A NEW LOOK AT 'PORTABILITY' FOR SURVEY MODEL SAMPLING AND IMPUTATION

James R. Knaub, Jr., Energy Information Administration
US Dept. of Energy, EI-524, Washington DC, 20585

Key Words: regression weights, factors of residuals, heteroscedasticity, establishment surveys

Background:

The relationship, $y_i = bx_i + e_{0_i}x_i^{\gamma}$, can be very useful for modeling electric power data, especially when the variate of interest is the same as the regressor, but for a more frequent and more recent time period. For Energy Information Administration (EIA) electric power establishment survey data, using a variation of this model, where γ is set equal to 0.5, has been shown in various empirical tests to be robust for estimating totals and variances. (Totals seem to be generally quite accurate, although variances seem somewhat less accurate, but generally good. Variance estimates, especially in the cases where variance is particularly small, appear to tend to be overestimates. See Knaub(1993) and Knaub(1994b) for more on this model with γ set equal to 0.5.) The subject of this paper is to consider what happens with regard to this modeling when data are collected at varying levels of aggregation.

Note that although the robustness of using $\gamma = 0.5$ in the above model seems quite obvious for tested EIA data, conversations with Raymond L. Chambers and Kenneth R. W. Brewer (Australian National University) indicate otherwise for other data sources. Perhaps survey data source 'types' may someday be identified which would best be modeled using given values of gamma, and/or estimated values of gamma, and/or perhaps other expressions for the nonrandom factor of error. (That is, expressions other than x_i^{γ} might be used, and thus one would have corresponding regression weights with formats other than $x_i^{-2\gamma}$, and therefore, the variance function would be something other than $x_i^{2\gamma}$.) Note that in this paper, the model error term is considered to be factored into the "nonrandom factor of error" and the "random factor of error" as shown, respectively, here: $e_{0_i}x_i^{\gamma} = x_i^{\gamma} \cdot e_{0_i}$

K.R.W. Brewer noted that the e values are not actually errors, but residuals, and further, that there is a distinction to be made "....between randomness and homoskedasticity. The e_{0_i} are both random and homoskedastic (equal in variance). The $(e_{0_i})(x_i)^{\gamma}$ are equally random, but they are heteroskedastic, and the variance function describes the way in which their variances differ from unit to unit."

Just as a variety of distributions and/or parameters are used in engineering/reliability work, depending upon the nature of the data, so also might the nonrandom factor of 'error' expression change due to characteristics of the survey data that we may be able to use to categorize these data. This approach would seem to agree with Thompson (1995), when he advocated using a nonparametric data analysis (perhaps Exploratory Data Analysis) and a model to "iteratively" improve on each other. Here, if characteristics of the data could be used to help determine the model format, and then the data used to estimate one or more parameters, then test results could be compared to what would have been obtained using other expressions and/or parameter values. Thus, our 'rules' for guessing expressions might be improved.

Introduction:

In Brewer, et.al.(1977), the use of γ as a "portable" parameter is discussed with regard to estimating variance from one survey based on another, where the x-variate represents a cluster size. This has become of interest to me for use in electric power surveys, except that I am not using the cluster size as the x-variate. My current interest in this topic was peaked when it first appeared that future sampling of electric power generators may be feasible at the owner/operator level

as opposed to the plant/facility level. Each owner/operator may operate multiple plants. There are advantages to collecting data at the more aggregate level. The burden to the respondents and the processing time for preparing monthly reports should be substantially improved at that level, as well as smoothing the problem of occasional plant "down time" for maintenance, when addressing the generation of electricity. Although this scenario was the catalyst for this study, these data may not ultimately be collected this way. As a possibility, however, this is discussed in Knaub(1995). Regardless, this study has resulted in information that has possible significance to other data collections.

Question:

In the case of using a more aggregate data collection level, it may be important to determine if any information on a less aggregate data collection level might be of use, or *vice versa*. By relating a census to a previous census for a given variate and data collection level, and then repeating at another data collection level, could we form any conclusions that might be useful for imputation and/or sampling? (Note that the estimated value of γ for a given data collection level, for a given variate, does not appear to be stable over time and/or subsamples. At least, that has been the experience for Nancy Kirkendall at the EIA, and my experience also, using various EIA data. In that sense, one does not see the kind of 'portability' that would allow use of one set of data to estimate γ, and another set to estimate totals and variances of totals, as would ordinarily be desirable.)

Research/Results:

Starting with real data, I have established empirically that when x represents the data of interest in the previous census, and y represents such data from a current census, the value estimated for γ for the more aggregate level of data collection is usually closer to 0.5 than the estimated value of γ for a less aggregate data collection level. Also, when I artificially generate data to follow the zero-intercept model, $y_i = bx_i + e_{0_i} x_i^{\gamma}$, and then study what happens when the x_i values are clustered and a model of the same format is applied, I generally estimate γ for this aggregate data collection level modeling to be closer to 0.5 than the original γ value. (Therefore, generally, $\gamma_a \leq \gamma_d$, where γ_d is the γ value for the more disaggregate data collection level model, and is usually between 0.5 and 1.0. Further, γ_a is the estimated value for the more aggregate data collection level model.)

