Properties of Some Sample Designs Based on Imperfect Frame Information

Randall K. Powers and John L. Eltinge

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

Powers.Randall@bls.gov

Key words: Estimation error; Frame information; Lack of fit; Probability-proportional-to-size (pps) designs.

1. Introduction: Probability-Proportional-to-Size Designs

In the design of large-scale surveys, one often uses unequal probabilities of selection for sampling within strata. Standard derivations (e.g., Cochran, 1977, Section 9A.3) indicate that under idealized conditions for estimation of a population total for a high-priority survey variable Y, the selection probability for a unit i would be proportional to the corresponding unit value Y_i. In general, the values Y_i are not known during the design phase of a study. For such cases, one often uses a "probability proportional to size" design in which the unit sizes are set equal to a gross size measure that is available on the sampling frame. In establishment surveys, common examples include the total number of employees, the total wages paid, or the total sales of a given unit i in a specified reference period.

However, in some cases standard gross-size measures may not provide good approximations to the idealized probabilities that would be proportional to the target variables Y_i. For cases that involve this type of imperfect frame information, one may (under conditions) minimize the variance of a standard probability-weighted expansion estimator of a population mean or total by using selection probabilities proportional to the unit-level size measures

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

where $\mu(X_i)$ and $[\sigma(X_i)]^2$ are the conditional mean and variance, respectively, of Y_i given the available frame information X_i. See, e.g., Godambe (1955), Brewer (1963), Thomsen et al (1986), Kott and Bailey (2000), Holmberg and Swensson (2001) and references cited therein.

Practical use of a size measure based on expression (1.1) depends on (a) timely information available to estimate the functions $\mu(X_i)$ and $[\sigma(X_i)]^2$; (b) the adequacy of fit in estimation of these functions; and (c) consistency of the idealized size measures across different survey variables Y_i.

The remainder of this paper addresses some aspects of issues (a) through (c) through numerical examples and simulation work. Section 2 presents examples involving three industries and several forms of size measures defined by the general expression (1.1). Section 3 presents the results of a related simulation study. Section 4 reviews the main ideas presented in this paper, and suggests some potential areas for future work.

2. An Example: Frame Information for An Establishment Survey

2.1. Data from the BLS Quarterly Census of Employment and Wages

To explore the issues (a) through (c) outlined in Section 1, we carried out an empirical study based on data from one state as provided in the BLS Quarterly Census of Employment and Wages (QCEW). The QCEW forms the basis for several large-scale establishment surveys managed by the Bureau of Labor Statistics. However, to allow a direct evaluation of competing size measures, the unequal-probability designs considered in this paper are simpler than those used in the BLS establishment surveys.

We considered three industries at the state level, and focused on three quarterly measures. First, e_{1i} is the total employment in unit i for quarter 1; this is an example of a gross size measure that one commonly uses in pps designs. Second, e_{2i} is the total employment in unit i for quarter 2. In addition, Y_{1i} and Y_{2i} are the total payroll amounts for unit i in quarters 1 and 2, respectively.

To avoid confidentiality issues, this study omitted units that were beyond the 95th percentile in any of their e or Y values. Table 0 presents basic descriptive statistics for the resulting trimmed populations. Note especially that each of the variables have skewed distributions, and that industry B has a somewhat larger skewness coefficient than industries A and C.

2.2 Regression Models for Mean and Variance Functions

For each industry, we fit several regression models for use in construction of the size measure S_i. First, we considered a simple linear regression of y_2 on y_1:

$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \varepsilon_{y2y1i} \tag{2.1}$$

