WEIGHTED MULTIPLE REGRESSION ESTIMATION FOR SURVEY MODEL SAMPLING

James R. Knaub, Jr., Energy Information Administration
US Dept. of Energy, EI-524, Washington, DC 20585
e-mail: jknaub@eia.doe.gov and Knaub@gnn.com

Key Words: Variance of Totals, Establishment Surveys, Cutoff Sampling

Abstract:
Model-based inference has performed well for electric power establishment surveys at the Energy Information Administration (EIA), using cutoff sampling and weighted, simple linear regression, as pioneered by K.R.W. Brewer, R.M. Royall, and others. However, 'nonutility' generation sales for resale data have proved to be relatively difficult to estimate efficiently. Design-based inference would be even less efficient. A weighted, multiple linear regression model, using a cutoff sample, where one regressor is the data element of interest as captured in a previous census, and another regressor is the nameplate capacity of the generating entity, has proved to be extremely valuable. This is being applied to monthly sampling, where regressor data come from previous annual census information. Estimates of totals, with their corresponding estimates of variance, have been greatly improved by this methodology. **This paper is an abbreviated version of an article found in the electronic journal, *InterStat*, located on the Internet at http://interstat.stat.vt.edu/InterStat.**

1 Introduction:

Monthly data are currently collected for retail sales of electricity, and for utility generation. A "data gap," or "data need," was identified in that prior to 1996, the EIA had no monthly data for 'nonutility' generation that would help explain why total retail sales data may exceed utility generation data published by the EIA. (There are other factors such as imports and exports of electricity.) The data need is addressed by collecting data on sales by 'nonutilities' to the power grid (i.e., 'nonutility' sales for resale).

Unlike sales by utilities, where that is the primary business, 'nonutility' sales to the grid may not be made to satisfy a particular set of customers. Often, these sales are not the primary business of the 'nonutility.' To use design-based sampling would require an inordinate number of observations before it could theoretically produce a noticeably better representation of the smallest establishments, and may never actually do so because of the proportionately larger nonsampling error one may expect from such establishments. Model-based inference should make efficient use of regressor (auxiliary) data. Burden and timeliness could also be problems that favor the use of a cutoff model-based sample for data collection from these highly skewed establishment data. However, even model-based sampling may require larger sample sizes than one may comfortably gather when dealing with these data, due to high variability. Respondent burden and resource considerations prompted the EIA to try collecting these data with a small, cutoff sample. Note that model sampling is relatively practical to administer, especially since no special imputation procedure needs to be invoked. Cutoff model sampling is particularly simple.

A methodology that has already been seen to work well for EIA electric power data, especially utility sales and revenue, has involved a weighted, zero-intercept, simple linear regression model,

$$y_i = \beta x_i + e_{0_i} x_i^{\gamma} ,$$

where e_{0_i} is the random factor of the residual. (See Knaub(1995).) Cutoff samples have performed very well, and are very practical for these highly skewed data.

When the same methodology was applied to a test for the collection of monthly 'nonutility' sales to the grid, results were not encouraging when using the 50 megawatt cutoff level dictated by burden considerations. This yields n about 400 from N about 2000. Stratification may help, and is now being considered, but clearly, from Knaub(1996), more help was needed.

Several papers addressing related topics can be found over the last few years in the ASA Proceedings of the Section on Survey Research Methods. (These papers include Knaub(1994), and Knaub(1995).) Also, see Brewer(1963), Royall(1970), and Knaub(1993). Linear regression modeling for finite population sampling has proved to be quite useful. The weight used is generally of the form $1/x^{2\gamma}$.

After much experimentation, a better format has yet to be found. Royall(1970) sets γ equal to 0, ½ and 1. These formulations are prevalent in the econometrics literature. (For example, see Maddala(1977) and

Maddala(1992).) To date, the relative performance of $1/x^{2\gamma}$ as the regression weight remains excellent. Whether gamma should be estimated, or set equal to ½, or some other value, is a subject for study in each case. These remarks also appear to substantially apply to the case of multiple regression.

2 Testing for an Appropriate Model for Nonutility Sales to the Grid:

The single regressor model, as shown in Knaub(1996), was not very successful when applied to 'nonutility' sales for resale. This used previous census data as the regressor for a current sample of these wholesale sales. However, because test data results were not good, a second regressor was then sought. Nameplate capacity (a way of rating the capacity for the generation of electricity) was decided upon because it (1) was available, (2) was expected to be positively correlated with 'nonutility' sales for resale, and (3) could never be a negative number (and always positive for existing generators). This was practical to implement for a monthly sample, and test results were good. (See Knaub(1996).) Zero-intercept modeling has proved to be desirable for these data.

