THE EGYPTION POPULATION AND
FAMILY PLANNING REVIEW
ISSR,CAIRO UNIV. Vol.37,No.1,2004

Calibration Estimators of The Population Parameters Using Stratified Random Sample

By: *Mounira A. Hussein**

I. Introduction:

One of the fundamental question in sampling is how to effectively use the complete auxiliary information in the estimating stage. In sample surveys recent advance literature in estimation using auxiliary information can be classified to three main approaches according to the methodology used in estimating stage. These approaches are design based approach, model based approach and model assisted approach (Wu and sitter, 2001). The design based approach judges estimators with reference to the sampling distribution. The model based approach used a model for the variable of interest. If the model is not satisfied, the estimator will not be efficient. The model assisted approach estimators are approximately (asymptotically) design unbiased irrespective of whether the working model is correct or not and are particularly efficient if the working model is correct. In this approach inferences and the asymptotic framework are design based with the working model is only used to increase the efficiency.

Stratification is a common technique to increase the precision of the finite population estimators (Hussein, 1999). It is suggested when it is possible to divide a heterogeneous population into homogenous subpopulations. It is useful when the data of known precision are wanted or when sampling problems differ markedly in different parts of the population (Cochran, 1977). In this paper, we consider the use of more complex model in obtaining model assisted estimators for stratified random sample. The proposed model-calibration estimators can handle any linear or non linear model and reduce to the conventional calibration estimators of Deville and Samdal (1992) in the linear model case. The conventional calibration estimator of Deville and Samdal can be presented as follow.

---

* Associate professor, faculty of commerce, Menoufia University, Shebein Elkom, Egypt.

Consider a finite population consisting of N identifiable units. Associated with the i^{th} unit the study variable y_i and a vector of auxiliary variables, X_i . the values $X_1, X_2, \ldots X_p$ are known for the entire population but y is known only of i^{th} unit is selected in the sample. Deville and Sarndal (1992) introduce the following calibration estimator for the total.

$$\hat{Y}_C = \hat{Y}_{HT} + (X - \hat{X}_{HT})'\hat{B}$$

where

$$\hat{X}_{HT} = \sum_{i \epsilon S} di\, Xi$$

where d_i are basic design weights

$$\hat{B} = \left\{ \sum_{i \epsilon S} di\, q_i X_i X_i^{'} \right\}^{-1} \sum_{i \epsilon S} d_i q_i X_i y_i$$

The uniform weights $q_i = 1$ are used in most applications but unequal weights can also be motivated.

This model is suggested only when relationship between y and x is linear. If a curved relationship exists between y and x, the so constructed calibration estimator could be very inefficient. Sitter and Wu (2001) proposed a unified model- assisted framework. This framework proposed a new model calibration estimators can handle any linear or non linear working models and reduce to the conventional calibration estimators of Deville and Sarndal in the linear model case. The unified framework can also be used with any sampling design. This framework can be presented as follow.

Assume the relationship between y and X can be described by a super population model through the first and second moment.

$$E(y|\ X_i) = \mu(X_i, \theta)$$

$$V(y_i|\ X_i) = V_i^2 \sigma^2 \qquad i = 1, \ldots N$$

where $\theta = (\theta_0, \ldots \theta_p)'$ and σ^2 are unknown super population parameters. $\mu(X, \theta)$ is a known function of X and θ, the V_i is a known function of X_i. E and V denote the expectation and variance with respect to the super

population model. It is also assumed that $(y_1, X_1), \ldots, (y_N, X_N)$ are mutually independent.