Second, we fit two separate variance function models, based on the squared residuals from the initial model fit. In the first variance model,

$$\left(\hat{\varepsilon}_{y2y1i}\right)^2 = \gamma_0 + \gamma_1 \hat{y}_{2i} + u_i \tag{2.2}$$

we regressed the squared residual from model (2.1) on the predictor from the initial fit. In the second variance model,

$$\left(\hat{\varepsilon}_{y2y1i}\right)^2 = \gamma_0 + u_i \tag{2.3}$$

we regressed the squared residual from model (2.1) on a simple intercept alone. In addition, we tried a direct regression of the square of y_2 on the square of y_1, with no intercept:

$$y_{2i}^2 = \omega_1 y_{1i}^2 + \delta_i \quad (2.4)$$

Table 1 presents the results of these regression fits for industry A. Note especially that the initial mean-function regression has a relatively high value of R^2, equal to 0.872. In contrast with this, the first variance-function model has a relatively small value of R^2, equal to 0.138. Tables 2 and 3 presents the results of the fits of the regression models (2.1) through (2.4) for industries B and C, respectively. The results are qualitatively similar to industry A.

Finally, we also fit similar models that used the outcome variable y_2, but used e_1 as the predictor instead of y_1.

$$y_{2i} = \beta_0 + \beta_1 e_{1i} + \varepsilon_{y2e1i} \quad (2.5)$$

$$\left(\hat{\varepsilon}_{y2e1i}\right)^2 = \gamma_0 + \gamma_1 \hat{y}_{2i} + u_i \quad (2.6)$$

$$\left(\hat{\varepsilon}_{y2e1i}\right)^2 = \gamma_0 + u_i \quad (2.7)$$

$$y_{2i}^2 = \omega_1 e_{1i}^2 + \delta_i \quad (2.8)$$

Tables 4 through 6 present the numerical results of these models fits for industries A, B and C, respectively. Note that for a given industry, the numerical results for models (2.5) through (2.8) tend to be qualitatively similar to those presented for models (2.1) through (2.4), but with somewhat lower R-squared values.

2.3 Four Size Measures

Based on the results of Section 2.2, we considered four size measures:

Measure *a* is the simple gross size measure e_1, the employment total in the baseline quarter. Measure *b* is the size measure s_i computed from the mean-function model (2.1) and the complex variance function model (2.2). Measure *c* is the size measure s_i computed from the mean-function model (2.1) and the simpler intercept-only variance function model (2.3). Measure *d* is the size measure s_i computed from the direct model (2.4) for the regression of y_2^2 on y_1^2.

Figure 1 shows a plot of y_2 and the final three size measures against the gross size measure e_1, based on a one-in-thirty subset of the original full population for industry B. Note that there is one prominent outlier. In addition, the final three size measures tend to be relatively close together. Figure 2 shows the corresponding plot for industry B with

both axes presented on the logarithmic scale. Note especially that for several cases, the log of the size measure b is equal to zero. This corresponds to cases in which the complex variance function fit led to negative values, which were truncated to the value zero on the log scale.

Figure 3 shows a plot of y_2 and the final three size measures against the gross size measure e_1, based on a one-in-three-hundred subset of the original full population for industry C. Figure 4 shows the corresponding plot for industry C, based on a one-in-three-hundred subset of the original population. The plot is qualitatively similar to the corresponding plot for industry B, with one important exception: for size measure b, there are fewer values truncated to zero on the log scale, and thus fewer problems with lack of fit.

2.4 Direct Evaluation of Design Effects from Finite Population Quantities

For each of the three populations identified in Section 2.1 and each of the four size measures defined in Section 2.2, we evaluated the efficiency of the associated "probability proportional to size" design through computation of the design effect ratios,

$$\Delta = V(\hat{Y}|pps)/V(\hat{Y}|srs) \qquad (2.9)$$

where $V(\hat{Y}|pps)$ is the variance of the standard probability-weighted estimator of a population total under a probability-proportional-to-size design, as given in expression (9A.11) of Cochran (1977) with a sample of size 1; and $V(\hat{Y}|srs)$ is the variance of a standard expansion estimator of a population total under simple random sampling with a sample of size 1.

