JACKKNIFING SURVEY DATA

Marietta P. Morada
and
Gloria A. Cubinar*

ABSTRACT

The study draws a 50-percent sample from establishments with twenty to forty nine average total employment (ATE) and engaged in wholesale and retail trade in the Metropolitan Manila in 1987. To come up with more refined estimates of revenue and cost, the ratio estimator is used with ATE as the auxiliary variable. Biases are estimated through the jackknife procedure. Results of the study highlight the reduction in the variances when compared to simple random estimates.

Keywords and Phrases: Bias, jackknifing, ratio estimator, stratified sampling.

1. Introduction

The Annual Survey of Establishments in the Philippines (ASE) is a nationwide survey covering large establishments in operation during the reference year. It is one of the major sources of comprehensive statistics on the structure, levels and trends of economic activities in the country and in each region. Specific items of data collected on establishment level include employment, compensation, revenues, costs, fixed assets, capital expenditures and inventories.

The survey stratifies the population by region, industrial classification (according to the 1977 Philippine Standard Industrial Classification or PSIC) and size. The latter is determined from the average employment or total number of persons engaged (ATE). For the 1987 round of surveys, covered are establishments with ATE of 10 or more. Sampling, however, is applied only for establishments with ATE of 10-19 persons while establishments with ATE 20 and over are covered on a 100-percent basis. Included in the latter group are establishments with average monthly sales/revenue (AMS/R) of at least a million, irrespective of ATE.

For the wholesale and retail trade sector, however, establishments with ATE of 20-49 persons still number to more than a thousand in 1987, representing about 70 percent of the total sample size for the sector. As may be expected, more than half of this number are in the Metropolitan Manila, where nonresponse and other nonsampling errors are most likely to creep in if only due to the large number of establishments in the region that are covered in the survey.

This paper aims to show that when sampling is used, increase in precision of the estimates may be gained

* Supervising Statistical Coordinator and Senior Statistician of the National Statistics Office, respectively.

through the use of the ratio estimator with its bias minimized using jackknife procedure.

2. Data and Methods

The study uses the results of the 1987 Annual Survey of Establishments for the wholesale and retail trade sector. Variables considered are ATE, revenue and cost. For purposes of the paper, excluded are establishments filing consolidataed reports for they present problems which cannot be addressed by improvement in the estimation procedure.

The wholesale and retail trade sector has specifically been selected for the study due to the expected variations in the relationship between ATE on the one hand, and cost and revenue on the other. As has previously been observed, revenue due to wholesaling activities is not very much dependent on ATE while direct relationship between these two variables is expected among establishments engaged in retailing. With this type of data, the characteristic of the estimation procedure can be displayed when it is properly used and when conditions for its applicability is violated.

2.1 Sample Selection

Defining the set of all wholesale and retail trade establishments with 20-49 ATE as the population and stratifying it according to the 3-digit industrial classification, the following notations are used:

N_h = total number of establishments in the h^{th} industrial classification;

n_h = corresponding sample size; and

y_{hi} = value of the y-variable for the i^{th} establishment in the hth industrial classification.

The sample size n_h is derived as follows:

$$n_h = \begin{cases} N_h/2 & ; \text{ if } N_h/2 \text{ is an integer} \\ [N_h/2] + 1 & ; \text{ if } N_h/2 \text{ is not an integer.} \end{cases}$$

Here, $[N_h/2]$ is the largest integer less than or equal to $N_h/2$.

Once n_h has been determined, the sample is drawn using simple random sampling without replacement.

For stratum with $N < 6$, no sampling is done since the number of samples would be too few. Minimum sample size allowed is 4.

2.2 Jackknife Estimate of the Population Mean and Total

The ratio estimator, though proven to be more efficient when the auxiliary variate is highly correlated with the variable of interest, is yet to be used in national surveys. This could probably be due to the fact that large sample sizes almost assure the statisticians of the validity of the sample mean.