Under this unified framework, auxiliary information should be used through the fitted values assuming any model, linear, curved, exponential, this can be done for any number of the auxiliary variables. Sitter and Wu (2001) suggested the following two model-calibration estimators for the mean which treated the fitted values as one auxiliary variable as follows:

$$(1) \qquad \hat{\bar{Y}}_{MC} = \bar{Y}_{HT} + \left\{ N^{-1} \sum_{i=1}^{N} \hat{\mu}_i - N^{-1} \sum d_i \hat{\mu}_i \right\} \hat{B}_N$$

where

$$\hat{B}_N = \sum_{i \in S} d_i q_i (\hat{\mu}_i - \bar{\mu})(y_i - \bar{y}) / \sum d_i q_i (\hat{\mu}_i - \bar{\mu})^2$$

$$\bar{y} = \sum_{i \in S} d_i q_i y_i / \sum d_i q_i$$

$$\bar{\mu} = \sum_{i \in S} d_i q_i \hat{\mu}_i / \sum d_i q_i$$

$$\bar{Y}_{HT} = N^{-1} \sum d_i y_i$$

$$(2) \qquad \hat{\bar{Y}}^{*}_{MC} = \hat{\bar{Y}}_{HT} + \left\{ N^{-1} \sum_{i=1}^{N} \hat{\mu}_i - N^{-1} \sum d_i \hat{\mu}_i \right\} \hat{B}^{*}_N$$

where

$$\hat{B}^{*}_N = \sum_{i \in S} d_i q_i \hat{\mu}_i / \sum_{i \in S} d_i q_i \hat{\mu}_i^2$$

Both $\hat{\bar{Y}}_{MC}$ and $\hat{\bar{Y}}_{MC}$ are model-assisted and can handle any linear or non linear models. That is they are both design – consistent irrespective of whether the model holds and particularly efficient if the model is correct.

II. Study Objectives:

This paper is organized to accomplish the following objectives:-

1. Developing the mathematical formula of two model-assisted estimators for the mean and total . These estimators can handle any linear or non-linear model using stratified sample.

2. Constructing the mathematical formula of their mean square error and comparing their mean square error with other estimators cited in the introduction.
3. Using the new proposed estimators in estimating the total fertility rate in developed and underdeveloped countries.

III Calibration Estimators Under Simple Random Sample:

Before proposing the calibration estimator for the stratified random sample, it is helpful to develop the estimator first for simple random sample as follows:

First, we start with Deville and Sarndall calibration estimator as follows:

$$\because \hat{Y}_C = \hat{Y}_{HT} + (X - \hat{X}_{HT})'\hat{B}$$

and $$\hat{X}_{HT} = \sum_{i \in S} d_i X_i$$

$$\hat{Y}_{HT} = \sum_{i \in S} d_i y_i$$

$$\hat{B} = \left[\sum_{i \in S} d_i q_i X_i X_i^1\right]^{-1} \sum d_i q_i X_i y_i$$

Using the uniform weights $q_i = 1$ which used in most applications. Also, under simple random sample the design weights $d_i = \frac{1}{\pi_i}$ where $\pi_i = P_r(i \in \delta) = \frac{n}{N}$ where n is the sample size and N is the population size.

Plugging for d_i and q_i we get the following.

$$\hat{X}_{HT} = \sum_{i \in s} \frac{N}{n} X_i = N \sum_{i \in s} X_i / n = N\overline{\chi} = \hat{X}$$

$$\hat{Y}_{HT} = \sum_{i \in s} d_i Y_i = N \sum_{i \in \delta} Y_i / n = N\bar{y} = \hat{Y}$$

$$\therefore \hat{B} = \left[\sum_{i\epsilon s} \frac{N}{n} X_i X_i^{l}\right]^{-1} \sum \frac{N}{n} X_i y_i$$

$$\therefore \hat{B} = \left[\sum_{i\epsilon s} X_i X_i^{l}\right]^{-1} \sum X_i y_i$$

So, under simple random sample, Deville and Sarndall calibration estimator for the total population take the following mathematical formula

$$\hat{Y}_C = \hat{Y} + (X - \overline{X})\hat{B}$$

where $\hat{Y} = N\bar{y}$ $\hat{X} = N\bar{x}$

$$\hat{B} = \left[\sum_{i\epsilon\delta} X_i X_i^{l}\right]^{-1} \sum X_i y_i$$

which is the simple linear regression estimator.