When the clusters are composed of cases where the x values are nearly equal, the result is that $\gamma_a \approx \gamma_d$. That is, if plants neighboring in size are covered by the same owner/operator, then the model applied to the more aggregate level should use about the same γ value. Otherwise, γ generally decreased for the more aggregate data collection level cases. Knowledge of this approximate upper bound might be helpful.

Generally, estimates of γ for our data are from 0.5 to 1.0, and very often, from 0.7 to 1.0. However, estimates also appear to often vary for different ranges of x (Knaub(1994b)), so the x^{γ} expression for the nonrandom factor of error, using EIA electric power data, is usually not very closely followed (in my experience). From Knaub(1993), different estimators for γ yield approximately the same results when the model is strictly followed, but tend to disagree when the model is not so strictly followed. (Michael L. Cohen suggested a simulation study to test the sensitivity of departures in two particular measures of γ found in Knaub(1993), when varying degrees of model failure exist, but this has not been done at this time.) It seems that this model does have a lot to offer, but strict adherence to the error structure is sometimes unadvisable. If, however, as stated earlier, this model

is still basically used, but with γ set equal to 0.5, performance is usually good for these EIA data. Thus, if the estimated value for γ actually comes closer to 0.5, this might be expected to further improve the usefulness of that value of γ.

If we let k represent clusters and q represent the members of the cluster, then let us represent a situation where $\gamma_d = \gamma_a$ (i.e., when we have portability) by letting $x_{kq} = x_k$ for all q. (There may be other ways to obtain portability, but this is the only one apparent from studying the data.) Using our standard model, but summing for each cluster,

$$\sum_q^{n_k} y_{kq} = y_{k*} = \sum_q^{n_k} (b_d x_{kq} + e_{0_{kq}} x_{kq}^{\gamma_d}) \text{ is obtained.}$$

Letting $x_{kq} = x_k$ *for all* q, means that

$$y_{k*} = (\sum_q^{n_k} b_d) x_k + (\sum_q^{n_k} e_{0_{kq}}) x_k^{\gamma_d}.$$ Also, for simplicity, letting clusters be of equal size, $n_k = \eta$ *for all k,* and letting $x_{k*} = \eta x_k$ be the more aggregated level regressor values, write $y_{k*} = b_d x_{k*} + (\overline{e_{0_{kq}}}/\eta^{\gamma_a - 1}) x_{k*}^{\gamma_a}$, where $b_a = b_d$ is the slope for the more aggregate level model, $\gamma_a = \gamma_d = \gamma$, *and* $\overline{e_{0_{kq}}}$ is an "average" representation of what would be obtained for the less aggregate level random factor of error.

The error term in the more aggregate level model is $\eta \ \overline{e_{0_{kq}}} \ x_k^{\gamma}$, which is written in terms directly traceable to the less aggregate data collection level model. This is a situation where γ is portable, which seems to be nearly true in many EIA electric power data cases. However, generally, it appears that

$$\gamma_a \leq \gamma_d.$$

In practice, it may be obvious that in some situations, results are far more sensitive to γ than for other situations. When estimates of prediction error variances (Knaub(1994a)) are substantially impacted by small changes in the γ value used, then the nonrandom error factor may have been relatively large.

An aside:
For estimates of the variance of the slope in the current zero-intercept model, artificial data strictly adhering to the model show an interesting phenomenon when the actual value of gamma is 1. In that case, the estimated variance of the slope does not change regardless of the value of gamma used in the estimations. That is, when the true value of gamma is 1, and the model is strictly followed, the estimated variance of the slope is unaffected by the value set for gamma. K.R.W. Brewer, in unpublished correspondence, noted that if "h" represents a selected value for gamma, and "γ" represents the 'true' value for gamma, then "... $\sigma^2(b(h))$ has $\sum x^{2\gamma - 2h}$ in the numerator and $\sum x^{2-2h}$ in the denominator. When gamma is equal to anything except unity, the choice of h matters, but when gamma = 1 the expressions in numerator and denominator cancel out, so the choice of h has no effect." This confirms the results of examples using artificial data. Thus, if the model were strictly applicable, we may still do well if we estimate low for γ when $0.5 \leq \gamma \leq 1.0$. If the true value of γ is 'large' (say, close to 1), it may appear to be better to underestimate it, than to overestimate a γ value that is actually 'small' (say, close to 0.5). Perhaps this is part of the reason for the successful use of $\gamma = 0.5$, whether or not the data 'strictly' adhere to the model, given at the beginning of this paper.

