Applied Multiple Regression for Surveys with Regressors of Changing Relevance: Fuel Switching by Electric Power Producers

James R. Knaub, Jr.
US Dept. of Energy, Energy Information Administration, EI-53, Washington, DC 20585

Keywords:
prediction; imputation; small area estimation; regressors; regression weights; model performance measure

Abstract:
This research concerns multiple regression for survey imputation, when correlation with a given regressor may vary radically over time, and emphasis may shift to other regressors. There may be many applications for this methodology, but here we will consider the imputation of generation and fuel consumption values for electric power producers in a monthly publication environment. When imputation is done by regression, a sufficient amount of good-quality observed data from the population of interest is required, as well as good-quality, related regressor data, for all cases. For this application, the concept of 'fuel switching' will be considered. That is, a given power producer may report using a given set of fuels for one time period, but for economic and/or other practical reasons, fuel usage may change dramatically in a subsequent time period. Testing has shown the usefulness of employing an additional regressor or regressors to represent alternative fuel sources. A performance measure found in Knaub(2002) is used to compare results. Also, the impact of regression weights and the formulation of those weights, due to multiple regression, are considered.

Introduction:
The Energy Information Administration (EIA) collects energy data, including data on generation of electricity and consumption of fuels, for the electric power industry in the US. An annual census is collected, as well as a monthly sample. The distribution of electric power plants is extremely skewed with regard to measures of plant size, as is common with establishment surveys. Thus a monthly census would require inordinate resources to appropriately limit nonsampling error for the smallest plants. These resources could be put to better use. In addition, it may not even be possible to collect high quality data for many of the smallest plants on a monthly basis, regardless of EIA resources, because the respondents may not be willing and/or able to provide such data so frequently. The smallest respondents appear to have the greatest difficulty in providing high quality data monthly. Thus cutoff sampling has been developed which could be thought of in terms of mass imputation for the smallest plants, or as small area estimation using regression and a form of 'borrowing strength' that is described in terms of "estimation groups" vs. "publication groups" in Knaub (1999, 2000, 2001). The cutoff levels are based on plant capacity and vary by sub-populations/estimation groups. An end-of-year census can be used to verify annual totals of previous monthly estimates.

Data are required by fuel type, but many plant generators are capable of using more than one type of fuel. This is referred to as "fuel switching." The best regressor is often the same data element for which predictions are made, but from a previous annual census. However, when fuel switching is present, the volume of generation or consumption for a fuel in the most recently completed and released census, for example, generation from natural gas-fired, regulated plants in the Northeast, may have largely been replaced in the current month, say, by petroleum-fired generation for those plants.

How do we adequately predict for these data elements? Generally, as stated above, the best predictor for a given data element in a sample is the same data element from a recent census. In the case of fuel switching, this may no longer be true. Generator nameplate capacity may often be a good additional regressor, but its correlation to many data elements has appeared to be weak. One could try stratifying plants by those that may fuel switch and those that cannot switch fuels. However, a preliminary investigation indicated that this would be a substantial task, and the process could remain maintenance intensive. This is not desirable in a monthly data publication production mode. It would be better to find regressors that would help improve accuracy when there is fuel switching, but would not degrade performance substantially when fuel switching is not present. One may be able to determine the fuels that could possibly be involved in fuel switching, and model all others in the usual way. However, as discussed below, that may not be necessary.

This method will estimate nonzero amounts of fuel for some respondents that have switched away from that given fuel, and will further overestimate in additional cases. However, it will also underestimate for that same fuel in other cases. The overall accuracy in estimating for (sub)totals is what is of greater importance. No individually imputed number should be published, but only used internally for analysis and data editing.

Note that data for fuels that can be switched could be added and treated at a more aggregate fuel level, but that would probably more severely limit publication possibilities. In fact, this has been done.

Possible Regressors:
What regressors would be best? Will collinearity be a problem? What regression weights would be helpful? These are questions that could vary with data sets, but a fairly general solution is needed for a survey production environment. Some testing has proved helpful.