Table 7 presents the resulting design effect ratios for estimation of the mean or total of y_2 Note that for size measures *a*, *c* and *d*, the design effects are all less than one. In addition, the more refined size measures produce design effects that are approximately one-third of the design effects produced with the gross size measure *a*. In other words, use of the more refined size measures can produce substantial efficiency gains, relative to those obtained from use of the gross size measure. However, we obtained a different result from the size measures *b* computed from the complex variance model (2.2). This complex model produced a small number of extreme values, which in turn led to very large design effects. Thus, caution is warranted in the possible use of these more complex variance models to produce a size measure.

Table 8 presents related design-effect results for estimation of the population total or mean of the second-quarter variable e_2. This was a variable for which the size measures *b*, *c* and *d* were not originally intended. Thus, the design effects in this table reflect the efficiency properties that would be obtained for variables that were not the top priority in design optimization. As observed in the previous table, direct use of the size measure *b* is problematic. On the other hand, use of the refined size measures *c* or *d* led to moderate

losses of efficiency in estimation for the second-quarter variable e_2, relative to use of the gross size measure *a*. Thus, if we use the refined size measures *c* or *d*, we obtain substantial efficiency gains for estimation of the mean or total of the high-priority variable y_2, while paying a moderate price in efficiency loss for the lower-priority variable e_2.

3. A Simulation Study

As a complement to these calculations of exact design effects presented in Section 2.4, we carried out a simulation study. For each industry and each size measure, we produced 10,000 replications of a probability-proportional-to-size selection with sample size equal to 1. Based on the simulation distributions of the resulting estimators, we computed two efficiency measures.

The first efficiency measure was similar to the design-effect ratio defined by expression (2.9), but with the numerator and denominator both computed from the simulation-based variances. The second efficiency measure was defined similarly, but with the variances replaced by the squares of the interquartile ranges of the respective estimators. This second efficiency measure is relatively insensitive to outliers.

Table 9 presents the simulation-based design effect results for estimation of the mean of y_2, for industries A, B and C. Note that in this case, all three of the refined size measures – *b*, *c* and *d* – produced design effects that were substantially less than the design effect obtained through use of the gross size measure *a*. Table 10 presents the corresponding results for estimation of the mean of the variable e_2.

4. Discussion

4.1. Summary of Ideas and Numerical Results

In summary, this paper has explored the use of various measures of unit size in the implementation of "probability proportional to size" sample designs. In keeping with some previous literature, we have contrasted use of simple gross size measures with more refined size measures based on estimation of mean and variance functions for a survey variable y, conditional on a vector of known frame variables.

The paper has placed primary emphasis on an empirical study of the properties of the resulting point estimators. That study identified some cases in which one may obtain substantial efficiency gains for some regression-based size measures, relative to a standard gross-size measure. Thus, for cases is which we have one higher-priority survey variable for which we can develop good mean-function and variance-function approximations, it is worthwhile to consider size measures based on these approximations, instead of gross size measures.

However, the empirical study also indicated the potential for sensitivity to lack of fit, especially in estimation of a complex variance function for use in the refined size measure.

4.2. Possible Areas for Future Research

There are several potential extensions of this work. For example, under forms of the designs considered here, one could explore the properties of more refined point estimators based on ratio estimation and calibration weighting. In addition, one could consider cases in which one must account for measurement errors in either the survey variable or the frame variables.

Acknowledgements and Disclaimer

The authors thank Rachel Harter, Anders Holmberg, Avi Singh, Daniell Toth and Erica Yu for helpful discussions of size measures in unequal-probability sampling; and thank Phil Kott for providing the Godambe (1955), Brewer (1963).and Kott and Bailey (2000) references. Any opinions expressed in this paper are those of the authors and do not constitute policy of the Bureau of Labor Statistics.

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.

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

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

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. 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.

