

10th National Convention on Statistics (NCS)
EDSA Shangri-La Hotel
October 1-2, 2007

**Estimation Procedure for the Quarterly Survey
of Philippine Business and Industry**
by

Gloria A. Cubinar and Erniel B. Barrios

For additional information, please contact:

Author's name	:	Gloria A. Cubinar
Designation	:	Statistician IV
Affiliation	:	National Statistics Office
Address	:	Sta. Mesa, Manila
Tel. no.	:	(0632) 713-7065
E-mail	:	G.Cubinar@census.gov.ph
Co-author's name	:	Erniel B. Barrios
Designation	:	Professor
Affiliation	:	University of the Philippines Diliman
Address	:	Diliman, Quezon City
Tel. no.	:	(0632) 929-2875
E-mail	:	ernielb@yahoo.com

Estimation Procedure for the Quarterly Survey of Philippine Business and Industry¹

by

Gloria A. Cubinar and Erniel B. Barrios²

ABSTRACT

A model-based estimation procedure in the nonparametric bootstrap framework is proposed to provide estimates for the Quarterly Survey of Philippine Business and Industry (QSPBI), considering the non-response situation among establishment surveys.

Intensive stratification (industry and employment size), probability proportional to size sampling (pps) and the current design are simulated assuming the results of first quarter of 2005 QSPBI as frame. Current design and probability proportional to size (pps) sampling with ATE as the measure of size yield reliable estimates especially when the sample size is small. As the sample size increases, the behavior of estimates from the three sampling strategies does not differ much.

The non-sampled part of the population and the non-responding samples are best predicted through gamma regression or generalized linear model (panel data). Resampling from the Monte Carlo population, and the nonparametric bootstrap yield low bias for the estimates of the population characteristics.

I. Introduction

The QSPBI is one of the regular surveys conducted by the National Statistics Office (NSO) particularly the Industry and Trade Statistics Department (ITSD). This survey aims to provide quarterly data on gross revenue/sales, employment and compensation for each of the industry major groups (3-digit code) of the 1994 Philippine Standard Industrial Classification (PSIC) and some selected industry sub-sectors or group of industry sub-sectors (2/4/5-digit PSIC code). Specifically, the survey data are used in constructing a national index as indicator of quarterly economic trends. The National Statistical Coordination Board (NSCB) also uses the results as direct inputs in the generation of the Quarterly National Accounts (QNA) in updating the estimates of Gross Value Added (GVA).

The 2004 QSPBI covers 17 administrative regions and 12 sectors namely: Mining and Quarrying (C); Manufacturing (D); Electricity Gas and Water (E), Construction (F), Wholesale and Retail Trade and Repair Services (G); Hotels and Restaurants (H); Transportation, Storage and Communications (I); Financial Intermediation (J); Real Estate, Renting and Business Activities (K); Education (M); Health and Social Work (N); Other

¹ Paper to be presented at the 10th National Convention on Statistics on October 1-2, 2007 at the EDSA Shangri-La Hotel.

² Statistician IV of Industry and Trade Statistics Department, Philippine National Statistics Office and Professor, School of Statistics, University of the Philippines-Diliman

Community, Social and Personal Services (O). However, for this study, it focuses only on the Wholesale and Retail Trade Sector.

Access to samples and cooperation among respondents are difficult to secure among the sample establishments. In the Philippines, even after 45 days of survey operation, non-response is still reported at 20-55%. Ninety days after, non-response rate could still be observed in the vicinity of 2-10%. Because of these problems, non-probability sampling scheme was adopted since 2001. Reduction in the sample size, manageability in the field operation, lessening respondent's burden and coming up with more "representative" samples are also the reasons for the change in the sampling design.

The non-probability nature of the samples, however, impedes the estimation of the population characteristics generated from this survey. Thus, utilization of the QSPBI results is confined only to the needs of NSCB and the planned construction of national indices. Given this scenario, this study explores an alternative estimation procedure that could provide reliable estimates on the current design. Thus, a model-based estimation procedure using the results of the 2001 to 2005 QSPBI is proposed.

II. Related Literature

Probability-sampling designs and inferences based on random sample are widely accepted protocols in sample surveys. Recent advances in the finite population sampling theory however, have supported the proposition that many sampling problems can be realistically formulated as prediction problems under appropriate superpopulation probability models, see for example [6]. This proposes an alternative framework to the inference based on the implied sampling distribution of the selection strategy. The appeal of the theory based on the random sampling plan springs from the fact that the mathematical correctness of the results depends only on the process of random selection, under the complete control of the researcher, see for example [2]. On the other hand, the correctness of results in the prediction theory depends on the validity of superpopulation probability model. The principal benefit from prediction theory-based inference is that it is robust to the sample selection plan. In fact, nonrandom selection plans are considered optimal.