Artificial test data were constructed to simulate conditions where there is no fuel switching (the "No Fuel Switching Occurs" test data set), and conditions where fuel switching does occur in many instances (the "Some Fuel Switching Occurs" test data set). (See the graphs below.) Under this latter scenario, there are cases where no fuel switching is evident, cases where it is evident that plants would have switched to the fuel type or energy source of interest, and cases where plants would have switched away from the energy source of interest. There may be times when, due to a drop in price for a given fuel, many plants will switch from one fuel to another in a more systematic fashion. Under the "Some Fuel Switching Occurs" test data set used, however, as just stated, all 'switching' was 'allowed.' A set of regressors that can deal effectively with this, yet remain useful for the "No Fuel Switching Occurs" test data set case, without stratifying for these characteristics, might be helpful in a monthly data publication production environment. Putting aside nameplate generating capacity as a regressor, and putting other data or processing problems aside that may have complicated any real data that would have been obtainable for this study, the author chose to use artificial data that could demonstrate the efficacy of the resulting methodology. Further testing in the future, using 'real' data, and perhaps more artificial data, would be advisable. (See the last table.) In these test data sets, y represents monthly data (say generation, or perhaps fuel consumption) for a given energy source (the fuel type or energy source of interest). There were four regressors used: x_1, ostensibly the data element of interest collected on a previous annual census; x_2, representing a related data element for which 'switching' could take place; and x_3 and x_4, which are generally unrelated to y. The values for y could represent generation using a given energy source, or they could be fuel consumption values. The values for x_1, x_2, x_3 and x_4, or at least x_2, x_3 and x_4, could all represent generation of electricity, to avoid having to convert to common units when considering a sum of these values as a regressor, and/or in regression weights, as will be done further below. Since x_2 is to be considered a better regressor than x_1 for predicting y in many cases, it will sometimes be more important than x_1, but sometimes it will be virtually irrelevant. The idea is to find a regression model form that will work well under all scenarios, and not halt production due to collinearity or any other problem.

Regression Weights:
The linear model used is as follows:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + e_{0i} z_i^{\gamma},$$

where z_i is a measure of 'size,' and usually $0<\gamma<1$. This format is discussed in Knaub(1997). The intercept is assumed to be zero since y should approach zero as all x's approach zero. Brewer(2002) also finds a zero intercept to be more useful. Further, he has interesting things to say about γ.

In terms of regression weight, $w_i = z_i^{-2\gamma}$, one has the following:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + e_{0i} w_i^{-1/2}$$

What would be a good z if x_1 represents the same fuel type as y, and there may or may not be fuel switching? An estimate for y, $z_i = \hat{y}_i$, was used in Knaub(1999) and elsewhere, but this requires that for each model application, that is, for each "estimation group" in Knaub(1999, 2000, 2001), the model must be applied at least once prior to full implementation. For the preliminary application, one could use $\hat{y}_i = \sum_j x_{ji}$, so that all coefficients are set to unity. This could make total generation for all fuels, like nameplate capacity for a plant, a measure of size for that plant. Another possibility is to start with ordinary least squares (OLS) regression, letting all data

points be weighted equally. It would be odd if an estimated value of 10 and an estimated value of 50,000 and all other estimated values had the same standard error, say 100, which is what is assumed with OLS regression. However, this could be used to obtain better z_i values by using the resulting estimated coefficients, b_j, in $z_i = \hat{y}_i = \sum_j b_j x_{ji}$. The first exercise of the regression software for a given estimation group could provide these b_j values, and they could be updated from time to time, or automatically in the software each time.

When there is no fuel switching, the coefficient for the regressor representing the same fuel in a past census that is of interest in the current sample, b_1 here, will be most influential, so that it is likely that $\sum_i b_1 x_{1i} > \sum_{j \neq 1} \sum_i b_j x_{ji}$. When there is fuel switching, more than one regressor may be very influential, and thus it may be best to at least loosely account for that when determining the z that will be used in determining the regression weights, $w_i = z_i^{-2\gamma}$.

