Effects of Imperfect Unit Size Information on Complex Sample Designs and Estimators

Randall K. Powers and John L. Eltinge

Office of Survey Methods Research, U.S. Bureau of Labor Statistics

Powers.Randall@bls.gov

Abstract: Work with sample surveys often makes extensive use of measures of size. Two prominent examples are the use of "probability proportional to size" sampling; and use of size measures in adjustment of survey weights through, e.g., ratio estimation, post-stratification or calibration weighting. However, many survey applications use size variables that are imperfect approximations to the idealized size measures that would produce optimal efficiency results. This paper explores the effects that alternative size measures may have on the efficiency of some standard design-estimator pairs. Principal emphasis is placed on numerical results of a simulation study that uses size measures and economic variables available through the Quarterly Census of Employment and Wages of the Bureau of Labor Statistics.

Key words: measures of size; ratio estimation; regression estimation; sampling with probabilities proportional to size; unequal-probability sampling.

1. Introduction

Large-scale sample surveys often use auxiliary information in an effort to improve the efficiency of the procedure defined by a given (design, estimator) pair. However, the available auxiliary information is often imperfect, and it is of interest to study the extent to which imperfections in that information may lead to problems with the proposed procedure. For example, Clark (2013, 2014) and others have studied the effects of imperfect population information on the properties of stratified sample allocation methods.

In many cases, sample surveys also use unequal-probability designs in which selection probabilities are proportional to a measure of unit size that are available through the frame (i.e., the list of sample units). Under standard conditions (e.g., Cochran, 1977), the resulting "probability proportional to size" (pps) designs are more efficient than equal-probability designs. However, it is generally recognized that imperfections in the unit size information can lead to degradation in the performance of the resulting pps design. Powers and Eltinge (2013) used a simulation study to explore this issue through the following steps. First, consider a population of size N. For a given unit i, one has a auxiliary variable X_i available for all population units, and a survey variable Y_i which

one intends to collect from each unit selected for the sample. Under conditions, the optimal "size" measure to use in a probability-proportional-to-size design is

$$s_i = \{[\mu(X_i)]^2 + [\sigma(X_i)]^2\}^{1/2} \quad (1.1)$$

where $\mu(X_i)$ and $[\sigma(X_i)]^2$ are the conditional mean and variance, respectively, of Y_i given X_i. For general background on probability-proportional-to-size designs, see, e.g., Cochran (1977, Section 9A.3), Godambe (1955, 1982), Brewer (1963), Thomsen et al (1986), Kott and Bailey (2000), Holmberg and Swensson (2001) and references cited therein.

Second, Powers and Eltinge (2013) applied the general idea of a size measure (1.1) to data from the BLS Quarterly Census of Employment and Wages (QCEW) in specified industries within a given state for one year. For a given unit i, primary attention focused on e_{1i}, the employment count from the first quarter of the year; and on y_{1i}, y_{2i}, y_{3i} and y_{4i}, total wages paid in the first through fourth quarters, respectively. Five size measures were considered. The measure that assigned a size of one to each unit was labeled (1). The remaining size measures all used expression (1.1), but with different choices for the mean and variance function. A second measure, labeled (a), was based on a mean function computed from the simple linear regression of y_2 on y_1:

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \varepsilon_{y2y1i} \quad (1.2)$$

and a variance function model based on regression of the squared residuals fro model (1.2) onto the associated predicted value, also computed from model (1.2):

$$\left(\hat{\varepsilon}_{y2y1i}\right)^2 = \gamma_0 + \gamma_1 \hat{y}_{2i} + u_i \quad (1.3)$$