Two factors prove to be deterrent to the popularity of the ratio estimator, namely:

1. it is unwieldy since the values of the auxiliary variate must be known for the entire population; and

2. it is biased of order 1/n.

Note however that for large n, the bias becomes negligible.

Since ATE is known from the frame, its population parameters are readily available. Thus the ratio estimate of the mean of a stratum may be used, i.e.,

$\bar{y}_{hR} = (\bar{y}_h/\bar{x}_h)\bar{X}_h$; where

$\bar{y}_h$ = stratum mean of the y-values of units in the sample.

$\bar{x}_h$ = corresponding stratum mean of the x-values, and

$\bar{X}_h$ = stratum population mean.

The estimate of the population total is given by

$$Y_{hR} = N_h\bar{y}_{hR}.$$

The strata here are composed of establishments in the same 3-digit industrial classification in the Metropolitan Manila. It may be expected, therefore, that there would be strata with small sample sizes. The bias in such cases may not be negligible. In order to reduce the bias to order $1/n^2$ and to estimate the bias, the jackknife procedure is used.

Letting y_{hi} be the observed value of the y-variate for the i^{th} individual in the h^{th} stratum, define $y_{h(i)}$ as follows:

$$\bar{y}_{h(i)} = (1/(n_h-1)) \sum_{\substack{j=1 \\ j\neq i}}^{n_h} y_{hj} \; ;$$

$$i=1,\ldots,n_h$$

and

$$\bar{y}_{h(i)R} = (\bar{y}_{h(i)}/\bar{x}_{h(i)})\bar{X}_h;$$

$$i = 1,\ldots n_h$$

The jackknife estimate of the population mean is, therefore, given by

$$\bar{y}_{(.)R} = (1/n_h) \sum_{i=1}^{n_h} \bar{y}_{h(i)R};$$

and of the stratum total, by

$$Y_{hR} = N_h\bar{y}_{h(.)R} .$$

Quenouille's estimate of bias of the ratio estimator is given by (Efron, 1982:6)

$$BIAS = (n_h - 1)(\bar{y}_{h(.)R} - \bar{y}_{hR}).$$

Thus, the bias corrected estimate of the mean is

$$\bar{y}^*_{hR} = \bar{y}_{hR} - BIAS$$

$$= n_h\bar{y}_{hR} - (n_h - 1)\bar{y}_{h(.)R},$$

and the bias corrected estimate of the stratum total is

$$Y^*_{hR} = N_h\bar{y}^*_{hR}.$$

2.3 Variance of the bias-corrected estimate of the mean

The estimate of the variance of the ratio estimate of population total is (Cochran, 1977:155):

$$v(Y_{hR}) = (N_h^2(1-f_h)/n_h)(s_{yh}^2 + R_h s_{xh}^2 - 2R_h s_{x\,y\;h\;h}),$$

where $f_h = n_h/N_h$,

$$s_{yh}^2 = (1/(n_h-1)) \sum_{i=1}^{n_h} (y_{hi}-\bar{y}_h)^2,$$

$$s_{xh}^2 = (1/(n_h-1)) \sum_{i=1}^{n_h} (x_{hi}-\bar{x}_h)^2,$$

$$s_{x\,y\;h\;h} = (1/n_h - 1) \sum_{i=1}^{n_h} (y_{hi}-\bar{y}_h)(x_{hi}-\bar{x}_h),$$

$$R_h = \bar{y}_h/\bar{x}_h.$$

Hence, after correction for bias of the estimate, the variance reduces to

$$v(\hat{Y}^*_{hR}) = v(Y_{hR}) - BIAS^2$$

3. Findings

3.1 The Population

The population considered in the present study is composed of 655 wholesale and retail establishments in Metropolitan Manila employing an average of 20 to 49 number of persons in 1987. Of these establishments, 345 are in wholesaling while 310 are in retailing business.