The residual term in $y_i = b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i} + z_i^{\gamma} e_{0i}$ may be amenable to being partitioned into various parts, with respect to different regressors, but there may be appreciable interactions, and thus far, this does not appear to be practical.

Another possibility would be to use $w_i = (x_{1i} + a)^{-2\gamma}$, where a is a nominal constant, so that w_i is never zero. If we use $w_i = c_i^{-2\gamma}$, where c is the plant capacity, then w_i should not be zero in any case either, as long as zero capacity is not carried in any data record. If it were, then this could be "planned capacity" that has not yet been built, and it should be ignored here.

Results should dictate what regression weights would eventually be used, or dictate a procedure for periodically updating them. Perhaps the best option would be to start with an OLS estimate of y, say $\hat{y}$, and let $z_i = \hat{y}_i$. Then obtain the WLS estimate, y^*, using that. Updates should be made as often as practical. It would be best if software were written to update z and then run a second time with the new z for each estimation group.

In the study below, however, regression models were tested using $z_i = \hat{y}_i = \sum_j x_{ji}$. The purpose of the exercise was to determine appropriate regressors.

In the study results below, when all four regressors, x_1, x_2, x_3 and x_4 were used, or when x_1 and the total of the other regressors, $toto = x_2 + x_3 + x_4$, were the regressors used, then $w_i = z_i^{-2\gamma}$, where $\gamma = 0.5$, and $z_i = x_{1i} + x_{2i} + x_{3i} + x_{4i}$. However, when only the x_1 regressor was employed, then z_i was set equal to x_{1i}, and when only toto was used as a regressor, then z_i was set equal to $x_{2i} + x_{3i} + x_{4i}$. Knaub(1993), Knaub(1997, pp. 9 and 10), and Knaub(2001) discuss the robust nature of using $\gamma = 0.5$, which is the ratio estimate. The degree of 'fuel switching' and the formula for z are further considerations, along with the "thermometer effect" shown in Knaub(2002), that may make $\gamma = 0$ a good starting place at times. Here we study $\gamma = 0.5$ to see the impact with these artificial test data. Remember, however, that the regression weights are a function of both γ and z.

Comparison of Results:

When using prediction-oriented software (Knaub(1999, 2000, 2001)), the sum of square errors (SSE) may be only the sum of the squares of the 'random factors' of the residuals (Knaub(1993, 1997, and others)). (KRW Brewer notes that these are "factors," not "components" as misstated in Knaub(1993).) That would be the sum of the squares of the e_{oi} shown above. To compare results for fuel switching and no fuel switching cases, for different regressors, we need to be able to compare under weighted least square (WLS) conditions. Also, because the regressor coefficients are estimated, their variability must be taken into account. To compare overall performance then, one could estimate the variance of the total when all observations are replaced by imputed values. Thus we may use the model-based form of the relative standard error under a superpopulation (RSESP), as described in Knaub(2002), as a performance measure. These

values are shown in the table below. In all cases here, $\gamma = 0.5$. Graphs are found below that illustrate the artificial data being used as input to this process. (The last table concerns some actually observed data.)

Conclusions:

Examination of these study results indicates that adding a regressor related to each fuel that could possibly be substituted ('switched') for each other, provides very good predictive 'power' for a variate relevant to any one of those fuels. If 'fuel switching' has occurred, this is very helpful. If fuel switching has not occurred, there is little if any harm done here. Collinearity was investigated as a concern, but it does not appear to be a problem. In an additional test, two regressors were added with all zero values. The prediction software program used declared that there was collinearity, but then dropped those two regressors and produced the same results that would have occurred had those two pseudo-regressors never been introduced. If, however, there were any danger of collinearity that could interfere with monthly production, then it may also be noted that a very large portion of the benefit obtained by adding regressors for each of the possible substitute fuels could still be obtained by combining regressors for all substitutes into one additional regressor. To do that, the regressors to be combined should probably all be in the same units. That would mean perhaps using electricity generation as regressors for each fuel type, even when the variate of interest is consumption for the fuel of interest. That is because generation can be in common units, but consumption units frequently vary. (Fuels can be in the various states: gas, liquid, solid.)