A third measure, labeled (b), used the same mean function-model (1.2) and an interept-only simplification of the variance model (1.3):

$$\left(\hat{\varepsilon}_{y2y1i}\right)^2 = \gamma_0 + u_i \quad (1.4)$$

The final measure, labeled (d), was based on the regression of the square of y_2 on the square of y_1, with no intercept:

$$y_{2i}^2 = \omega_1 y_{1i}^2 + \delta_i \quad (1.5)$$

2. Ratio and Regression Estimators

Powers and Eltinge (2013) carried out a simulation study to evaluate the properties of simple expansion estimators of population means under pps designs with size measures (1), (a), (b), (c) and (d); detailed descriptions of the populations used for the study are provided in that previous paper. The current paper extends that work for the same populations by evaluating the properties of the ratio estimator

$$\hat{\bar{Y}} = \hat{R}\bar{e}_1 \quad (2.1)$$

where $\hat{R}$ is the customary weighted sample ratio and $\bar{e}_1$ is the known population mean of the employment counts e_{1i}; and the regression estimator

$$\hat{\bar{Y}}_{LR} = \hat{\beta}_0 + \hat{\beta}_1\bar{e}_1 \quad (2.2)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the result of a weighted regression of y_{2i} on e_{1i} with weights determined by the inverses of selection probabilities.

3. Numerical Results

In the current work, for each of the five size measures, we used SAS PROC SURVEYSELECT to produce 10,000 without-replacement samples of size $n = 5, 10$ and 30. Tables 1 through 3 present ratio estimation results for Industry B for estimation of the means of y_2, y_3 and y_4, respectively. Within each table, the first two columns specify the sample size and unit size measure under consideration. The third through fifth columns report the simulation-based estimates of the bias, standard deviation and mean squared error of the ratio estimator. The sixth column reports the ratio defined by the squared bias divided by the mean squared error. The final two columns report two efficiency measures. The first is

$$scale_1 = \{MSE(n, size)\}^{1/2}/\{MSE(n = 30, size = size1)\}^{1/2}$$

where $MSE(n, size)$ is the simulation-based mean squared error of the specified estimator for a sample size n and a size measure $size$. Thus, $scale_1$ allows relatively simple comparisons of the competing design options to the reference design defined by the constant size measure (1) and a sample size of 30. The second efficiency measure is

$$scale_2 = scale_1(30/n)^{1/2}$$

where the additional factor $(30/n)^{1/2}$ adjusts for differences in sample sizes, and thus in a sense makes the results comparable across differing sample sizes. Three features of Tables 1 through 3 are of special interest. First, the size measure (d) based on model (1.5) has produced results that are clearly inferior to those for the other size measures, as reflected in the diagnostics in the final three columns. Second, the size measure (a) leads to ratio estimators that are somewhat more efficient than those produced through designs that use the constant size measure (1), but both (1) and (a) lead to mean squared errors that are somewhat larger than those for (b) and (c). Third, the mean squared errors associated with size measures (b) and (c) tend to be relatively close.

Tables 4 through 6 present related results for Industry C. Again in this case, use of the size measure (d) is problematic. However, for this industry, use of the constant size measure (1) led to smaller mean squared errors than those obtained through use of the size measure (a).

Finally, Figure 1 presents a plot of the regression estimate (2.2) against the ratio estimates (2.1) for the 10,000 replications of the samples of size $n = 5$ based on size measure (1) for industry C. Note that the plot displays a generally ellipsoidal pattern with the major axis approximately following a line that has a slope of one and an intercept of zero; and with a modest indication of right skewness in both the vertical (regression estimation) and horizontal (ratio estimation) dimensions. Figures 2 through 4 present related plots for size measures (a), (b) and (c), respectively. Each of the final three figures deviate somewhat from the approximate ellipsoidal pattern noted for Figure 1.

4. Discussion

The tables and figures presented here have restricted attention to ratio and regression estimation of a mean under unequal-probability sampling from a single stratum. In related work that is not detailed here, we also carried out simulation studies for stratified sampling. Separate and combined ratio and regression estimation were considered for designs that used four distinct forms of allocation of sample sizes across strata: equal allocation; proportional allocation; Neyman allocation based on variances of the second-quarter wage variable; and Neyman allocation based on the variances of a related regression residual.