CONCLUSIONS:

Collecting data at a more aggregate level may mean that using the square root of x as the nonrandom error factor might be even more robust than ever in the future for the EIA. This is because the estimated value of gamma tends toward 0.5 as the degree of aggregation for the data collection rises. Thus the estimated value of gamma, and the value 0.5, found to perform very well for much of the data collected by the EIA (Knaub(1993), Knaub(1994)), would come more closely into agreement. Also, when the model is more strictly followed, it is probably better to underestimate than to overestimate gamma, at least with regard to estimating the variance of the 'slope' parameter.

Still, when any estimated value of gamma is substantially different from 0.5, it may be difficult to determine the values for gamma that one might prudently use when estimating totals and variances of totals.

Near portability of γ often occurs when differing levels of aggregation for data collection are used. This is a rare degree of stability for γ among these data. Perhaps a strong use for this knowledge will become apparent in the future.

Suggestion:

Perhaps the optimal value of γ to use, in the estimation of variances of totals, might normally be found between 0.5 and the estimated value of γ. (This was suggested in Knaub(1993), but might be more urgently needed when employing more aggregate data collection levels. The smaller numbers of observations at more aggregate levels will mean that refinements may have more of an impact.)

POSSIBLE FUTURE CONSIDERATION:

How would all of this be affected by the use of an expression for the nonrandom factor of error other than x^{γ} ?

Acknowledgement:

Thanks to K.R.W. Brewer for helpful correspondence, although he can in no way be considered responsible for any errors or inadequacies in this paper.

References:

Brewer, K.R.W., Foreman, E.K., Mellor, R.W., and Trewin, D.J. (1977), "Use of Experimental Design and Population Modelling in Survey Sampling," Bulletin of the International Statistical Institute, 47, pp. 173-190.

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. (1994a), "A Formulation for the Variance of the Prediction Error for Weighted Least Squares Simple Regression," unpublished manuscript.

Knaub, J.R., Jr. (1994b), "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), "Planning Monthly Sampling of Electric Power Data for a Restructured Electric Power Industry," Data Quality, 1, pp. 13-20.

Thompson, J.R. (1995), "Postmodern Data Analysis: The End of Statistics" (presented as a Washington Statistical Society Seminar, Washington, D.C., May 8, 1995).

Comments via e-mail are welcomed. Please send your messages to JKNAUB@EIA.DOE.GOV.

Appendix 1

Thanks to Phil Kott for bringing the following to my attention while in Orlando:

In W.G. Cochran's first edition (1953) of Sampling Techniques, pages 210 - 212, he makes use of the

assumption that " $E(e_i^2) = az_i^g$ " in a comparison of two estimators for totals. This seems to have launched a great deal of study using x^γ as the 'nonrandom factor of error,' notably in works by K.R.W. Brewer and R.M. Royall. I also note that econometrics often relies on this. Perhaps, however, it is now time that we explore some other functional forms as well.

Appendix 2

The following is an example of the use of artificial data as mentioned in this paper:

Let Dataset I be made up of 120 points, which consists of three sets of the same 40 points. These points were artificially generated to lie on a regression line with slope = 1, and γ = 1. (Thus, in Appendix 1, we would have g=2.) The data points were generated to reflect only the nonrandom factor of 'error,' so that a nearly exact value of γ could be represented. For every point 'above' the regression line, there is one symmetrically placed below it, and there are other similar pairs of points, all with residuals of the form $\pm c_j x_i^\gamma$. For one or more values of "j" (here, j=1,2), let i range (here, i=1,2,3,...,10) so that in this case there are a total of 40 points, repeated twice, or a grand total of 120 data points. Here, γ is 1. The 40 data points are as follows:

x	y	x	y
1000	1100	1000	900
2000	2200	2000	1800
3000	3300	3000	2700
4000	4400	4000	3600
5000	5500	5000	4500
6000	6600	6000	5400
7000	7700	7000	6300
8000	8800	8000	7200
9000	9900	9000	8100
10000	11000	10000	9000
1000	1050	1000	950
2000	2100	2000	1900
3000	3150	3000	2850
4000	4200	4000	3800
5000	5250	5000	4750
6000	6300	6000	5700
7000	7350	7000	6650
8000	8400	8000	7600
9000	9450	9000	8550
10000	10500	10000	9500

Dataset II is formed by adding coordinates for each set of three identical points. (That is, the numbers above are each multiplied by 3, and only those 40 points are used.) Here, γ = 1 again.

For Dataset III, data points that are usually similar, but not identical, were combined. Here, coordinates from the 120 data points in Dataset I are added in groups of three in succession (so that the first y value is 1100 + 900 + 2200 = 4200). As predicted, the value for γ for the third dataset is smaller than in the other two cases. The Iterated Reweighted Least Squares method estimates γ = 0.84, and my method estimates γ = 0.83. (See Knaub (1993).) (Also note that in this case, the variance of the estimate of the slope parameter is smallest, being even smaller than for the case of n=120.)