In probability sampling, randomization is introduced in the sample selection plan where implied sampling distribution forms the basis for statistical inference. Random selection avoids unnecessary assumptions about the population and the sample. The aim of sample design is to provide sampling scheme that uses all available information in the most

efficient way to produce estimates of population parameters. If auxiliary information is available, however, there is a great deal of scope for designing a sampling scheme to produce more accurate estimates at lowest possible cost. Estimation procedures that incorporate auxiliary information could be more efficient, yet bias is introduced. The extent of bias is generally dependent on the relevance of the auxiliary information to the sampling distribution of the survey variable.

The relationship between the auxiliary variable and the survey variables can be used in two ways. The traditional design strategy is to use probability sampling and techniques such as stratification, systematic selection, and varying selection probabilities. Measures of error are based on the sampling distribution of estimates considering all possible samples. The second approach is to analyze the relationship between the auxiliary and survey variables in terms of a model and to use the selection and estimation schemes that minimize a measure of error for the distribution of the estimator over all possible replications of the population given the model, see for example [6].

Survey sampling is perhaps unique in being the area of statistics where inferences are primarily based on the randomization distribution rather than on statistical models for the survey outcome. The debate between randomization-based and model-based inference is sharply drawn [4], [5], [3].

Many survey statisticians adopt both design and model-based philosophies of statistical analysis. For example, descriptive inference about finite population quantities based on large probability samples are carried out using design-based methods, but models are used for problems such as non-response or small area estimation.

III. Methodology

The data used in modeling survey data is the matched responding sample establishments for 2001-2004 QSPBI results for Wholesale and Retail Trade comprising of 120 sample establishments. Variables include region, PSIC, ATE, ATECODE, total employment, total compensation and total revenue. On the other hand, results of the 2005 QSPBI, First Quarter for Wholesale and Retail Trade Sector is used as the frame and in simulating the sampling strategies namely current design, stratified SRSWOR and probability proportional to size (pps). It includes 520 sample establishments with the following information: region, PSIC (*Phil Standard Industrial Code*), employment size (ATECODE), total employment (ATE) from the frame which is used as the auxiliary

information, and the 2005 QSPBI first quarter survey results for total employment, total compensation and total revenue.

Using the complete data set of responding sample establishments of 2001-2004 QSPBI for all the quarters, regression models were postulated and estimated for use in model-based estimation. These are (1) linear regression (OLS); (2) poisson regression with log link; (3) gamma regression with log link; and (4) generalized linear model for panel data.

For each regression models, the dependent variables are total employment, total compensation and total revenue while ATE or average total employment from the frame, region, industry stratum, employment size (ATECODE), quarter and year are considered as independent variables.

3.1 Model Building

Poisson regression generally works in cases where the dependent variable or the response variable y_i is skewed as in the case of counts, example total employment. The Poisson distribution $f(y) = \frac{e^{-\mu} \mu^y}{y!}$, $y=0, 1, 2, \dots$ with parameter $\mu > 0$ is a reasonable probability model for count data. The Poisson regression model can be written as $y_i = E(y_i) + e_i$, $i = 1, 2, \dots, n$. The assumption that the expected value of the observed response can be written as $E(y) = \mu$ and that there is a function g that relates the mean of the response to a linear prediction, say $g(\mathbf{m}_i) = \mathbf{h}_i = \mathbf{b}_0 + \mathbf{b}_1 x_1 + \dots + \mathbf{b}_k x_k = \mathbf{x}' \mathbf{b}$. The function g is usually called the link function. The relationship between the mean and the linear predictor is $\mathbf{m}_i = g^{-1}(\mathbf{h}_i) = g^{-1}(\mathbf{x}' \mathbf{b})$.

Gamma regression model, on the other hand, assumes that the response variable y_i is nonnegative and would be expected to have an asymmetric distribution with a long right tail. The Gamma distribution $f(y|\mathbf{a}, \mathbf{b}) = \frac{1}{\Gamma(\mathbf{a}) \mathbf{b}^{\mathbf{a}}} x^{\mathbf{a}-1} e^{-x/\mathbf{b}}$, $0 < x < \infty$, $\alpha > 0$, $\beta > 0$ is a reasonable probability model for nonnegative data. Although, the canonical link of gamma distribution is the inverse link; the log link is often very effective choice as the link function.

The generalized linear model analyzes panel data that consist of time series observations on each of several individuals or cross-sectional units. It has the form of

$y_{it} = \mathbf{a}_i + \mathbf{b}_t + \mathbf{g}_{it} + \mathbf{e}_{it}$, where $\mathbf{e}_{it} \approx NID(0, \mathbf{s}^2)$, independent of $\mathbf{a}_i \approx NID(0, \mathbf{s}_a^2)$ and $\mathbf{b}_t \approx NID(0, \mathbf{s}_b^2)$.