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 0: Descriptive Statistics for Three Industries

		Industry		
Parameter	Variable	A	B	C
Population Size	-------	123	2978	27,163
Mean	E_1	5.59	23.74	16.16
	Y_1	26,733	126,078	59,437
	Y_2	27,623	125,654	58,939
Standard Deviation	E_1	5.36	30.84	15.23
	Y_1	26,878	187,944	60,590
	Y_2	29,394	192,247	60,266
Skewness	E_1	1.75	2.46	1.35
	Y_1	1.35	2.80	1.61
	Y_2	1.45	2.88	1.60

Table 1: Regression Model Fits for y_2 Using Predictor y_1 for Industry A

Models for y_2	Intercept (Std Error)	Slope (StdError)	R^2	MSE
Mean Function Model (2.1)	323 (1346)	1.021 (0.036)	0.872	$1.12x10^9$
Variance Function Model (2.2) Model (2.3)	$5.21x10^7\left(6.98x10^7\right)$ $1.10x10^7(4.9x10^8)$	2086 (1796) -------	0.011 -------	$2.96x10^{17}$ $2.97x10^{17}$
Direct Model for y_2^2: Model (2.4)	7096 (8725)	0.995 (0.113)	0.838	$1.96x10^{18}$

Table 2: Regression Model Fits for y_2 Using Predictor y_1 for Industry B

Models for y_2	Intercept (Std Error)	Slope (Std Error)	R^2	MSE
Mean Function Model (2.1)	525(1027)	0.992 (0.005)	0.941	$2.17x10^9$
Variance Function Model (2.2) Model (2.3)	$-1.01x10^9\left(2.60x10^8\right)$ $2.17x10^9\left(2.32x10^9\right)$	25280 (1157) -------	0.138 -------	$1.39x10^{20}$ $1.61x10^{20}$
Direct Model for y_2^2: Model (2.4)	66207 (11412)	0.911 (0.016)	0.886	$3.44x10^{21}$

Table 3: Regression Model Fits for y_2 Using Predictor y_1 for Industry C

Models for y_2	Intercept (Std Error)	Slope (Std Error)	R^2	MSE
Mean Function Model (2.1)	2412 (150)	0.951 (0.002)	0.914	$3.11x10^8$
Variance Function Model (2.2) Model (2.3)	$-9.52x10^7\left(1.25x10^7\right)$ $3.11x10^8\left(9.04x10^6\right)$	6898 (151) -------	0.071 -------	$2.06x10^{18}$ $2.22x10^{18}$
Direct Model for y_2^2: Model (2.4)	29024 (939)	0.778 (0.005)	0.881	$2.86x10^{19}$

Table 4: Regression Model Fits for y_2 Using Predictor e_1 for Industry A

Models for y_2	Intercept (Std Error)	Slope (Std Error)	R^2	MSE
Mean Function Model (2.5)	2545 (2222)	4483 (287)	0.668	$2.89x10^8$
Variance Function Model (2.6)	$5.01x10^7\left(8.72x10^7\right)$	8492(2388)	0.095	$4.01x10^{17}$
Model (2.7)	$2.85x10^8\left(5.98x10^7\right)$	-------	N/A	$4.40x10^{17}$
Direct Model for y_2^2: Model (2.8)	-------	$2.04x10^7\left(1.52x10^6\right)$	0.596	$4.85x10^{18}$

Table 5: Regression Model Fits for y_2 Using Predictor e_1 for Industry B

Models for y_2	Intercept (Std Error)	Slope (Std Error)	R^2	MSE
Mean Function Model (2.5)	-7545 (1937)	5611 (50)	0.810	$7.02x10^9$
Variance Function Model (2.6)	$-9.53x10^8\left(8.56x10^8\right)$	63399 (4004)	0.078	$1.43x10^{21}$
Model (2.7)	$7.01x10^9\left(7.21x10^8\right)$	-------	N/A	$1.55x10^{21}$
Direct Model for y_2^2: Model (2.8)	-------	$3.34x10^7\left(4.08x10^5\right)$	0.692	$9.33x10^{21}$