Under the unified framework introduced by Wu and Sitter (2001), the information about the auxiliary variable should be used through the fitted values $\mu(X_i,\theta)$ i=1, ...,N

Assume we have a vector of auxiliary variables X_i and assume any appropriate model explain the relationship between the dependent variable y and the auxiliary variables $X_1, X_2, \ldots X_p$. This model can be a linear or non-linear model. The fitted values can be treated as one auxiliary variable and can be used to construct calibration estimators for the simple and stratified sample as follows:

For simple random sample

$$\hat{Y}_C = \hat{Y} + (u - \hat{u})\hat{B}$$

where $\hat{Y} = N\bar{y}$ $\hat{u} = N\bar{u}$

$\bar{y}$ and $\bar{u}$ are the sample means of the dependent variable and predicted values respectively

IV Calibration Estimators Under Stratified Sample:

(a) The separate calibration estimator:

For stratified sample, two type of regression estimators of the total can be formulated from the fitted values a follows.

$$\hat{Y}_{Ch} = \hat{Y}_h + (U_h - \hat{U}_h)\hat{B}_h$$

$$\hat{Y}_{CS} = \sum_{h=1}^{L} W_h \hat{Y}_{ch}$$

where

$$\hat{B}_h = \left[\sum_{i=1}^{n_h} U_i U_i^{!}\right]^{-1} \sum_{i=1}^{n_h} U_i$$

This estimator is appropriate when it is thought that the true regression coefficients B_h vary from stratum to stratum.

(b) The combined calibration estimator:

When we can assume that the regression coefficients does not change from stratum to stratum, the following estimator is more appropriate and can be formulated from the fitted values as follows:

$$\hat{Y}_{St} = \sum_{h} W_h \hat{Y}_h$$

$$\hat{U}_{st} = \sum_{h} W_h U$$

$$\hat{Y}_{cc} = \hat{Y}_{st} + (U - \hat{U}_{st})\hat{B}$$

where

$$\hat{B} = \left[\sum_{1}^{n} U_i U_i^{'}\right]^{-1} \sum_{1}^{n} U_i$$

V The Mean Square Error of the Calibration Estimators Under Stratified Random Sample.

In the model assisted approach, inferences and the asymptotic framework are design based with working model is only used to increase the efficiency. So the mean square error for the calibration estimator under stratified random sample can be constructed as the Same as the mean square error of the regression estimator in stratified sample. The only exception is treating the predicted value as one auxiliary variable as follows.

$$M(\hat{Y}_{Ch}) = V(\hat{Y}_{ch}) = \frac{N_n - n_h}{n_n N_n}\left(S_{yh}^2 - 2b_h S_{yuh} + b_h^2 S_{uh}^2\right)$$

$$\therefore M(\hat{Y}_{Cs}) = \sum_{h=1}^{L} \frac{N_h^2}{N^2}\left(\frac{N_n - n_h}{n_n N_n}\right)\left(S_{yh}^2 - 2b_h S_{yuh} + b_h^2 S_{uh}^2\right)$$

where b_h is the stratum regression coefficients.

VI Numerical Example:

In this example, the new proposed estimators are used in estimating total fertility rates. Data for 185 countries all over the world is considered (United Nations, 2000). The special characteristics of birth statistics suggest many of the variables that are useful in the analysis. The variable of prime importance is the age of the child's mother. Two variables are included in the analysis, percentage of births to women under age 20 and percentage of births to women age 35 years or over. Some characteristics of interest in the analysis of natality relate to the place of occurrence in terms of urban- rural residence. The variable percentage urban is included in the analysis to measure socio-economic and place of residence difference in natality. The following table illustrate some descriptive statistics for the variables included in the analysis.

Table (1)

Descriptive Statistics for the Variables Included in the Analysis

The Variable	N	Minimum	Maximum	Mean	Standard deviation
Total fertility	185	1.24	8.80	3.7843	1.89847
Percentage Urban	185	5.66	100.00	53.6838	24.2079
Percentage of births (under 20)	185	1.00	27.00	11.5892	6.22092
Percentage of births (35 and over)	185	3.00	27.00	12.5838	4.71869

We used the total fertility rate as dependent variable and percentage of urban, percentage of births to women under age 20 and percentage of births to women 35 and over as auxiliary variables. Assuming the linear relation between the dependent variable and the auxiliary variables we get the following fitted model.