Adequacy of the models are assessed based on the residual analysis, goodness of fit based on scaled deviance and the mean absolute percentage error defined as $MAPE = \frac{|Y - \hat{Y}|}{Y} \times 100$. Since MAPE measures the predictive ability of the model, the lower the value of MAPE the better.

3.2 Predictive Estimation

Predictive capability of the four models considered is also evaluated by estimating the non-sampled observations. The non-sampled observations are predicted through the models developed in this study, which in return are combined with the sampled observations. These combined observations became the Monte Carlo population where the point estimate and the standard error are computed through the nonparametric bootstrap. Following the arguments of (Guarte and Barrios, 2006), estimation based on the current design (non-probability) is feasible through the nonparametric bootstrap.

3.3 Sampling Strategies

Two probability sampling strategies are simulated: stratified SRSWOR and probability proportional to size sampling. Stratification tends to select samples that represent behavior of various sectors. Probability proportional to size sampling, on the other hand, will select samples that are key drivers of sectors they represent. For stratified SRSWOR, 3-digit industry classification (1994 PSIC) and employment size (ATECODE) are considered the stratum and stratification variable, respectively while for pps scheme, the size refers to the average total employment or ATE recorded in the frame. The current design that draws sample establishments based on the top contributors of the production in a sub-sector is used as a benchmark for the sampling strategies.

The sampling strategies are evaluated through simulation assuming 2005 QSPBI First Quarter survey results to be the frame. This simulation will not only assess the viability of the sampling strategies in implementation but also to exhibit some desirable characteristics of the estimation procedure.

Model-based estimation is often constrained by how well the postulated model fits the data. Model fit may sometimes be affected by how “representative” the sample is of the population characteristics needed in model building. A good fitted model may not necessarily imply efficiency in model-based estimation. Thus, the simulation will help in the assessment of the linkages between sample selection, model-fitting, and predictive estimation. The effect of sample selection on the models is likewise addressed/assessed through simulation.

The same is true on the effect of non-response or extent of allowance for non-response where it is also simulated to determine the threshold level that will still generate estimates with minimum bias. Non-response rates of 10-50% (common among establishment surveys) are simulated.

IV. Results and Discussions

The OLS model enables to capture the industry and regional differences of total employment as it revealed significant parameter estimates. For year and quarter indicator variables, the parameter estimates are not significant. This may imply that the technological advancement that the year may explain is not possibly detected in 5-year period or the seasonality that may have been captured by the quarterly dummy variables could have been aptly accounted by the industry stratum, which may exhibit their own seasonality. ATECODE or employment size showed similar behavior as it yield insignificant parameter estimates.

Interestingly, the generalized linear model for total compensation yield significant parameter estimates only in ATE and quarter variables. This means that the model did not capture the uniqueness of the industry, regional differences and the concept of economies of scale is not true. The same holds true for total revenue except for ATE, which shows insignificance while those economies of scale concept are true.

All variables considered (ATE, industry stratum, region, ATECODE, year and quarter) are significant in Poisson model. The significance of the parameter estimates points out that this model adequately accounts for differences in industry stratum, regional location, employment size (ATECODE) and the seasonality (year and quarter). Gamma models, on the other hand, show similar behavior except for total employment, where the parameter estimates of the year and quarter indicator variables are not significant.

Except for OLS regression model for total employment, the results for the current design through model-based estimation procedure are very encouraging as it posted low percentage difference from the population mean as compared to the proposed pps and stratified SRSWOR. This implies that although the estimated means are moderately biased, current design still provides the means closer to the population means. The same holds true for the reliability of the estimates as their CV values recorded the least. These findings are somehow true with the three variables and for all the sample sizes in review.

For the current design, gamma regression and generalized linear model perform slightly well in the mean estimation of total employment and total compensation, as the estimated means are generally closer to the population mean and the recorded CV's are generally less than 10%. Although the percentage difference from the population means of total revenue recorded by the generalized linear model and the OLS regression models have double-digit figure, these models still provide the nearest estimated means of total revenue from the population means among the models considered in this study. Furthermore, large deviations of the estimated means of total revenue could be attributed to the weak linear relationship of ATE with the total revenue.

For design-based strategies, probability proportional to size (pps) sampling gives the closest estimated means to the population means and provides the least MSE values for the three variables as compared with the stratified SRSWOR.

Coefficient of variations of the three sampling strategies, on the other hand, had values that did not differ much from each other. As expected, large variation could be seen in total revenue, which affirms the behavior of the establishment-based data.