One could also consider additional point estimators based on, e.g., post-stratification (Cochran, 1977; Little, 1993; and references cited therein); and calibration weighting (Deville and Sarndal, 1992; Kott, 2006; Kott and Liao, 2012; and references cited therein). These alternatives may be of interest for cases in which one has especially rich auxiliary information available through the frame, and for cases that involve substantial levels of nonresponse. In addition, probability-proportional-to-size sampling can produce cases in which some sample units are highly influential due to severe skewness of the underlying size measures. For such cases, practical attention may center on alternative estimators that reduce some extreme weights, and it would be of interest to study the extent to which the one may link the weight-modification approaches with the estimated mean and variance functions that have contributed to a given set of size measures s_i .

5. Acknowledgements and Disclaimer

The authors thank Phil Kott and Michail Sverchkov for helpful discussions of calibration weighting and literature references; and thank Mike Buso for comments on an earlier version of this work. The views expressed here are those of the authors and do not necessarily reflect the policies of the United States Bureau of Labor Statistics.

6. References

Brewer, K.R.W. (1963). Ratio Estimation and Finite Populations: Some Results Deductible from the Assumption of an Underlying Stochastic Process. *Australian Journal of Statistics* **5**, 93-105.

Clark, R.G. (2013). Sample Design Using Imperfect Design Data. *Journal of Survey Statistics and Methodology* **1**, 6-23.

Clark, R.G. (2014). Practical Approaches to Sample Design Using Imperfect Design Information. Paper presented to the Australian Statistical Conference, July 7-10, 2014.

Cochran, W.G. (1977). *Sampling Techniques, Third Edition.* New York: Wiley.

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

Godambe, V.P. (1955). A Unified Theory of Sampling from Finite Populations. *Journal of the Royal Statistical Society, Series B* **17**, 269-278.

Godambe, V.P. (1982). Estimation in Survey Sampling: Robustness and Optimality. *Journal of the American Statistical Association* **77**, 393-403.

Holmberg, A. and B. Swensson (2001). On Pareto πps Sampling: Reflections on Unequal Probability Sampling Strategies. *Theory of Stochastic Processes* **7**, 142-155.

Kott, P.S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology* **32**, 133-142.

Kott, P.S. and D. Liao (2012). Providing double protection for unit nonresponse with a nonlinear calibration-weighting routine. *Survey Research Methods* **6**, 105-111.

Kott, P.S. and J.T. Bailey (2000). The Theory and Practice of Maximal Brewer Selection with Poisson PRN Sampling. *Proceedings of the Second International Conference on Establishment Surveys*, 269-278.

Little, R.J.A. (1993). Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association*, **88**, 1001-1012.

Powers, R.K. and J.L. Eltinge (2013). Properties of Some Sample Designs Based on Imperfect Frame Information. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Thomsen, I., D. Tesfu and D.A. Binder (1986). Estimation of Design Effects and Intraclass Correlations When Using Outdated Measures of Size. *International Statistical Review* **54**, 343-349.

Table 1: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_2 for Industry B. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	-1709.32	40211.86	40248.18	0.00180	2.57277	6.30197
5	a	56.98	32820.56	32820.61	0.00000	2.09798	5.13898
5	b	4703.23	27322.32	27724.16	0.02878	1.77220	4.34099
5	c	2998.79	27282.77	27447.08	0.01194	1.75449	4.29760
5	d	15185.91	45323.34	47799.76	0.10093	3.05548	7.48438
10	1	-1299.81	27565.69	27596.32	0.00222	1.76403	3.05539
10	a	14.61	24075.94	24075.94	0.00000	1.53900	2.66562
10	b	2549.71	18103.47	18282.14	0.01945	1.16864	2.02415
10	c	1508.04	18322.36	18384.32	0.00673	1.17517	2.03546
10	d	9327.48	33012.30	34304.72	0.07393	2.19285	3.79812
30	1	-443.83	15637.62	15643.92	0.00080	1.00000	1.00000
30	a	-29.80	13901.51	13901.54	0.00000	0.88862	0.88862
30	b	1557.65	11057.24	11166.41	0.01946	0.71379	0.71379
30	c	481.73	10875.11	10885.78	0.00196	0.69585	0.69585
30	d	4693.22	21910.10	22407.11	0.04387	1.43232	1.43232