The establishments in the first category are predominantly engaged in wholesaling of construction materials and supplies (614), machinery and equipment including transport equipment (616) and of products not classifiable in any definite group (619). This is understandable since the Metropolitan Manila remains to be the country's major source of construction materials and transport equipments. On the other hand, the distribution of retail establishments by sector is more even.

Table 1 presents the population parameters for employment, revenues and cost by sector. The nondependence of wholesaling activities on employment may be observed as values of revenue and cost within the employment range 20-49 vary widely. This, however, may not readily be observed among retail establishments among which the sectoral means do not vary as much.

The applicability of the ratio estimator hinges on the degree of correlation between the auxiliary and the main variable. As mentioned earlier and as is observed, low correlation between ATE and revenue or cost is expected among wholesaling establishments. However, contrary to expectation, the same is also true for retail establishments among which high correlation is expected to be observed.

3.2 Simple Random Sample Estimates

Table 2 gives the simple random sample estimates of means and the corresponding coefficients of variation. Among the wholesaling establishments, percent deviation from the population means remain to be within acceptable values except for sectors 616 and 618 where percent deviations of 37 and 26 percent may be observed. For sector 618, this may be explained by the sample size (n=4) which is rather small

considering that the population variance shown in Table 1 is very high.

Sector 616 is the second largest wholesaling sector (N=72) and it is surprising to observe a very large difference between the estimated mean and the population mean even with n=36. Examination of the data shows, however, that the deviation is mainly due to an outlier which was not drawn into the sample. Deleting the outlier, the estimated mean would be 15 percent of the true value.

Estimated means for retail sectors show erratic pattern. For some sectors (621, 624, 627 and 628), deviations from population means are less than 15 percent. However, for the remaining sectors, estimated means deviate from population means by about 20 percent for sector 625 to as much as about 35 percent for sectors 622 and 623. These observations are not as wide among retail sectors.

3.3 The Ratio Estimates

The ratio estimate of the form

$$\bar{y}_R = (\bar{y}/\bar{x})\,\bar{X}$$

departs from the s.r.s. estimates depending on the value of $\bar{X}/\bar{x}$. From table 3, $\bar{X}/\bar{x}$ for various sectors lies in the interval (0.9000, 1.1000); thus, leading to ratio estimates that are close to s.r.s. estimates.

However, due to large positive covariances between ATE and revenue and cost, substantial reduction may be noted in the approximate variances of ratio estimates when compared to variances of s.r.s. estimates, giving maximum c.v. of 0.44 (or 44%) for 612 as against the maximum s.r.s. estimate c.v. of 2.36 (or 236%). Moreover, c.v.'s for several sectors fall within acceptable level of 10 percent.

3.4 The Bias-Corrected Estimates

The biases, as estimated through the jackknife procedure, are subtracted from ratio estimates to arrive at the bias-corrected estimates presented in Table 4 for revenue and in Table 5 for cost. For both variables, the estimated biases are insignificant, falling within at most about 5 percent of the corresponding ratio estimates.

4. Conclusion

4.1 Data Quality

Results of the study point out several problems besetting the survey. If measurements may be assumed to be fairly accurate and reflective of business activity of the establishments, the ATE is a very weak stratification variable since within 20-49 ATE, wide variations in revenue and cost may be observed. In such a case, the sample mean, though unbiased, would expectedly be subject to large sampling variability. A reliable estimate may only be made by getting a very large sample. Considering the population size for each sector, a 100 percent coverage

appears to be necessary. This has a very relevant implication on the current sampling design of the QSE which assigns a 25% sampling rate to the stratum.

Data analysis, however, points to problems related to measurement. Considering that enterprises maintaining branches all over the country usually have their main units in the Metropolitan Manila, there may be reasons to suspect that reports submitted by some establishments actually contain transactions made in its branches. This could possibly be the reason why even after controlling for economic activity and ATE, some extreme value may still be observed.

