanich (2006). Firstly, this paper analysed the lacks between DSEM in theory and application. One assumptions was added to the model. Secondly, expanded DSEM, to measure a general census coverage error in China, was constructed based on ideas of the framework and the added assumptions. At last, matching properties were proved and estimation approaches to measure the components of CCE were also discussed in this presentation.

Key Words: Expanded dual system estimation model; Matching properties; Census coverage error;

1. Introduction

Capture-recapture models originated in the 17th century, and the modern development dated form Peterson in 1896. Two-sample capture-recapture method was used to estimate the population size of fish in a lake by matching records between two captures in table 1.

Table 1 Matching Records

		Recpture		Total
		Yes	No	
Capture	Yes	$x_{11}\ (p_{11})$	$x_{12}\ (p_{12})$	$x_{1+}\ (p_{1+})$
	No	$x_{21}\ (p_{21})$	$x_{22}\ (p_{22})$	$x_{2+}\ (p_{2+})$
Total		$x_{+1}\ (p_{+1})$	$x_{+2}\ (p_{+2})$	N

x_{ij} is the number of matching records, p_{ij} is the probability of each cell. The method relies on the following classical assumptions:

(1) The population is closed, that N is the same for each capture;

(2) Each fish is equally likely to be caught in one capture;

(3)The two captured sample are independent, that $p_{ij} = p_{i+}p_{+j}$ or $p_{11}p_{22} = p_{12}p_{21}$;

(4)Marked fish in the first capture can be identified in recapture;

Then, the population size N is estimated by $\hat{N} = \frac{x_{1+}x_{+1}}{x_{11}}$ (1)

Marks (1978) discussed the method in post-enumeration survey (PES) for census

evaluation, named dual system estimation (DSE). Wolter (1983) discussed the theory of DSE model to estimate the census total true population and measure the census coverage error (CCE). Extensive recent discussion is presented by applied in the U.S. census (Hogan 1992, 2003; U.S. Census Bureau 2004; Mulry and Kostanich 2006; National Research Council 2009). This paper discusses how to expand this method to all census records on the base of foreign research and application.

2. Theory and Application of DSE Model

2.1 Model Theory and Application

The theory of capture-recapture model was used to estimate the census population, the census equals first capture and the PES equals the recapture. The four original model assumptions will be transformed into as follows:

(1)The total population of census and PES is closed.

(2) Any investigation unit has a same probability of registered in census or PES;

(3) Census and PES are independent of each other;

(4) The information of Records can be matching between census and PES;

In U.S. census application，PES got a sample of blocks, the census enumeration records of the sample blocks called E records, the PES records of the same sample block called P records. Matching results between E and P are shown in Table 2.

表 2　Matching Records of DSE

		In P records		Total
		Yes	No	
In E records	Yes	M		N_C
	No		N_{00}	
Total		N_P		N

In the directly form of (1), that DSE of N is $\hat{N}_{DSE} = N_C \frac{N_P}{M}$ (2)

Equation (2) is just a application form of capture-recapture model theory, not true DSE. Census and PES are not perfect, not all the records in E or P can be used to match. Real matching records are the “correct records” in E and P, so defined them as CE and CP. Then, in order to satisfy assumption (2), DSE was used in every post-stratification of units in the sample blocks, we need to use sample to estimate the population. True DSE of post-stratification s is defined as

$$\hat{N}_{DSE}^{(s)} = \hat{N}_{CE} \frac{\hat{N}_{CP}}{\hat{M}} \qquad (3)$$

The estimation of true population N is $\hat{N} = \sum_s \hat{N}_{DSE}^{(s)}$.

2.2 The lack of DSE

First, the four assumptions of DSE model directly came from the capture-recapture theory, but not enough to guarantee the validity of the model can be used in practice. In addition, not all records in census and PES can participate in model estimation, it is

necessary to add some assumption to clarify the premise of which records in all census and PES can participate in model estimation

Secondly, on the basis of the matching records in DSE model, we can only estimate net census coverage error through the difference between estimation value and the original census results. How to estimate the decomposition factors of census coverage error provides a breakthrough for our improvements of model theory.