Table 2: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_3 for Industry B. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	-1437.71	47089.88	47111.83	0.000931	2.65827	6.51141
5	a	119.11	38867.86	38868.05	0.000009	2.19312	5.37202
5	b	4962.96	28981.06	29402.94	0.028491	1.65905	4.06383
5	c	2946.07	29739.24	29884.81	0.009718	1.68624	4.13043
5	d	15464.39	46908.04	49391.41	0.098031	2.78690	6.82647
10	1	-977.08	31552.81	31567.93	0.000958	1.78121	3.08515
10	a	190.51	27445.30	27445.96	0.000048	1.54863	2.68231
10	b	2471.48	19515.09	19670.97	0.015786	1.10993	1.92245
10	c	1626.51	20265.17	20330.34	0.006401	1.14713	1.98689
10	d	9463.79	35152.70	36404.33	0.067581	2.05410	3.55781
30	1	-280.38	17720.51	17722.73	0.000250	1.00000	1.00000
30	a	61.06	15952.42	15952.54	0.000015	0.90012	0.90012
30	b	1588.29	11982.30	12087.10	0.017267	0.68201	0.68201
30	c	614.08	12073.89	12089.50	0.002580	0.68215	0.68215
30	d	4941.77	23383.56	23900.04	0.042753	1.34855	1.34855

Table 3: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_4 for Industry B. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	-1979.52	44200.17	44244.47	0.002002	2.66715	6.53316
5	a	-197.63	36838.03	36838.56	0.000029	2.22071	5.43960
5	b	4396.58	29207.38	29536.44	0.022157	1.78052	4.36136
5	c	2293.85	30278.52	30365.29	0.005707	1.83048	4.48375
5	d	13577.53	47902.93	49789.96	0.074363	3.00144	7.35201
10	1	-1011.65	29902.52	29919.62	0.001143	1.80362	3.12396
10	a	-137.60	26537.18	26537.53	0.000027	1.59974	2.77083
10	b	2052.06	19395.09	19503.34	0.011070	1.17570	2.03638
10	c	1269.46	20319.10	20358.72	0.003888	1.22727	2.12569
10	d	8549.25	37521.16	38482.82	0.049354	2.31983	4.01806
30	1	-415.39	16583.47	16588.67	0.000627	1.00000	1.00000
30	a	30.07	15701.91	15701.94	0.000004	0.94655	0.94655
30	b	1405.76	11839.84	11923.01	0.013901	0.71874	0.71874
30	c	499.27	11962.14	11972.55	0.001739	0.72173	0.72173
30	d	4424.10	23588.70	23999.99	0.033980	1.44677	1.44677

Table 4: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_2 for Industry C. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	924.74	16381.75	16407.83	0.003176	2.85166	6.98510
5	a	216.52	15662.82	15664.32	0.000191	2.72243	6.66857
5	b	1592.96	12466.21	12567.58	0.016066	2.18423	5.35024
5	c	1385.32	12471.89	12548.59	0.012187	2.18093	5.34215
5	d	3830.61	15267.90	15741.10	0.059220	2.73578	6.70127
10	1	463.61	10291.09	10301.53	0.002025	1.79039	3.10105
10	a	-3.89	10378.90	10378.90	0.000000	1.80384	3.12434
10	b	998.52	8594.62	8652.43	0.013318	1.50378	2.60462
10	c	738.58	8346.58	8379.19	0.007769	1.45629	2.52237
10	d	2720.97	12808.78	13094.60	0.043178	2.27582	3.94184
30	1	134.82	5752.21	5753.79	0.000549	1.00000	1.00000
30	a	-29.37	6162.54	6162.61	0.000023	1.07105	1.07105
30	b	423.76	5047.79	5065.55	0.006998	0.88038	0.88038
30	c	265.21	4773.45	4780.81	0.003077	0.83090	0.83090
30	d	1148.66	8066.17	8147.54	0.019876	1.41603	1.41603