The benefits of using another regressor or regressors when fuel switching is possible appear substantial. It would not be necessary to attempt to stratify by those plants that could or could not have fuel switching. Also, it would not be necessary to attempt to separate fuels that can be switched with one another, from those that can not be so substituted, although this might be operationally helpful. Simplicity is maintained, which is very beneficial in a monthly data production environment. The suggestion that may be most difficult to implement would be programming to obtain z-values for regression weights with one iteration, and completed results on a second or subsequent iteration. However, this could be very worthwhile too.

Acknowledgements:

The author thanks other Federal Government and contractor staff, particularly Dr. Orhan Yildiz, for helpful discussions.

References:

Brewer, KRW (2002), Combined Survey Sampling Inference: Weighing Basu's Elephants, Arnold: London.

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. (1997), "Weighting in Regression for Use in Survey Methodology," InterStat, April 1997, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 1997.)

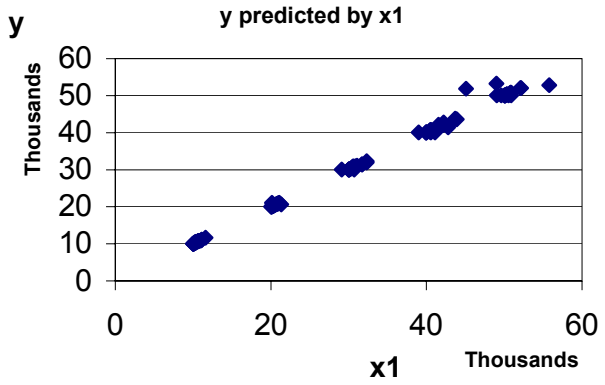
Knaub, J.R., Jr. (1999), "Using Prediction-Oriented Software for Survey Estimation," InterStat, August 1999, http://interstat.stat.vt.edu, partially covered in "Using Prediction-Oriented Software for Model-Based and Small Area Estimation," in ASA Proceedings of the Survey Research Methods Section, 1999, and partially covered in "Using Prediction-Oriented Software for Estimation in the Presence of Nonresponse," presented at the International Conference on Survey Nonresponse, 1999.

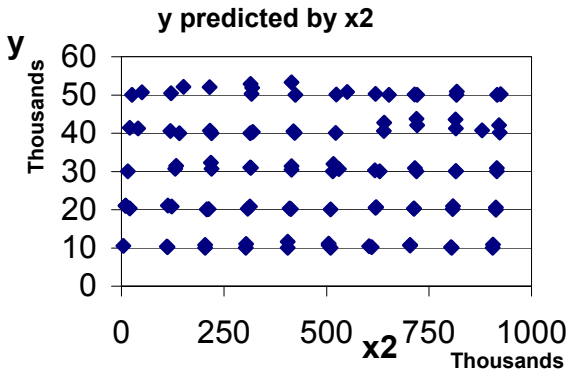
Knaub, J.R., Jr. (2000), "Using Prediction-Oriented Software for Survey Estimation - Part II: Ratios of Totals," InterStat, June 2000, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2000.)

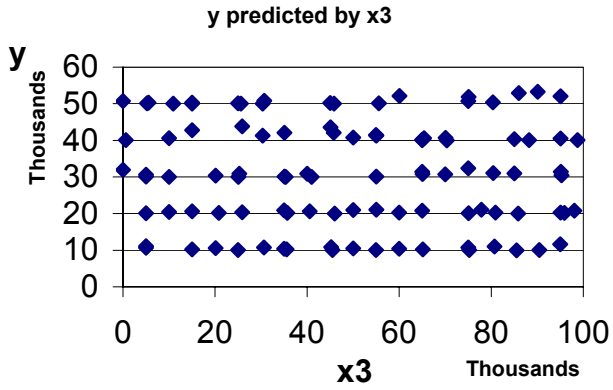
Knaub, J.R., Jr. (2001), "Using Prediction-Oriented Software for Survey Estimation - Part III: Full-Scale Study of Variance and Bias," InterStat, June 2001, http://interstat.stat.vt.edu. (Note shorter, more recent version in ASA Survey Research Methods Section proceedings, 2001, JSM CD.)