3. Expanded DSE Model

3.1 Expanded Model assumption

In order to meet the assumptions (4), Childers (2001) defined units with complete name and at least two characteristics as"sufficient information for matching ". In this section, this paper puts forward the following expanded assumption:

(5) "correct enumeration" records in census and PES can participate in model estimation;

In order to meet the assumptions (5), this paper defines unit with "belong to the target population of the census, only enumeration with sufficient information for matching" as "correct enumeration". Further, this paper defines E records and P records of sample blocks which meet the assumptions (4) as E-sample and P-sample, and defines 0-1 variable of unit U_{ij} as follows:

$$E_{ij}=\begin{cases}1 & only\ enumeration\ of\ U_{ij}\ in\ E\text{-}sample\\ 0 & otherwise\end{cases} \qquad P_{ij}=\begin{cases}1 & only\ enumeration\ of\ U_{ij}\ in\ P\text{-}sample\\ 0 & otherwise\end{cases}$$

That, $\sum E_{ij}$ and $\sum P_{ij}$ are the "correct records" enumeration in E-sample and P-sample.

3.2 Matching Records in Expanded DSE Model

We can estimate the population in a post-stratification by matching records as shown in Table 3.

Table 3 Matching Records based on expanded assumption

		sufficient information records of P-sample		Total
		1	0	
Sufficient information records of E-sample	1	M	N_{10}	N_{CE}
	0	N_{01}	N_{00}	
Total		N_{CP}		N_{DSE}

Hogan (2003) defined units with name and at least one characteristics as"data-defined record" (*DD*).

Mulry and Kostanich (2006) give a series of definition in framework for CCE components as follows:

(1) those correctly enumerated in the census, *CE*;

(2) those enumerated in the census but in the wrong location, *WL*;

(3) those erroneously enumerated in the census, *EE*;

(4) those with insufficient information for matching to the P-sample, *II*;

(5) those that are not data defined in the census, *NDD*;

(6) those omitted in the census, *OM*;

Considering table 3, this paper defined units with"sufficient information for matching " as *SI*. The following additional relationships are used below:

$$N_C = DD + NDD = SI + II + NDD = N_{EE} + N_{CE} + II + NDD$$

$$= N_{EE} + N_{CE} + II_{EE} + II_{CE} + NDD_{EE} + NDD_{CE}$$

Where: $N_{CE} = CE + WL = CE_{11} + WL_{11} + CE_{10} + WL_{10}$

$II_{CE} = II_{01} + II_{00}$ $NDD_{CE} = NDD_{01} + NDD_{00}$ $OM_{CE} = NDD_{01} + NDD_{00}$

Then, $M = CE_{11} + WL_{11}$ $N_{10} = CE_{10} + WL_{10}$ $N_{01} = II_{01} + WL_{10}$

$$N_{CP} = CE_{11} + WL_{11} + II_{01} + NDD_{01} + OM_{01}$$

3.3 **Expanded DSE Model**

If consider matching records in Table 3 in all census and PES records, we can obtain matching records of extended DSE shown in Table 4. The shadow region "1, 0" corresponds to table 3.

Table 4 Matching Records of expanded DSE model

			In P records				Total			
			Yes			No				
			1	0						
In E records	YES					N_{EE}		SI	DD	N_C
		1		CE_{11}	CE_{10}		N_{CE}			
				WL_{11}	WL_{10}					
		0		II_{01}	II_{00}	II_{EE}	II			
				NDD_{01}	NDD_{00}	NDD_{EE}	NDD			
	No			OM_{01}	OM_{00}		OM			
			N_{EP}							
Total				N_{CP}	N_{PP}	N_{OP}		N_{DSE}		
			N_P							

True DSE of post-stratification s is defined as

$$\hat{N}_{DSE}^{(s)} = \hat{N}_{CE}\frac{\hat{N}_{CP}}{\hat{M}} = (\widehat{CE}_{11} + \widehat{CE}_{10} + \widehat{WL}_{11} + \widehat{WL}_{10})\frac{\widehat{CE}_{11} + \widehat{WL}_{11} + \widehat{II}_{01} + \widehat{NDD}_{01} + \widehat{OM}_{01}}{\widehat{CE}_{11} + \widehat{WL}_{11}} \quad (4)$$

4. Matching Properties of Expanded DSE Model

In order to discuss DSE could provides an unbiased estimate of the total population, Mulry and Kostanich (2006) put forward three assumption for matching properties, but failed to be proved. In this section, this paper will prove the two important of them through expanded DSE model.

Assumption 1, correct enumerations in the matching universe are included in the P-sample at the same rate as all correct enumerations.