Table 5: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_3 for Industry C. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	633.68	17710.19	17721.52	0.001279	2.70088	6.61579
5	a	361.76	19473.88	19477.24	0.000345	2.96847	7.27123
5	b	1699.77	14624.30	14722.75	0.013329	2.24385	5.49629
5	c	1487.49	14286.14	14363.38	0.010725	2.18908	5.36213
5	d	3903.09	16023.11	16491.64	0.056013	2.51344	6.15665
10	1	327.85	11506.26	11510.93	0.000811	1.75435	3.03862
10	a	40.89	11903.56	11903.63	0.000012	1.81420	3.14228
10	b	967.75	9740.34	9788.30	0.009775	1.49181	2.58388
10	c	816.46	9492.49	9527.54	0.007344	1.45206	2.51505
10	d	2579.96	12927.17	13182.11	0.038305	2.00905	3.47977
30	1	132.70	6560.04	6561.38	0.000409	1.00000	1.00000
30	a	-49.13	7954.02	7954.18	0.000038	1.21227	1.21227
30	b	445.35	5747.20	5764.42	0.005969	0.87854	0.87854
30	c	274.22	5381.60	5388.58	0.002590	0.82126	0.82126
30	d	1199.63	9592.81	9667.53	0.015398	1.47340	1.47340

Table 6: Simulation Evaluation of Properties of the Combined Ratio Estimator for Specified Sample Sizes and Unit Size Measures: Estimation of the Mean of y_4 for Industry C. Results from 10,000 Replications.

n	size	bias	stderr	rootMSE	biasratio	scale1	scale2
5	1	650.98	18211.89	18223.52	0.001276	2.73074	6.68892
5	a	307.06	19954.32	19956.68	0.000237	2.99045	7.32508
5	b	1605.22	15362.51	15446.14	0.010800	2.31456	5.66949
5	c	1352.65	14283.31	14347.22	0.008889	2.14989	5.26613
5	d	3841.72	16620.54	17058.76	0.050717	2.55621	6.26140
10	1	286.12	11740.96	11744.44	0.000594	1.75987	3.04818
10	a	95.03	12972.47	12972.82	0.000054	1.94394	3.36700
10	b	1020.64	9545.94	9600.35	0.011303	1.43858	2.49170
10	c	831.29	9810.60	9845.76	0.007129	1.47536	2.55540
10	d	2549.75	13388.08	13628.72	0.035001	2.04222	3.53723
30	1	170.94	6671.28	6673.47	0.000656	1.00000	1.00000
30	a	-135.27	7156.89	7158.17	0.000357	1.07263	1.07263
30	b	422.49	5662.83	5678.57	0.005535	0.85092	0.85092
30	c	290.26	5471.23	5478.92	0.002807	0.82100	0.82100
30	d	1126.39	9580.60	9646.59	0.013634	1.44551	1.44551

Figure 1: Plot of the regression-based estimates of the mean of y_2 against ratio-based estimates. Simulation results from 10,000 samples of size n=5 selected from industry C based on size measure 1

Note that 55 of the 10,000 points exceeded the plot bounds, and are thus omitted.