Although coefficient of variations recorded by the stratified SRSWOR procedure are sometimes lower than the CVs of pps, still pps fares better than the stratified SRSWOR since pps has consistently recorded lower MSEs and estimated means closer to the population values.

The three sampling strategies behave differently in terms of consistency of estimates. For PPS, the closest estimated means from the population means of total employment and total compensation could be observed when the sampling rate is 10.4% (n=52) and 5.4% (n=28), respectively; generalized linear model when the sampling rate is 5.4% (n=28) and 10.4% (n=52); respectively; Gamma when the sampling rate is 15.8% (n=82) and 10.4%

($n=52$), respectively; Poisson when the sampling rate is 27.9% ($n=145$) and 10.4% ($n=52$), respectively; and stratified SRSWOR when the sampling rate is about 51.5% ($n=268$) for both variables. As expected, higher sampling rate is needed for total revenue to attain the consistency of estimators for the three sampling strategies under study. These results show that pps and the model-based procedures could provide moderately consistent estimator for small sample size than with stratified SRSWOR which needs a larger sample size to achieve this property of the estimators.

Using the sample size $n=268$ (51.5% sampling rate), the effect of non-response is evaluated for the three sampling strategies. The nonresponse and the nonsampled establishments are predicted through the models developed in this study, which in turn are combined with the sampled observations. These observations became the Monte Carlo population where nonparametric bootstrap procedure is used to compute for the point estimate and the standard error of the mean.

Simulations on the effect of nonresponse for total employment show that the three sampling strategies behave differently. For stratified SRSWOR, a response rate of at least 50% could give an estimated mean less than $\pm 2\%$ from the population mean where the nonrespondents establishments are predicted based on generalized linear model. PPS and current design, on the other hand, show that a response rate of 70% and 80%, respectively could give an estimated mean less than 1% from the true mean. Gamma model for the aforementioned designs provided better prediction to the nonresponse. Thus, the threshold levels of nonresponse are 50%, 30% and 20% for stratified SRSWOR, pps and the current design, respectively.

For total compensation, the threshold levels are 10%, 50% and 40% nonresponse for the current, pps and stratified SRSWOR designs. Gamma and Poisson models provide the predicted values for nonresponse, However, deviations from the population value could range from a low of less than 1% to a high of more than 10%. As noted, variability of deviations is much higher in stratified SRSWOR. This is maybe due to the effect of weights, which could possibly explain the erratic behavior on the simulation exercise.

Higher response rate is needed to come up with the total revenue estimate with minimum bias. For instance, 90% response rate is needed in the current design; 80% for the pps design; and 60% in stratified SRSWOR to achieve estimate within less than 5% from the true mean.

It is interesting to note that the current design needs higher response rate for all the variables considered since 41.5% or 252 nonsampled observations are also predicted through the models developed in this study.

V. Conclusions

Four regression models are used in building the models for the QSPBI data of Wholesale and Retail Trade Sector, including Linear Regression (OLS), Poisson Regression, Gamma Regression, and Time Series Cross Section Regression. The models moderately fit the data. However, Gamma model is the most adequate for this survey as its MAPE is consistently lower and scaled deviance is near unity. The results are consistent for all three variables analyzed.

Predictive capabilities of the Gamma, OLS, generalized linear model and Poisson models are evaluated on the nonsampled observations of the current design, and the simulated strategies (PPS and stratified samples). Gamma, Poisson and generalized linear model yield good predictive estimates. Probability proportional to size sampling (pps) provides more reliable but moderately biased estimates than in stratified SRSWOR procedure especially for small sample size.

The nonparametric bootstrap procedure is applied in the Monte Carlo population of the current and the simulated designs to compute for the point estimate and the standard error of the mean. The results show that this procedure, used along with the model-based procedure in predicting nonresponse and nonsampled observations could provide an alternative framework of survey estimation in probability or nonprobability samples.

References

- [1] Guarte, J. and Barrios, E., 2006, Estimation Under Purposive Sampling, *Communications in Statistics-Simulation and Computation*, 32, 277-284.
- [2] Kalton, G., 2000, Developments in Survey Research in the Past 25 Years. *Survey Methodology*, 26, 3-10.
- [3] Kish, L., 1995, The Hundred Years' Wars of Survey Sampling. *Statistics in Transition*, 2, 813-830.
- [4] Smith, T.M.F., 1976, The Foundations of Survey Sampling: A Review. *Journal of Royal Statistical Society*, A139, 183-204.
- [5] Smith, T.M.F., 1994, Sample Surveys 1975-1990; An Age of Reconciliation. *International Statistical Review*, 62, 5-34.
- [6] Zacks, S., 2002, In the Footsteps of Basu: The Predictive Modelling Approach to Sampling From Finite Population. *Sankya: The Indian Journal of Statistics*, 64A, 532-544.