Table (2)

The analysis of variance of the regression model

Model	SS	F	MS	F	Sig.
Regression	485.609	3	161.870	165.717	.000
Residual	167.798	181	.977		
Total	662.407	184			

$R^2 = .733$

Table (3)

Model coefficients and heir significant level

Model	Coefficients	Sig.
Constant	.364	.357
percentage urban	-.02079	.000
Percentage of births (women under 20)	.127	.000
Percentage of births (35 and over)	.243	.000

Assuming the model is appropriate, we calculated the predicted values for each unit of the population. Table (4) illustrates the observed and the predicted value for all the countries included in the analysis.

Table (4)

The observed value and the predicted values for each unit ordered descending according to the first variable

	Observed	Predicted		Observed	Predicted
1	8.80	3.91693	35	5.89	5.39245
2	7.60	6.64651	36	5.80	5.42407
3	7.40	6.18165	37	5.80	5.06642
4	7.20	6.49895	38	5.79	5.98100
5	7.20	6.17049	39	5.70	5.36358
6	7.20	3.60037	40	5.70	5.64495
7	7.10	5.67495	41	5.70	6.02045
8	7.10	6.32116	42	5.70	4.79956
9	7.10	5.99025	43	5.60	5.57469
10	7.00	5.36073	44	5.57	3.16791
11	7.00	5.89992	45	5.51	4.68948
12	7.00	5.72243	46	5.43	5.55803
13	6.90	5.01305	47	5.40	5.32929
14	6.80	6.92004	48	5.40	5.22846
15	6.80	5.88011	49	5.39	5.18819
16	6.80	6.64547	50	5.36	4.87990
17	6.70	5.97496	51	5.30	5.73476
18	6.69	6.27643	52	5.30	3.94566
19	6.58	6.48596	53	5.25	5.64855
20	6.55	5.38814	54	5.20	5.39025
21	6.50	6.42089	55	5.20	5.64409
22	6.50	5.62603	56	5.05	7.08854
23	6.45	5.82909	57	5.00	6.36059
24	6.39	4.25260	58	5.00	5.95155
25	6.37	6.87079	59	4.92	4.89844
26	6.30	4.99870	60	4.90	6.37373
27	6.29	5.19447	61	4.86	5.13846
28	6.10	5.74443	62	4.85	4.42260
29	6.06	5.75544	63	4.80	4.51402
30	6.00	4.56297	64	4.79	5.57279
31	5.90	5.04110	65	4.77	4.72573
32	5.89	5.41972	66	4.70	3.33476
33	5.89	5.85533	67	4.68	4.89919
34	5.80	5.68314	68	4.55	5.05244

Table (4) (continued)

	Observed	Predicted		Observed	Predicted
69	4.45	3.47397	107	2.92	3.18922
70	4.41	4.31001	108	2.90	4.56293
71	4.30	4.93680	109	2.88	3.68149
72	4.30	2.99930	110	2.85	2.39805
73	4.20	5.90017	111	2.83	3.59547
74	4.18	3.88329	112	2.74	3.27093
75	4.10	4.74685	113	2.70	2.64711
76	4.10	3.13325	114	2.70	3.65836
77	4.00	4.13236	115	2.68	2.45238
78	3.95	2.60842	116	2.61	3.97616
79	3.88	3.92997	117	2.60	1.41433
80	3.83	2.21307	118	2.55	3.10763
81	3.80	3.86236	119	2.54	2.20307
82	3.80	4.02944	120	2.51	2.32662
83	3.75	4.54319	121	2.44	3.08168
84	3.62	3.28207	122	2.43	2.28613
85	3.60	4.47132	123	2.38	2.18683
86	3.58	2.75882	124	2.35	3.34619
87	3.56	2.53202	125	2.33	3.06485
88	3.52	4.44443	126	2.29	3.20751
89	3.50	4.65328	127	2.20	4.20408
90	3.43	4.05652	128	2.20	3.58234
91	3.40	5.36066	129	2.19	2.37807
92	3.40	3.40383	130	2.18	2.82843
93	3.40	4.08886	131	2.17	1.67822
94	3.39	3.80814	132	2.15	2.74748
95	3.35	3.40559	133	2.14	2.46860
96	3.29	3.66625	134	2.12	2.02264
97	3.25	3.82413	135	2.10	-.05327
98	3.14	4.17111	136	2.10	2.37661
99	3.12	4.04805	137	2.10	1.84791
100	3.10	2.51995	138	2.08	2.28578
101	3.09	2.46206	139	2.05	2.34211
102	3.09	3.55114	140	2.05	2.86128
103	3.05	3.15742	141	2.01	4.17118
104	3.00	3.68458	142	2.01	1.92915
105	2.98	3.46208	143	1.95	3.01636
106	2.93	2.62587	144	1.94	4.04161