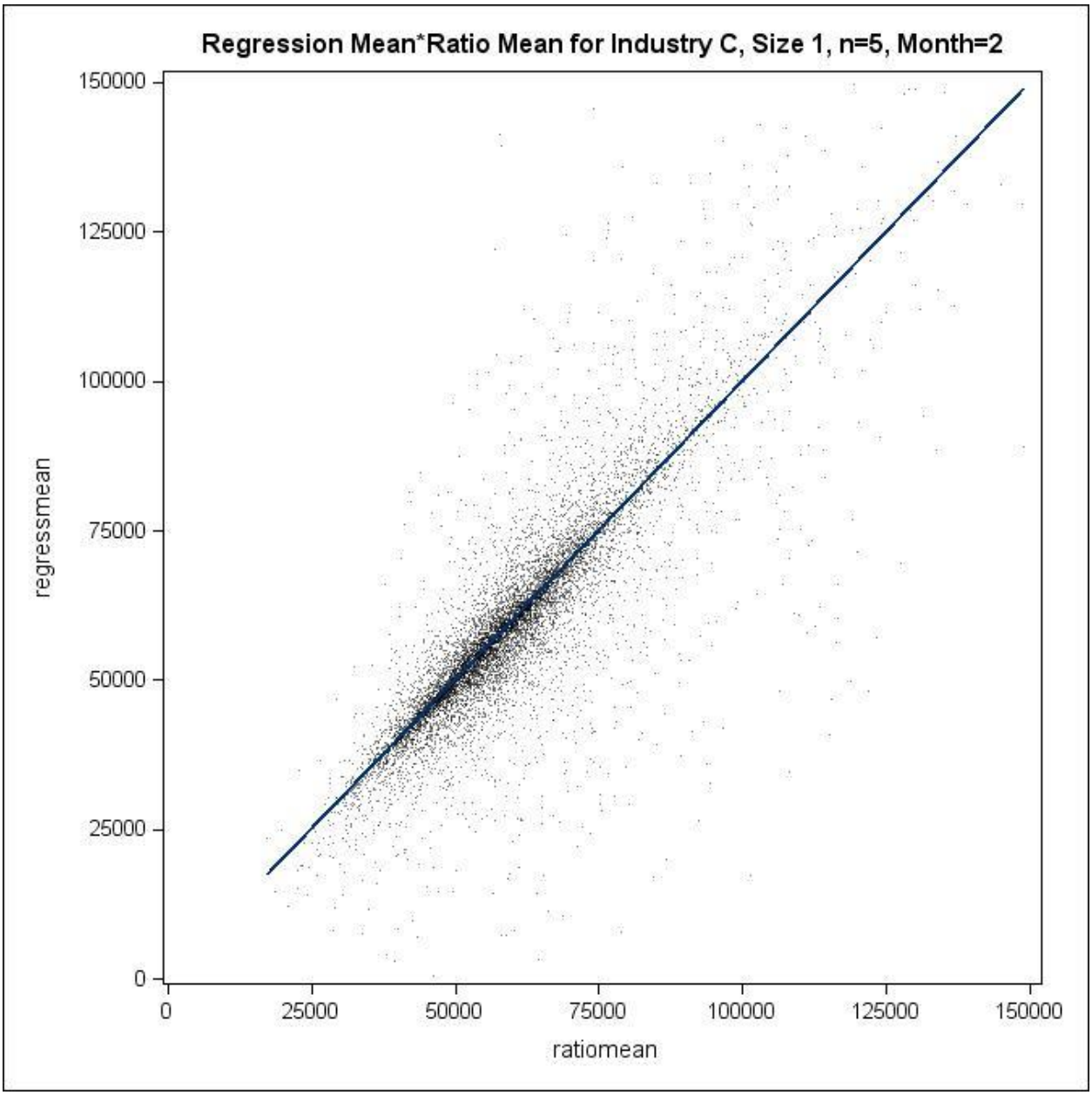

Figure 2: Plot of the regression-based estimates of the mean of y_2 against ratio-based estimates. Simulation results from 10,000 samples of size n=5 selected from industry C based on size measure a

Note that 219 of the 10,000 points exceeded the plot bounds, and are thus omitted.