The monthly sample survey is now in place and appears to be functioning well. The cv estimates of the estimated totals are based on an extension of what Royall and Cumberland call V_L . (See Royall and Cumberland (1981).) Although V_L is not considered to be a robust estimator, Knaub (1992) contains a figure which shows that it has compared well to V_D , and the latter estimator is considered to be a robust estimator of variance.

3 Formulations:

For the "MR2Z" (multiple regression - two regressors - zero-intercept) case:

$$y_i = \beta_1 x_i + \beta_2 c_i + e_{0_i}/w_i^{1/2},$$

where, for the case of 'nonutility' sales for resale,

y is sales (by a 'nonutility') for resale,
x is the corresponding sales for resale, taken from a census at an earlier time period,
c is the nameplate capacity, and
w is the regression weight such that the inverse square root is the nonrandom factor of each residual. (See Knaub(1995) if interested in a further reference to the nonrandom factor of the residuals, and a unique way to consider such factors.)

Let

$$Q_{MR2Z} = \sum_{i=1}^{n} \left[(y_i - \overset{*}{\beta}_1 x_i - \overset{*}{\beta}_2 c_i) w_i^{1/2} \right]^2 = \sum_i^n e_{0_i}^2 ,$$

and then set $\dfrac{\partial Q_{MR2Z}}{\partial \overset{*}{\beta}_j} = 0$ for j equal to 1 and j equal to 2, and solve for $\overset{*}{\beta}_1$ and $\overset{*}{\beta}_2$. The following results are obtained:

$$\overset{*}{T} = \sum_{i=1}^{n} y_i + \sum^{/} (\overset{*}{\beta}_1 x_i + \overset{*}{\beta}_2 c_i) ,$$ where "T" represents totals, asterisks represent weighted estimates, and the prime symbol indicates summation is over all N-n of the establishments not in the sample.

A multiple regression extension of V_L follows:

$$\overset{*}{V}_{(L)}(\overset{*}{T}) = \sum^{/} \frac{\overset{*2}{\sigma_e}}{w_i} + (\sum^{/} x_i)^2 \overset{*}{V}(\overset{*}{\beta}_1) +$$

$$(\sum^{/} c_i)^2 \overset{*}{V}(\overset{*}{\beta}_2) + 2(\sum^{/} x_i)(\sum^{/} c_i) \overset{*}{COV}(\overset{*}{\beta}_1, \overset{*}{\beta}_2) ,$$

where $\overset{*2}{\sigma_e} = \sum_{i=1}^{n} e_{0_i}^2/(n-2)$;

and where

$$V(\beta_1^*) = \sigma_e^{*2} b_{11}; \qquad V(\beta_2^*) = \sigma_e^{*2} b_{22};$$

$$and \quad COV(\beta_1^*, \beta_2^*) = \sigma_e^{*2} b_{12} = \sigma_e^{*2} b_{21} .$$

$$\beta_1^* = b_{11} \sum_{i=1}^{n} y_i x_i w_i + b_{12} \sum_{i=1}^{n} y_i c_i w_i ;$$

$$\beta_2^* = b_{21} \sum_{i=1}^{n} y_i x_i w_1 + b_{22} \sum_{i=1}^{n} y_i c_i w_i ;$$

$$b_{11} = g/(gr-h^2); \quad b_{12} = b_{21} = -h/(gr-h^2);$$

$$b_{22} = r/(gr-h^2);$$

$$g = \sum_{i=1}^{n} c_i^2 w_i, \quad h = \sum_{i=1}^{n} c_i x_i w_i, \quad r = \sum_{i=1}^{n} x_i^2 w_i$$

Estimation in the case of a non-zero intercept was a little more involved (with three covariances instead of one, and more involved expressions for the parameters), but it was all very easy to program in FORTRAN. (Some may prefer to 'adjust' the use of a packaged program, or write in another language such as SAS, but FORTRAN is highly reliable, flexible and very fast for coding, debugging, and executing.)

4 Conclusions:

Cutoff model sampling has performed well for highly skewed electric power establishment survey data, using simple linear regression with a zero intercept (as shown in Royall and Cumberland(1981)). However, it has been found that using nameplate capacity as a second regressor may have a dramatic impact in some cases, greatly improving estimates of total and variance of total. Improved stability, in the final model selected, appears to be a benefit, in that adjustments to weights can be used to improve estimations, but results are not wildly sensitive to such changes. This model can be convenient to apply to a monthly survey. Imputation is automatic. Using a cutoff sample will ensure that efforts to reduce nonsampling error will not be concentrated inefficiently on the 'smallest' respondents. Also, using the suggested multiple regression model, we will not have to separately account for sales for resale from facilities that had no sales for resale in the regressor data.