Table 6: Regression Model Fits for y_2 Using Predictor e_1 for Industry C

Models for y_2	Intercept (Std Error)	Slope (Std Error)	R^2	MSE
Mean Function Model (2.5)	3753 (269)	3415 (12)	0.745	$9.26x10^8$
Variance Function Model (2.6)	$3.85x10^7\left(2.75x10^7\right)$	15053(350)	0.064	$9.02x10^{18}$
Model (2.7)	$9.26x10^8\left(1.88x10^7\right)$	-------	N/A	$9.63x10^{18}$
Direct Model for y_2^2: Model (2.8)	-------	$1.31x10^7\left(5.29x10^4\right)$	0.692	$7.41x10^{19}$

Table 7: Design Effects for Estimation of $\bar{Y}_2$. Effects Computed Directly from Population Values Based on Expression (2.9)

	Industry		
Size Measure	A	B	C
a (employment counts e_1)	0.3398	0.1562	0.3240
b (predicted y_2 with complex variance model)	0.1462	233 Large	33.2 Large
c (predicted y_2 with intercept-only variance model)	0.1690	0.0571	0.0952
d (direct regression of y_2^2 on y_1^2)	0.1236	0.0430	0.0838

Table 8: Design Effects for Estimation of $\bar{E}_2$. Effects Computed Directly from Population Values Based on Expression (2.9)

	Industry		
Size Measure	A	B	C
a (employment counts e_1)	0.0738	0.0412	0.1407
b (predicted y_2 with complex variance model)	0.1860	404 Large	142 Large
c (predicted y_2 with intercept-only variance model)	0.1884	0.0951	0.1968
d (direct regression of y_2^2 on y_1^2)	0.2238	0.0695	0.1819

Table 9: Design Effects for Estimation of $\bar{Y}_2$. Effects Estimated from Simulation Based on Repeated Sampling Under Specified Designs

		Industry		
Size Measure	Efficiency Measure	A	B	C
a (employment counts e_1)	Variance	0.3505	0.1484	0.8186
	$(IQR)^2$	0.2343	0.2563	0.1258
b (predicted y_2 with complex variance model)	Variance	0.1455	0.0295	0.0649
	$(IQR)^2$	0.0285	0.0739	0.0427
c (predicted y_2 with intercept-only variance model)	Variance	0.1638	0.0575	0.0958
	$(IQR)^2$	0.0627	0.1679	0.0674
d (direct regression of y_2^2 on y_1 and y_1^2)	Variance	0.1260	0.0419	0.0833
	$(IQR)^2$	0.0353	0.1491	0.0752

Table 10: Design Effects for Estimation of $\bar{E}_2$. Effects Estimated from Simulation Based on Repeated Sampling Under Specified Designs

		Industry		
Size Measure	Efficiency Measure	A	B	C
a (employment counts e_1)	Variance	0.0787	0.0399	0.7714
	$(IQR)^2$	0.0177	0.0163	0.0141
b (predicted y_2 with complex variance model)	Variance	0.1906	0.0777	0.2001
	$(IQR)^2$	0.2367	0.1363	0.1096
c (predicted y_2 with intercept-only variance model)	Variance	0.1875	0.0944	0.1939
	$(IQR)^2$	0.2219	0.2100	0.1363
d (direct regression of y_2^2 on y_1 and y_1^2)	Variance	0.2203	0.070	0.1711
	$(IQR)^2$	0.2139	0.3759	0.1187