Another source of variability may be the ambiguity in classification of economic activity. The recent years have witnessed adjustments made by trade establishments to cope up with the economic slump. For example, while maintaining a store as a front, some establishments engaged in wholesale/retail of auto supplies and accessories are now also engaged in manufacturing of spare parts or assembling transportation units such as jeepneys. Unreported diversification or change of economic activity may certainly lead to variations in reported revenue and cost in the sector to which the establishment has been initially classified.

Another phenomenon that is not adequately measured is the proliferation of sales persons on commission basis. Many of these sales persons visit offices and private houses selling various household items using receipts of well-known large wholesale/retail establishments. Although the survey intends to exclude these workers from the reported ATE, no serious study has actually been conducted on how establishments treat the employment status of such sales person. To what extent does an establishment make use of this type of arrangement? Is this the reason why some establishments report unusually high revenues despite low employment?

Another issue that needs addressing is the way establishments report part-time workers. Are they being counted individually or are establishments counting them in terms of mandays spent at work?

4.2 Applicability of the Estimation Technique

The study, despite the limitations imposed on it by the quality of the data used, points out the strong features and ease of use of ratio estimation and jackknifing technique. This is very important in as much as the ASE is a repetitive survey and uses a subsample of a census. The availability of auxiliary information is therefore never a problem for the use of the same variable measured during the census can serve as auxiliary information to the estimation of the value of current transactions.

5. References

Cocharan, W. G. 1977. Sampling Techniques. Canada: John Wiley & Sons, Inc.

David, I. P. and B. V. Sukhatme. 1974. "On the Bias and Mean Square Error of the Ratio Estimator", JASA, Vol. 69 (346): 464-466.

Efron, Bradley. 1982. The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Table 1. Population Parameters

	N	$\bar{X}$	$\bar{Y}$ Revenue	S^2
Wholesaling				
611	4	28.2	25,926,366	1.7×10^{15}
612	24	32.6	60,345,704	1.5×10^{16}
613	9	28.6	12,487,859	1.7×10^{14}
614	62	31.3	29,264,582	1.8×10^{16}
615	12	29.9	13,903,033	5.6×10^{13}
616	72	31.8	20,316,035	2.8×10^{15}
617	23	31.9	51,047,920	4.3×10^{15}
618	6	28.3	24,777,231	3.9×10^{14}
619	133	31.3	20,800,155	6.2×10^{14}
Retailing				
621	26	31.3	11,756,078	2.0×10^{14}
622	31	37.2	32,938,018	7.7×10^{14}
623	36	30.5	8,606,792	7.5×10^{13}
624	46	29.0	9,780,380	8.2×10^{13}
625	32	30.8	27,877,370	2.1×10^{15}
626	28	31.8	17,361,661	2.8×10^{14}
627	56	27.3	26,447,077	2.4×10^{14}
628	33	28.8	24,075,106	2.3×10^{14}
629	22	33.8	6,135,632	3.1×10^{13}
	N	$\bar{X}$	$\bar{Y}$ Cost	S^2
Wholesaling				
611	4	28.2	22,924,293	1.4×10^{15}
612	24	32.6	59,236,638	1.3×10^{16}
613	9	28.6	11,562,140	1.6×10^{14}
614	62	31.3	27,295,782	1.5×10^{15}
615	12	29.9	13,432,093	6.6×10^{13}
616	72	31.8	16,023,996	1.2×10^{15}
617	23	31.9	43,068,583	3.9×10^{15}
618	6	28.3	23,669,723	3.7×10^{14}
619	133	31.3	18,566,973	6.2×10^{14}
Retailing				
621	26	31.3	10,939,850	1.6×10^{14}
622	31	37.2	31,548,962	7.2×10^{14}
623	36	30.5	7,970,645	6.3×10^{13}
624	46	29.0	9,267,503	8.2×10^{13}
625	32	30.8	26,508,430	1.9×10^{15}
626	28	31.8	16,831,533	2.7×10^{14}
627	56	27.3	25,420,469	2.2×10^{14}
628	33	28.8	23,031,492	2.0×10^{14}
629	22	33.8	4,952,603	2.5×10^{13}