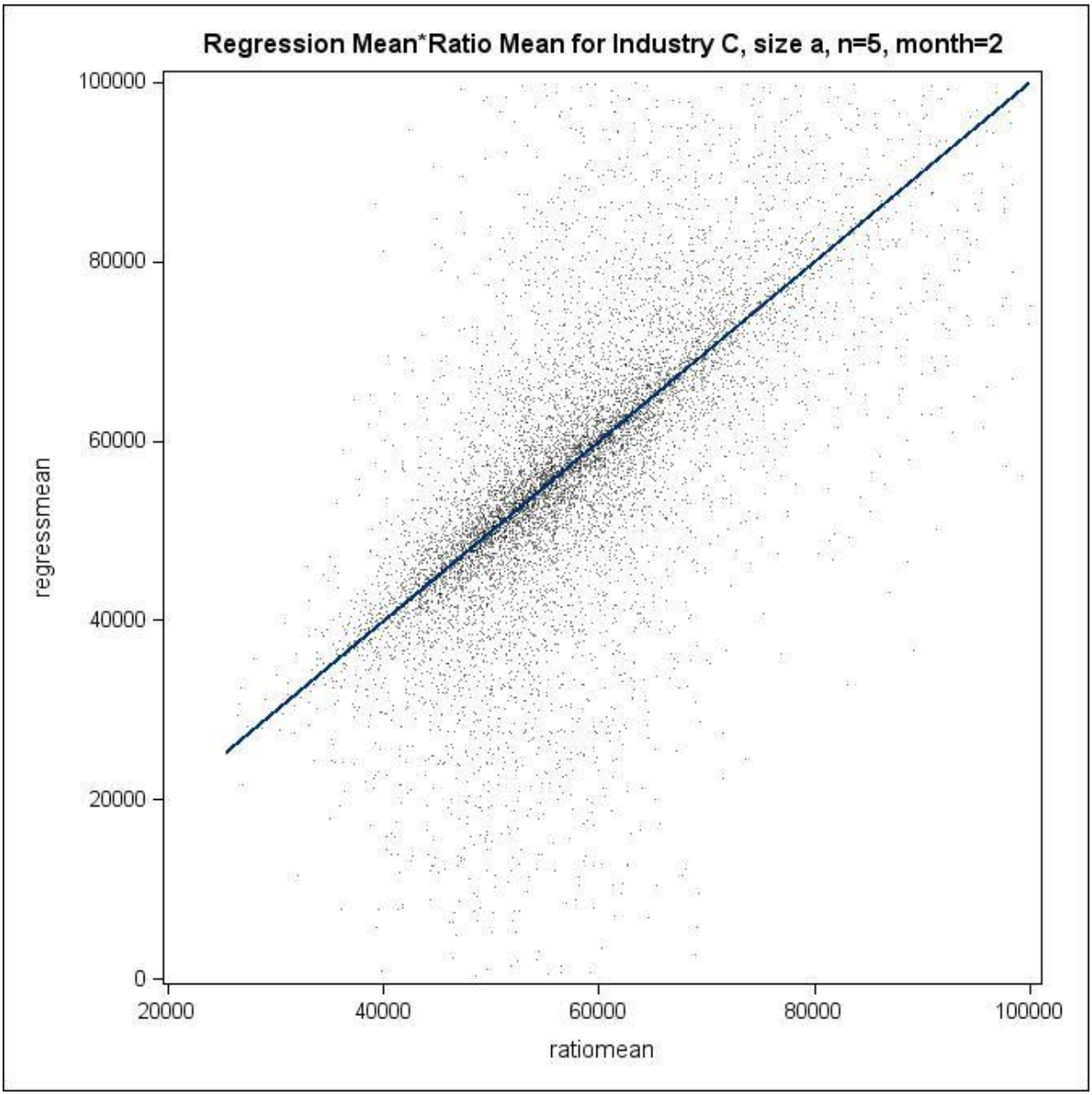

Figure 3: Plot of the regression-based estimates of the mean of y_2 against ratio-based estimates. Simulation results from 10,000 samples of size n=5 selected from industry C based on size measure b

Note that 187 of the 10,000 points exceeded the plot bounds, and are thus omitted.