Figure 1: Plot of y2i and size measures b, c, and d against e1i for Industry B

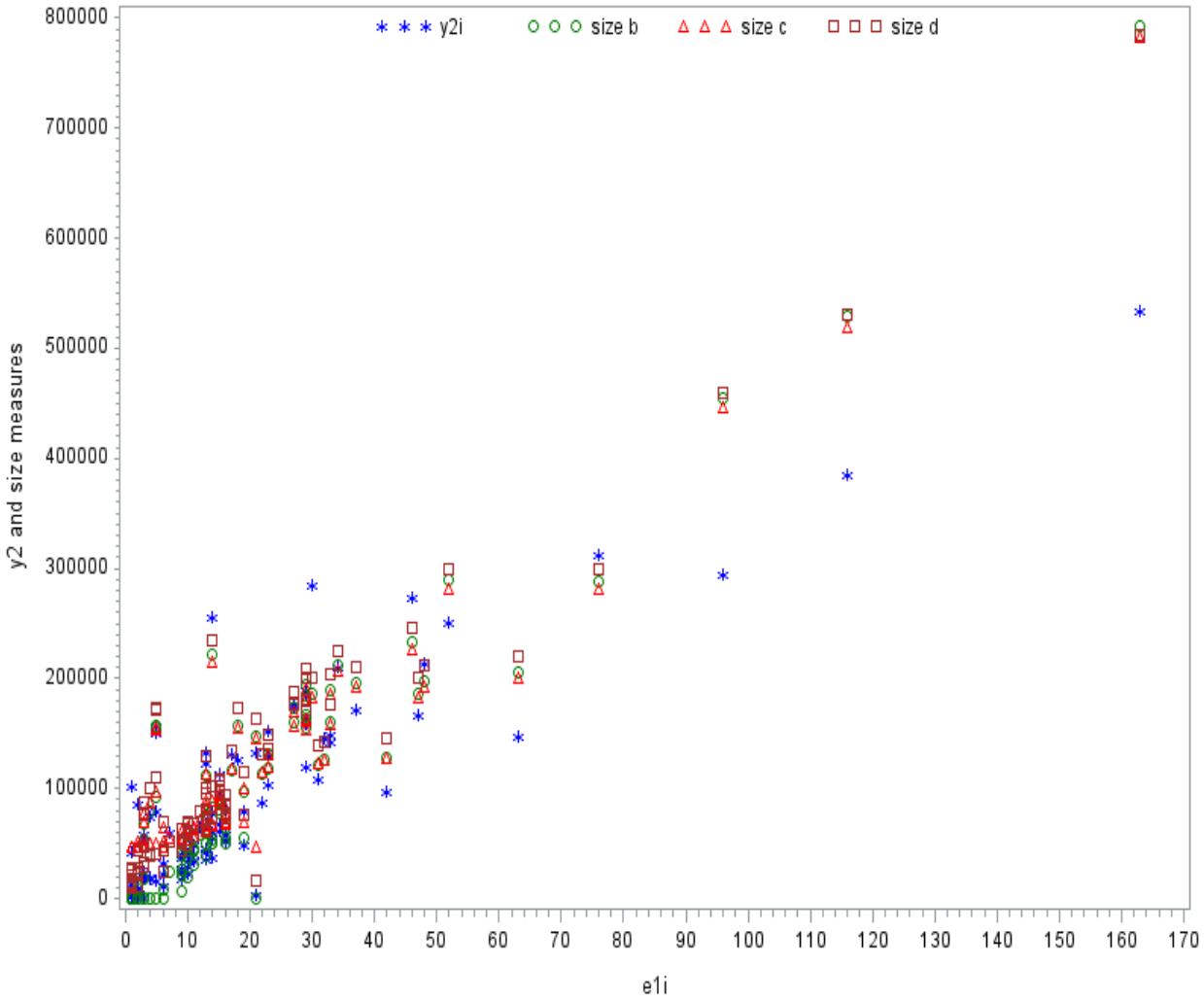

Figure 2: Plot of ln(y2i) and logarithms of size measures b, c, and d against ln(e1i) for Industry B