Table 2. Simple Random Sample Estimates

	$\bar{y}$ Revenue	c.v.*	$\bar{y}$ Cost	c.v.*
Wholesaling				
611	-	-	-	-
612	69,026,656	2.3576	65,028,243	2.2807
613	12,008,886	0.4352	11,076,598	0.4513
614	27,783,405	1.15445	27,963,800	1.2199
615	15,065,722	0.5455	14,337,989	0.5456
616	12,746,350	0.8382	11,145,608	0.8882
617	55,840,581	1.5241	46,752,233	1.7495
618	18,621,640	0.9454	17,632,361	0.9882
619	20,848,060	1.1941	18,501,555	1.3411
Retailing				
621	10,379,888	1.3120	9,282,182	1.2436
622	44,329,584	0.7444	42,978,932	0.7415
623	11,692,179	0.9521	10,519,117	0.9498
624	9,768,589	0.9836	9,086,397	1.0039
625	33,227,583	1.8147	32,014,004	0.5686
626	13,043,296	0.8951	12,968,403	0.9280
627	24,906,365	0.6197	23,827,951	0.6372
628	23,819,198	0.6040	22,390,349	0.5762
629	4,553,908	0.8490	3,528,681	0.8900

* c.v. = $s/\bar{y}$

Table 3. Ratio Estimates of Revenue and Cost

	$\bar{X}/\bar{x}$	$\bar{Y}_R$	$s_{y_R}^2$	c.v.
Wholesaling				
611	-	-	-	-
612	1.0966	75,697,299	1.1471×10^{15}	.4381
613	0.9211	11,061,949	1.5531×10^{12}	.1127
614	1.1066	30,745,489	1.5000×10^{13}	.1260
615	1.0200	15,365,324	7.0828×10^{12}	.1732
616	0.9845	12,548,902	1.5126×10^{12}	.0980
617	1.0638	59,401,429	3.2307×10^{14}	.3026
618	0.9290	17,298,791	2.0816×10^{13}	.2637
619	0.9977	21,135,057	4.3756×10^{12}	.0990
Retailing				
621	0.9819	10,192,300	7.3264×10^{12}	.2656
622	1.0216	45,288,631	2.4366×10^{13}	.1090
623	0.9910	11,586,654	2.7524×10^{12}	.1432
624	0.9024	8,815,556	2.0637×10^{12}	.1630
625	0.9788	31,502,304	1.0838×10^{14}	.3305
626	1.0125	13,206,337	5.7838×10^{12}	.1821
627	1.0214	25,438,410	3.3403×10^{12}	.0718
628	0.9740	22,005,940	2.1577×10^{12}	.0668
629	1.0013	4,560,045	5.7347×10^{11}	.1661

Table 3. Ratio Estimates of Revenue and Cost (concluded)

	$\bar{X}/\bar{x}$	$\bar{Y}_R$	$s_{y_R}^2$	c.v.
Wholesaling				
611	-	-	-	-
612	1.0966	71,312,485	9.5058x10^14	.4323
613	0.9211	10,203,174	1.2471x10^12	.1094
614	1.1066	30,945,118	1.7459x10^13	.0773
615	1.0200	14,623,119	6.1840x10^12	.1701
616	0.9845	10,972,956	1.2923x10^12	.1036
617	1.0638	49,733,535	2.9767x10^14	.3469
618	0.9290	16,379,788	2.1572x10^13	.2836
619	0.9977	18,459,042	4.4038x10^12	.1137
Retailing				
621	0.9819	9,114,432	5.3898x10^12	.2547
622	1.0216	43,908,759	2.3391x10^13	.1101
623	0.9910	10,424,179	2.2156x10^12	.1428
624	0.9024	8,199,919	1.8728x10^12	.1669
625	0.9786	30,351,738	9.8704x10^13	.3273
626	1.0125	13,130,508	6.3300x10^12	.1916
627	1.0214	24,336,959	3.3257x10^12	.0749
628	0.9740	21,807,625	1.8491x10^12	.0624
629	1.0013	3,533,436	3.6775x10^11	.1716