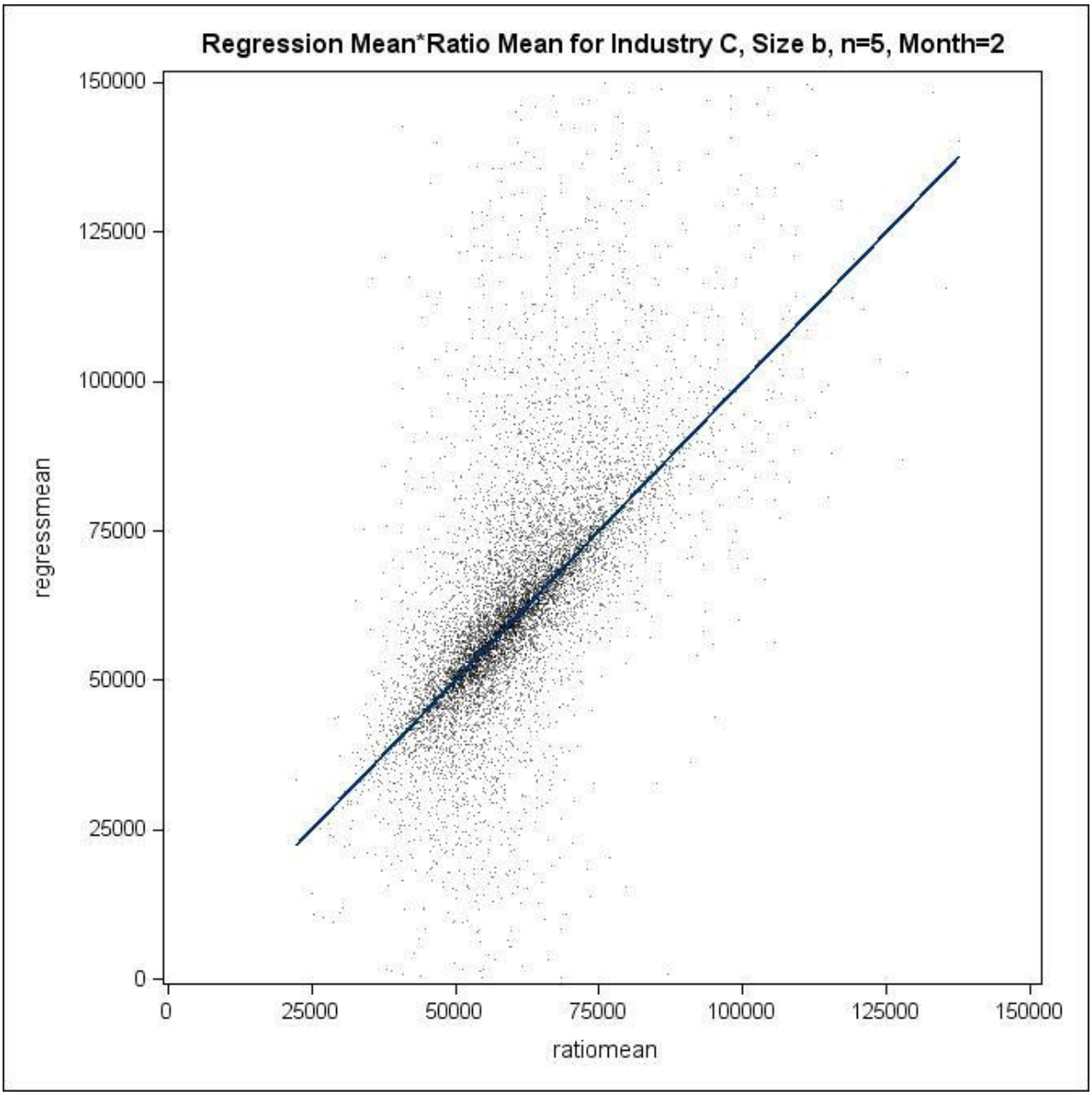

Figure 4: Plot of the regression-based estimates of the mean of y_2 against ratio-based estimates. Simulation results from 10,000 samples of size n=5 selected from industry C based on size measure c

Note that 97 of the 10,000 points exceeded the plot bounds, and are thus omitted.