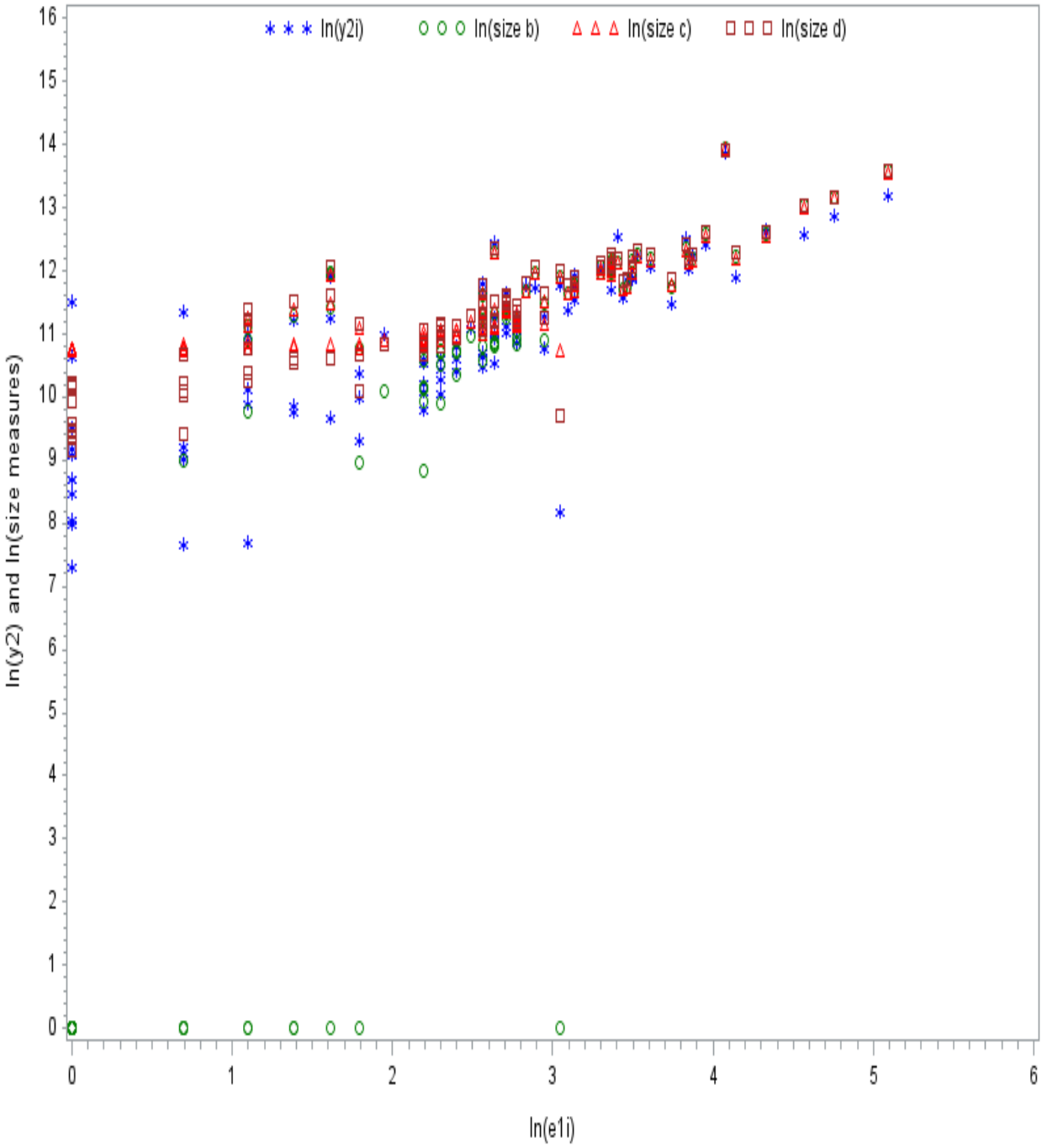

Figure 3: Plot of y2i and size measures b, c, and d against e1i for Industry C