Table 4. Bias-Corrected Estimates of Revenue

	$\bar{y}^*_R$	$s^2_{y_R}$	c.v.
Wholesaling			
611	-	-	-
612	75,053,964	1.1467x10^15	.4512
613	11,030,701	1.5522x10^12	.1129
614	30,744,549	1.5000x10^13	.1260
615	15,190,804	7.0523x10^12	.1748
616	12,549,114	1.5126x10^12	.0980
617	59,480,761	3.2307x10^14	.3022
618	17,278,863	2.0816x10^13	.2640
619	20,799,099	4.3756x10^12	.1006
Retailing			
621	10,140,340	7.3237x10^12	.2669
622	45,420,406	2.4349x10^13	.1086
623	11,584,835	2.7524x10^12	.1432
624	8,791,928	2.0631x10^12	.1634
625	31,589,304	1.0834x10^14	.3295
626	13,097,670	5.7720x10^12	.1834
627	25,449,696	3.3402x10^12	.0718
628	23,297,544	2.1481x10^12	.0629
629	4,567,445	5.7341x10^11	.1658

Table 5. Bias-Corrected Estimates of Cost

	$\bar{y}^*_R$	$s^2_{y_R}$	c.v.
Wholesaling			
611	-	-	-
612	70,578,895	9.5004x10^14	.4367
613	10,197,626	1.2470x10^12	.1095
614	30,944,578	1.7459x10^13	.1350
615	14,473,234	6.1615x10^12	.1715
616	10,973,108	1.2923x10^12	.1036
617	48,839,609	2.9687x10^14	.3528
618	16,385,166	2.1572x10^13	.2835
619	18,457,458	4.4038x10^12	.1137
Retailing			
621	9,056,304	5.3864x10^12	.2563
622	44,018,709	2.3379x10^13	.1098
623	10,422,547	2.2156x10^12	.1428
624	8,176,445	1.8722x10^12	.1673
625	30,434,988	9.8697x10^13	.3264
626	13,002,640	6.3134x10^12	.1932
627	24,345,815	3.3256x10^12	.0749
628	21,880,713	1.8334x10^12	.0619
629	3,544,006	3.6760x10^11	.1711

Table 6. Correlation Coefficients of Employment with Revenue and Cost by Stratum

PSIC	N	Correlation: ATE and Revenue	Correlation: ATE and Cost
611	4	.147	.159
612	24	-.044	-.238
613	9	.582	.592
614	62	.392	.360
615	12	-.471	-.446
616	72	.103	.133
617	23	-.117	-.124
618	6	-.018	-.049
619	133	.163	.123
621	26	.169	.148
622	31	.413	.388
623	36	.391	.401
624	46	-.016	-.039
625	32	.170	.174
626	28	-.058	-.098
627	56	.283	.260
628	33	.663	.626
629	22	.172	.207

Appendix A. Three-Digit Classifications of Establishments Engaged in Wholesale and Retail

Wholesaling

611 - Farm, forest and marine products
612 - Processed food, beverages and tobacco products
613 - Dry goods, textiles and wearing apparel
614 - Construction materials and supplies
615 - Office and household furniture, furnishings, and appliances and wares
616 - Machinery and equipment including transport equipment
617 - Minerals, metals and industrial chemicals except crude petroleum and petroleum products
618 - Petroleum and petroleum products
619 - Wholesaling, n.e.c.

Retailing

621 - Books, office, school supplies, including newspaper and magazines
622 - Food, beverages and tobacco
623 - Dry goods, textile and wearing apparel
624 - Construction materials and supplies
625 - Office, household furniture and furnishings, fixtures, appliances and wares
626 - Transportation, machinery and equipment, accessories and supplies
627 - Medical supplies and equipment stores
628 - Petroleum and other fuel products
629 - Retailing. n.e.c.