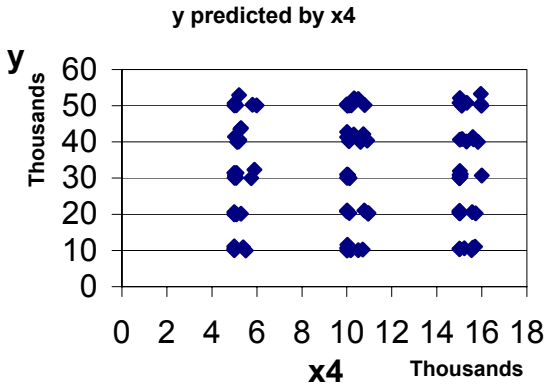
Knaub, J.R., Jr. (2002), "Practical Methods for Electric Power Survey Data," detailed version, InterStat, July 2002, http://interstat.stat.vt.edu with Korn and Graubard correction on superpopulations. (Note shorter version in ASA Section on Government Statistics proceedings, 2002, JSM CD.)

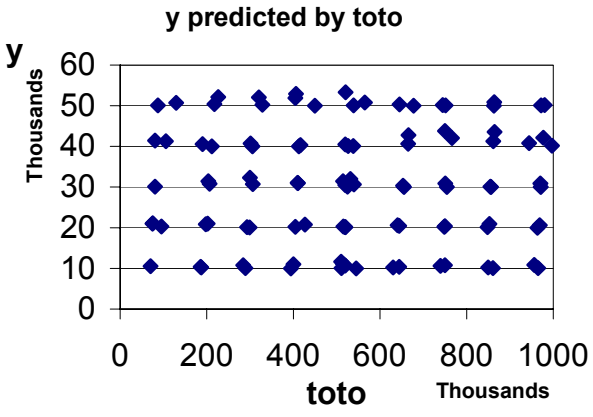
Test Data: No Fuel Switching Occurs

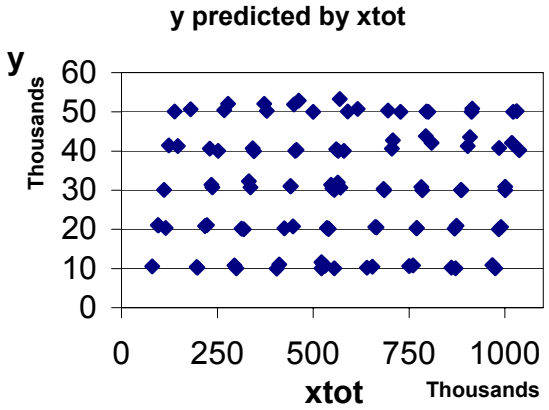






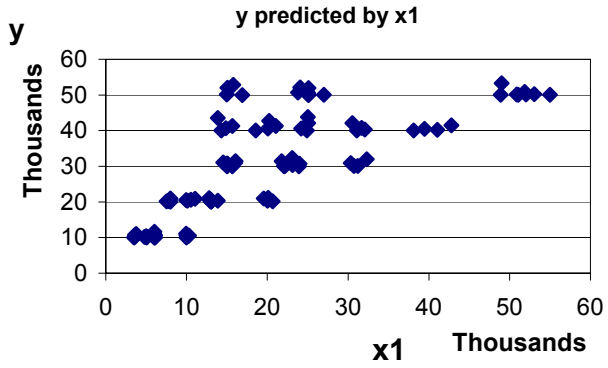


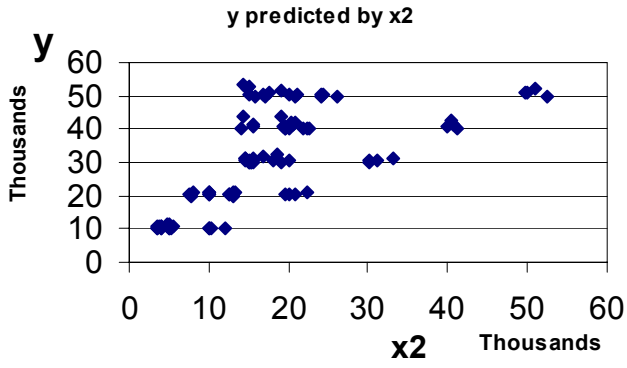


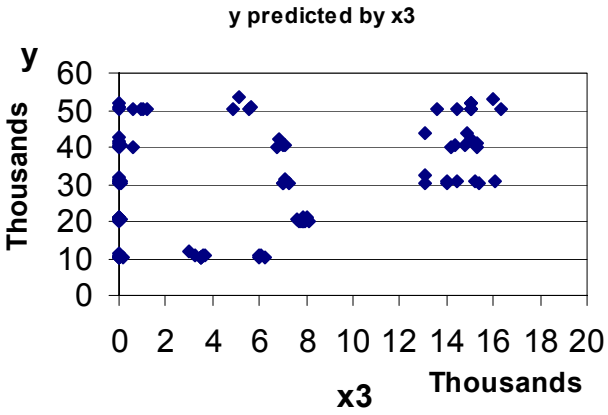


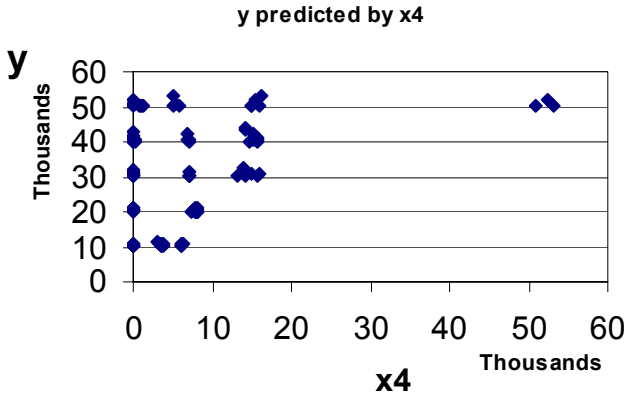
Note: toto = $x_2 + x_3 + x_4$ and xtot = x_1 + toto