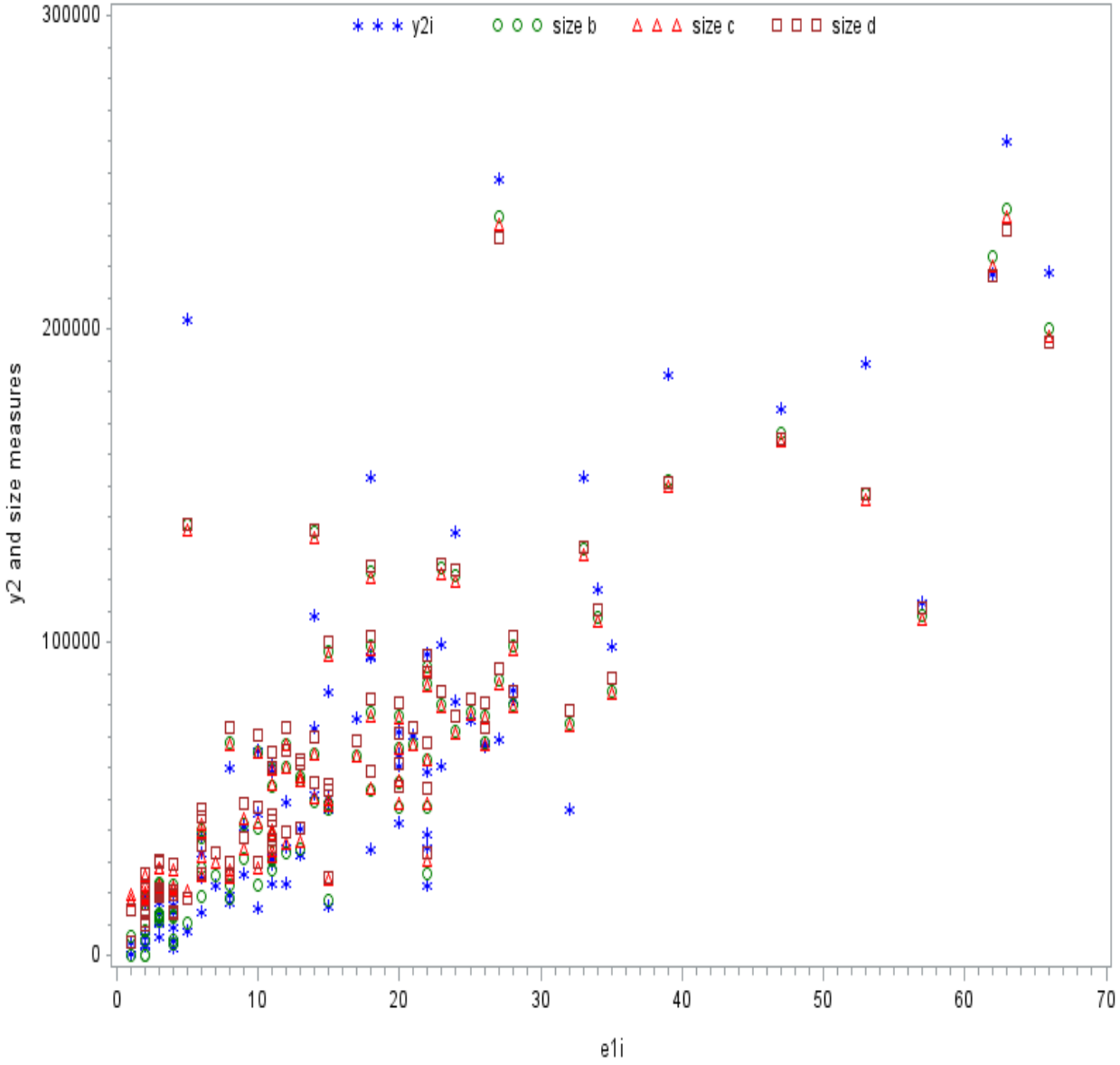

Figure 4: Plot of ln(y2i) and logarithms of size measures b, c, and d against ln(e1i) for Industry C