According to the assumption (3) of the model, there is $p_{11}p_{22} = p_{12}p_{21}$ and $\frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}}$ in table 1.Therefore:

$$\frac{p_{11}}{p_{12}} = \frac{p_{11} + p_{21}}{p_{12} + p_{22}} \tag{5}$$

$$\frac{p_{11}}{p_{11} + p_{12}} = \frac{p_{21}}{p_{21} + p_{22}} = \frac{p_{11} + p_{21}}{p_{11} + p_{12} + p_{21} + p_{22}} = p_{11} + p_{21} \tag{6}$$

With the form of (5)

In table 3, $$\frac{M}{N_{10}} = \frac{N_{CP}}{N_{10} + N_{00}} \tag{7}$$

In table 4, $$\frac{CE_{11} + WL_{11}}{CE_{10} + WL_{10}} = \frac{CE_{11} + WL_{11} + II_{01} + NDD_{01} + OM_{01}}{CE_{10} + WL_{10} + II_{00} + NDD_{00} + OM_{00}} \tag{8}$$

In（8）, OM_{01} and OM_{00} those omitted in the census can be neglected compared to the census enumeration. Therefore:

$$\frac{CE_{11} + WL_{11}}{CE_{10} + WL_{10}} \doteq \frac{CE_{11} + WL_{11} + II_{01} + NDD_{01}}{CE_{10} + WL_{10} + II_{00} + NDD_{00}} \tag{9}$$

According to the equation（5）, there is

$$\frac{CE_{11} + WL_{11}}{CE_{11} + WL_{11} + CE_{10} + WL_{10}}$$

$$= \frac{CE_{11} + WL_{11} + II_{01} + NDD_{01}}{CE_{10} + WL_{10} + II_{00} + NDD_{00} + CE_{11} + WL_{11} + II_{01} + NDD_{01}} \tag{10}$$

Assumption 2, the proportion of the total True Population correctly enumerated in the census equals the proportion of the P-sample enumerated in the census.

With the form of (6)

In table 3, $$\frac{M}{N_{10}} = \frac{N_{CP}}{N_{10} + N_{00}} \tag{11}$$

In table 4,

$$\frac{CE_{11} + WL_{11}}{CE_{11} + WL_{11} + CE_{10} + WL_{10}} = \frac{CE_{11} + WL_{11} + II_{01} + NDD_{01} + OM_{01}}{N} \tag{12}$$

Substituting（10）in（12）, therefore

$$\frac{CE_{10}+WL_{10}+II_{00}+NDD_{00}+CE_{11}+WL_{11}+II_{01}+NDD_{01}}{N}$$
$$=\frac{CE_{11}+WL_{11}+II_{01}+NDD_{01}}{CE_{11}+WL_{11}+II_{01}+NDD_{01}+OM_{01}} \quad (13)$$

5. Conclusions

According to (13), that $N=\frac{CE+WL+II_{01}+II_{00}+NDD_{01}+NDD_{00}}{CE_{11}+WL_{11}+II_{01}+NDD_{01}}$

$$\times(CE_{11}+WL_{11}+II_{01}+NDD_{01}+OM_{01}) \quad (14)$$

Substituting（=(10）in (14)，therefore

$$N=(CE+WL)\frac{CE_{11}+WL_{11}+II_{01}+NDD_{01}+OM_{01}}{CE_{11}+WL_{11}} \quad (15)$$

(15) is the total population of (4)，which demonstrated the accuracy of matching properties in expanded DSE model

This paper not only proves two important matching assumptions through expanded DSE model and its matching records, but also demonstrates the measurement error influence factors of CCE.

References

[1] Marks E S. The Role of Dual System Estimation in Census Evaluation[A]. Developments in Dual System Estimation of Population Size and Growth[C]. Edmonton: University of Alberta Press, 1978.156-188.

[2] Wolter K M. Coverage Error Models for Census and Survey Data[J]. Bulletion of the International Statistical Institute, 1983:415-431.

[3] Hogan H. The 1990 Post-Enumeration Survey: An Overview[J].The American Statistician, 1992,46（4）:261-269.

[4] Hogan H. The Accuracy and Coverage Evaluation: Theory and Design[J]. Survey Methodology, 2003,29（2）:129-138.

[5] U.S. Census Bureau. Accuracy and Coverage Evaluation of Census 2000: Design and methodology[R]. Issued September 2004.

[6] National Research Council. Coverage Measurement in the 2010 Census[M].Washington, DC: The National Academies Press.2009.

[7] Wolter K M. Coverage Error Models for Census Data. Journal of the American Statistical Association, 1981,81:338-346.

[8] Childers D. Accuracy and Coverage Evaluation: The Design Document. DSSD Census 2000 Procedures and Operations Memorandum Series, Chapter S-DT-1 (Revised).2001.

[9] Mulry M H and Kostanich D K. Framework for Census Coverage Error Components[J]. American Statistical Association Proceedings of the Joint Statistical Meeting, 2006:3461-3468.