Table (4) (Continued)

	Observed	Predicted		Observed	Predicted
145	1.93	2.52785	166	1.64	2.06504
146	1.92	.59097	167	1.62	1.16659
147	1.89	2.24222	168	1.60	1.83379
148	1.89	2.15065	169	1.60	2.17629
149	1.88	1.77776	170	1.59	1.67985
150	1.85	3.28239	171	1.58	2.06005
151	1.83	2.82878	172	1.53	2.34546
152	1.79	1.69310	173	1.53	2.26569
153	1.78	1.69943	174	1.53	2.37617
154	1.78	1.95300	175	1.52	2.82555
155	1.75	1.51547	176	1.50	2.69472
156	1.79	1.95934	177	1.50	2.35360
157	1.73	3.34576	178	1.48	1.05194
158	1.70	1.97713	179	1.47	1.73099
159	1.6	2.00499	180	1.38	2.07935
160	1.68	1.75633	181	1.36	2.02545
161	1.67	1.60329	182	1.32	2.03938
162	1.66	1.31979	183	1.30	1.50065
163	1.65	1.47050	184	1.27	1.95212
164	1.65	1.80545	185	1.24	2.27154
165	1.64	2.01582			

A simple random sample of size n=99 is drown from the total population. The calibrated estimator and its mean square error is calculated .

A stratified random sample of two strata (n_1=53 , n_2= 46) is drown from the total population. The calibration estimators for each strata and their mean square errors are calculated . The two stratified estimators and their mean square errors are calculated. Table (5) illustrates some descriptive statistics for simple random sample, stratified random sample for the observed and predicted values.

Table (5)

Simple random sample			Stratified random sample			
	Mean	Standard Deviation	First stratum		Second stratum	
			Mean	Standard Deviation	Mean	Standard Deviation
Observed	3.8398	1.61871	5.0047	1.09718	2.1652	.53599
Predicted	3.8482	1.95126	5.4096	1.34818	2.4683	.78936

Table (6)
Relation Efficiency of the Calibration Estimators From the Simple Random Sample

Method of estimation	The estimate (TFR)	The mean square error	The relative efficiency
1. Simple random sample (conventional estimator)	3.8482	.0123035	100
2. Simple random sample (conventional estimator for first stratum)	5.4096	.0111382	100
3. Simple random sample (conventional estimator for second stratum)	2.4683	.0026957	100
4. Simple random sample (calibration estimator)	3.7818	.0047029	261.6
5. Simple random sample (calibration estimator for first stratum)	4.6616	.0087805	126.8
6. Simple random sample (calibration estimator for second stratum)	2.159087	.0017459	159.9

Table (7)
Relation Efficiency of the Calibration Estimators From the Stratified Random Sample

Method of estimation	The estimate	The mean square error	The relative efficiency
1. Stratified random sample (conventional estimator)	3.76359	.004036	100
2. Stratified random sample (calibration combined estimator)	3.4079	.005426	74.9
3. Stratified random sample (calibration separate estimator)	3.5658	.003109	129.8

VII Conclusion:

Table (6) illustrates the gain in precision from using the calibration estimator in simple random sample. The calibration estimator is more precise than the mean per element for the entire population and for each strata. Table (7) illustrates the gain in precision from using the calibration estimator in stratified sample. In this example the separate calibration estimator is more efficient than the combined calibration estimator. This result is acceptable since b_n different from the first stratum to the second stratum where b_1=.657 and b_2=.403 respectively. In this approach the model assisted approach, the estimate can be improved by using more appropriate model. This example illustrates using the calibration estimator as an example for model assisted approach in addition to using stratification as a common technique for increasing the precision.