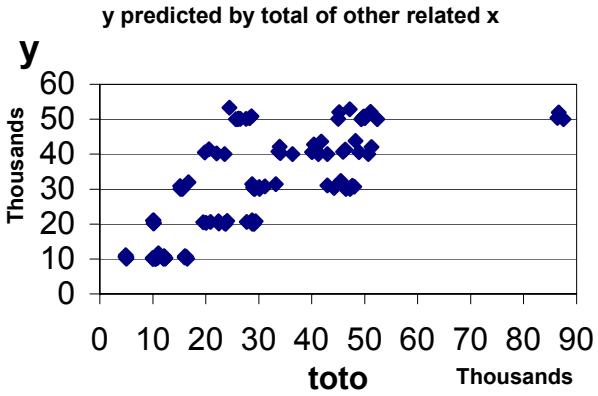
Test Data: Some Fuel Switching Occurs

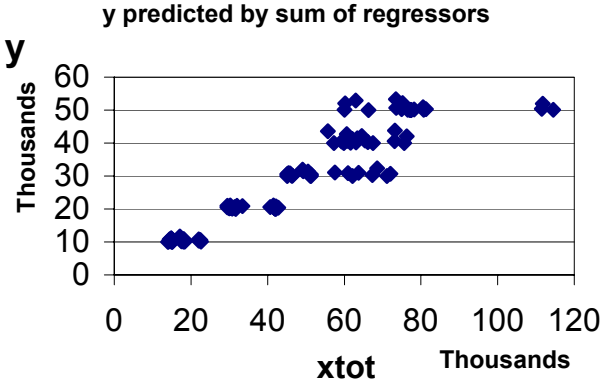












Note: $toto = x_2 + x_3 + x_4$ and $xtot = x_1 + toto$

RSESP* Estimates for Artificial Data

Fuel Switching Allowed

Regressors	Sum(e_i^2)	Var due to coefficients	total variance	RSESP
XF1, XF2, XF3, XF4	2,152,776,960	1,968,930,468	4,121,707,428	2.09%
XF1 and TOTO	3,061,120,000	2,432,898,000	5,494,018,000	2.41%
XF1 only	17,430,527,000	14,742,935,000	32,173,462,000	5.84%
TOTO only	15,178,169,000	15,298,390,000	30,476,559,000	5.68%

No Fuel Switching

Regressors	Sum(e_i^2)	Var due to coefficients	total variance	RSESP
XF1, XF2, XF3, XF4	83,476,608	112,104,923	196,581,531	0.46%
XF1 and TOTO	83,751,000	111,916,000	195,667,000	0.46%
XF1 only	82,695,000	56,826,000	139,521,000	0.38%
TOTO only	43,801,000,000	87,084,000,000	130,885,000,000	11.78%

*RSESP refers to a measure of relative standard error for the superpopulation from which the subject population and any sample were drawn. See “RSESP” in Knaub(2002) JSM CD & ***InterStat***. On page 4 of Knaub(1999) in ***InterStat***, the ‘exact’ variance estimate shown for totals is used here, except that summations above are over N rather than N-n.

Results from 'real' (actually collected) test data:

Natural Gas – A 'Switchable' Fuel

These are RSE estimates, in percent, from a census with simulated missing data. Here, regulated utilities are considered, by region, for generation of electricity using natural gas. Regressors used were XF1, natural gas generation from a previous census, C, the nameplate capacity, and XTOTO, the total of generation for other fuels in a previous census. The base of the regression weight, in the test cases shown below, is the sum of the regressors. Often, that may not be very effective, and the two-stage approach described earlier may be recommended.

REGION	RSE using XF1 only		RSE using XF1 and C	RSE using XF1 and XTOTO	RSE using XF1, C and XTOTO
	gamma = 0.8	gamma = 0	gamma = 0.8	gamma = 0.8	gamma = 0.8
USTotal	3.50	2.71	2.01	0.30	0.27
NewEngland	4.52	124.38	4.52	5.16	5.14
MidAtlantic	0.32	4.42	0.32	0.22	0.22
EastNoCent.	1.88	26.98	1.87	1.91	1.90
WestNoCent.	43.04	79.93	27.14	4.95	4.92
SouthAtlan.	4.94	2.76	3.17	0.39	0.39
EastCentral	11.97	4.05	0.83	0.94	0.11
WestCentral	4.10	3.01	2.63	0.33	0.33
MountainRe.	16.19	9.65	10.39	1.31	1.30
PacificCon.	31.00	11.06	19.93	2.46	2.46
PacificNon.	2.48	8.04	2.46	1.54	1.53
CA	40.86	14.59	26.30	3.25	3.24

Fun Facts:

$\sum y_i^* = \sum y_i$ when $\gamma = 0.5$, which is the model ratio estimate. That is, if all observed values are replaced by predicted values, using a model-based ratio estimate, then their sums will be equal.

When the data are grouped carefully for modeling purposes ("estimation groups" in Knaub (1999)), then otherwise nonignorable nonresponse can sometimes be made ignorable.

For many of the smallest electric power establishments, it appears that data collected frequently may often be imputed more accurately than they may be observed.

Sometimes an obviously incorrect value for γ will result in a lower estimate of RSE, so the lowest estimated RSE is not necessarily a good test for deciding which γ to use.

Only when gamma=0.5 can a lump sum of the regressor data for the unobserved cases be enough information, as shown on pages 4 and 5 of Knaub, in Proceedings of the Survey Research Methods Section, ASA 1991, found at the following URL: http://www.amstat.org/sections/ /srms/proceedings/papers/1991_133.pdf.