This methodology was next applied to generation values for utilities. In general, however, testing thus far indicates that using capacity as a second regressor may have very little impact, as the corresponding 'best' value for beta may be extremely small and often negative. However,

$$\frac{(nameplate\ capacity)\ (\beta(nameplate\ capacity))}{(sales\ for\ resale)\ (\beta(sales\ for\ resale))}$$

was generally fairly small in the above work on 'nonutilities,' yet using the second regressor was very helpful there.

5 Current Work:

Stratification may be particularly important for model-based sampling, as one model may not adequately describe data that should belong under more than one model. Work is being done to make such an appropriate distinction, in an attempt to both increase overall accuracy, and publish less aggregate results.

Another project underway is a study of the variance of the estimate of gamma. The need for such a study had been expressed to me in 1993, by Dr. Michael L. Cohen, in his role as a discussant for a Washington Statistical Society seminar. More recently, K.R.W. Brewer also suggested this. I am currently using one of his suggestions which has quickly produced results in the brief time devoted to this thus far. Results may impact, for example, on findings in Knaub(1993).

6 Acknowledgments:

Thanks are due to K.R.W. Brewer, N.J. Kirkendall and others for helpful input, and to D.R. Bellhouse for constructive criticism. These helpful comments occurred at various stages in the progression of this work. Any disagreement with the final product should be addressed solely to the author who is alone responsible for content and any flaws. (Thanks again, "Mr. Sanders.")

7 References:

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

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," Proceedings of

the Section on Survey Research Methods, American Statistical Association, pp. 876-881.

Knaub, J.R., Jr. (1993), "Alternative to the Iterated Reweighted Least Squares Method: Apparent Heteroscedasticity and Linear Regression Model Sampling," Proceedings of the International Conference on Establishment Surveys, American Statistical Association, pp. 520-525.

Knaub, J.R., Jr. (1994), "Relative Standard Error for a Ratio of Variables at an Aggregate Level Under Model Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 310-312.

Knaub, J.R., Jr. (1995), "A New Look at 'Portability' for Survey Model Sampling and Imputation," Proceedings of the Section on Survey Research Methods, Vol. II, American Statistical Association, pp. 701-705.

Knaub. J.R., Jr. (1996), "Weighted Multiple Regression Estimation for Survey Model Sampling," InterStat, May 1996, http://interstat.stat.vt.edu/InterStat.

Maddala, G.S. (1977), Econometrics, McGraw-Hill.

Maddala, G.S. (1992), Introduction to Econometrics, 2nd ed., Macmillan Pub. Co.

Royall, R.M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," Biometrika, 57, pp. 377-387.

Royall, R.M. and Cumberland, W.G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of its Variance," Journal of the American Statistical Association, 76, pp.66-88.

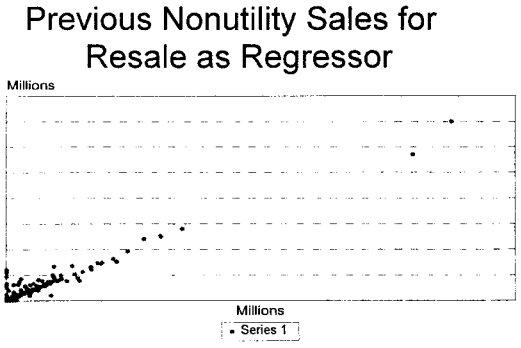
Addendum:

The following model and example graphs summarize the description of 'nonutility' sales for resale data.

Model: $y_i = \beta_1 x_i + \beta_2 c_i + e_{0_i} / w_i^{1/2}$

Graph 1 - Sales for Resale, y, as a Function of Previous Sales for Resale, x

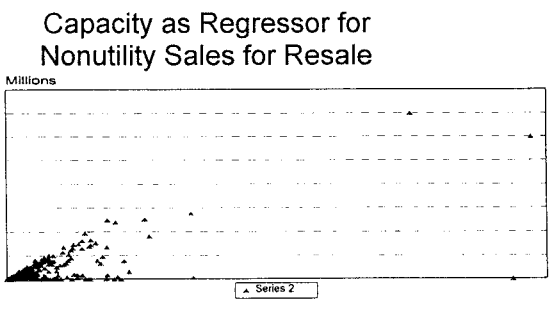
For 'nonutilities,' sales for resale data have considerable model variance. (Utility sales to end user data are generally less variable.) This graph (y vs x) shows a 'typical' example. Note that there are many overlapping points near the origin of this graph, and many of these



points would be excluded from most practical cutoff samples. Note also that data have been altered and scales removed to protect company sensitive information.

Graph 2 - Sales for Resale, y, as a Function of Nameplate Capacity, c

For y vs c (i.e., current sales for resale data as a function of nameplate capacity), large variance for these 'nonutility' data is apparent, and so is a great deal of heteroscedasticity. There are many overlapping points



near the origin. Most of them would be excluded in a cutoff model sample. Note that data here, as in Graph 1, have been altered and scales removed to protect company sensitive information.