VIII Reference

- Baker, S. G. (2000). *"Analyzing a Randomized Cancer Prevention Trial With a Missing Binary Outcome, an Auxiliary variable, and All or Non Compliance"* Journal of the American Statistical Association, Vol.95, No.449.

- Boehmke, J. F. (2003). *"Using Auxiliary Data to Estimate Selection Bias Model"*. University of Iowa, USA.

- Chang, S. G. (2000). *"The Asymptotic Distribution of Multivariate Product Estimator"*. Chinese journal of Mathematics, Vol.14. No.3.

- Cochran, W.G. (1977): "*Sampling Techniques*". 3rd Edition, John Wiley, New York.

- Deville. J. C. and Samdal C. E. (1992), *" Calibration Estimators is Survey Sampling"* Journal of the American Statistical Association, 87. 376-382.

- Dorfman A. H. (1994). *"A Note on Variance Estimation for the regression Estimator in Double Sampling"* Journal of the American Statistical Association, Vol.98, No.425.

- Hussein, M. A. (1988). *"On Increasing the Precision of Finite Population Estimators Using Auxiliary Variables"*. Unpublished Ph. D. Thesis. University of Iowa, USA.

- Hussein, M. A. (1992). *"Alternative Estimators for Stratified Random Sampling"*. The Egyptian Statistical Journal, ISSR, Cairo University, Vol.35, No.2.

- Hussein, M. A. (1995). *An Estimator in Cluster Sample Combined with Stratification"*. The Egyptian Statistical Journal, ISSR, Cairo University, Vol.46, No.2.

- Hussein, M. A. (1998). *"Estimating Finite Population Parameters Using Stratified Sample in The Presence of The Auxiliary Information"*. The Egyptian Population and Family Planning review Journal, ISSR, Cairo University, Vol.42, No.2.

- Hussein, M. A. (1999). *"Separate and Combined Ratio Estimators for The Median".* The Egyptian Population and Family Planning review Journal, ISSR, Cairo University, Vol.43, No.2.

- Hussein, M. A. (2001). *"Median Estimation using Auxiliary Information".* The Egyptian Population and Family Planning review Journal, ISSR, Cairo University, Vol.43, No.1.

- Hussein, M. A. (2002). *"Multivariate Product Type of Estimator and Its Applications in Family Planning Programs".* The Egyptian Population and Family Planning Review Journal, ISSR, Cairo University, Vol.35, No.2.

- Nove, A. (2001) *"Using Weights to Adjust or Sample Selection When Auxiliary Information is Available"* NBER, Working Paper No. t.275.

- Ouyang Z., Srivastya J. N. and Schreuder, H. T. (1992). *"IKA General Ratio Estimator and Its Applications in Model Based Inferences".* Ann Inst, Statis. Math., Vol.45, No.1.

- Sekkppan R. M. (1986). *"Estimation in Sampling from Finite Populations Under the General Linear Regression Model"* Journal of Indian Statistical Association , Vol.24.

- Srvastava, j. N. and Ouyang Z. (1992). *"Some Properties of a General Estimator in Finite Population Sampling".* Sankhya, The Indian Journal of statistical, vol.54.

- Sweet, E. M. and Sigmen, R. S. (1995).*"Evaluation of Model Assisted Procedures for Stratifying Skewed Population Using Auxiliary Data".*

- Sitter R. and WUC (2001), *"A Model Calibration Approach to Using Complete Auxiliary Information from Survey Data".* Journal of The American Statistical Association, Vol. 96, No.453, Theory and Methods.

- U. N. (2000). "World Population Monitoring 1998: Health and Mortality Selected Aspects". New York.