More Model Sampling and Analyses Applied to Electric Power Data

James R. Knaub, Jr.

U.S. Department of Energy, Energy Information Administration, EI-521, Washington, DC 20585

Key Words: Regressor, Variance estimation, Weighted average of estimates

Abstract:

This is the third in a series of papers which have dealt with adapted uses of linear model sampling and analyses for electric power industry data. Several applications are outlined, including monthly estimation of fuel costs per million BTU when neither total costs nor total BTU are known or estimated accurately and data are preliminary. Also included is the combination of two estimators, each of which uses a different regressor, for the estimation of generation expense. This latter process, unlike a single multiple regression, makes better use of the work that has been done by Royall and Cumberland, and others. A tentative conclusion is made regarding the use of this process for establishment surveys, which, along with a conclusion regarding heteroscedasticity, results in a simple, and perhaps often useful methodology for a large number of cases.

Introduction:

Along with Knaub(1990) and Knaub(1991), this paper presents a variety of applications of model sampling to electric power establishment survey data. Here, the model form chosen will be reviewed; some comments will be made regarding Knaub(1990) and Knaub(1991); study results will be shown for variance estimator performance under special, but often occurring conditions; and, finally, two new cases are considered and methodology developed.

Model form review:

The model used here is of the form

$$y_i = bx_i + x_i^{\gamma} e_{oi} .$$

From page 776, Knaub(1991), "...gamma may be obtained by fitting a homoscedastic linear regression to the result of dividing the absolute values of the residuals by x_i^{γ} to see when the slope approaches zero. The reasoning is that if each error is a multiple of a random error, then the absolute values of the errors, divided by the corresponding multipliers, should not be increasing or decreasing with increasing x."

The method of iterated reweighted least squares could have been used. However, just as the slope in the method that was used does not always equal zero for some value(s) of gamma, neither does the iterated reweighted least squares method always converge. These situations might be expected when the data appear to favor a nonlinear relationship, but this is conjecture and has not, as yet, been studied. Graphs of the 'slopes' mentioned above, plotted as functions of gamma, provide an interesting insight as to possible values of gamma that may be used, in addition to the gamma value which yields the slope closest to zero. (See Figure 3b.) However, sample sizes for establishment surveys used in this office are generally small enough that factors, such as nonsampling error, that may greatly impact upon gamma and other parameters, do not impact as greatly upon estimates of totals, ratios, and variances. This is apparently due largely to the highly skewed nature of establishment survey data. Thus, perturbances in the estimate of the 'best' value for gamma may not be very important. This is especially true if only very old and/or small sample data are all that are available for estimation of gamma.

Comments regarding previous papers in this series:

Knaub(1990) showed cases where gamma = 0.5 and gamma = 1.0 performed well. This seems to be true of a lot of electric power establishment survey data. Note also that Kirkendall (1992), in this same volume, finds gamma = 0.5 useful.

Knaub (1991) showed graphs for several cases. These cases intersect with those used in this paper.

Note that the use of n=2 was explored (Knaub (1991)) as if, other than the regressor data set, no other data were available. However, a referee from the *Journal of Official Statistics*, Statistics Sweden, made a good point when he suggested that the use of gamma = 1 when n=2 does not have any relation to the value of gamma for the underlying populations. Therefore, more information is needed; much more is preferred.

Errata from Knaub (1991): CVs for the examples on page 777 of that article are small, but actually, due to an error when programming, these estimates should have been smaller. Experience with these data, ironically, indicates that perhaps the erroneous estimates in the range shown, are superior. This is not likely to always be true, however, and the source code distributed at last year's Joint Statistical Meetings contained this error. Anyone interested may contact the author for the correction, as well as source code for a 'robust' CV estimator, and for the combination of two estimators, to be discussed later in this paper.

Variance estimator case study:

Consider regression using the zero-intercept model above, and using the variate of interest, but for a previously completed census, as the regressor. I considered monthly sample data, for which estimates had been made using a design based estimator (see Knaub (1989)), and compared results had the model

shown in this paper been used. The design-based estimator contained an element of modelling as it did make use of auxiliary data on the variates of interest for a previously completed census, but it did not assume that respondents of all 'sizes' had changed in the same proportions. When I used only the "certainty" stratum from the design based sample as a model based sample, approximately the same results were obtained using sample sizes a little more than half of that being used in accordance with Knaub (1989). Further, when using an annual census as regressor data, and a 'sample' from another annual census as if it were the only data available, a technique once suggested at the Energy Information Administration by Dean Fennell, these model and design based results were compared to the actual census results. The Royall and Cumberland (1978) V_D 'robust' variance estimator was then compared to the weighted least squares (WLS) variance estimator (used in Knaub (1990,1991)), using gamma = 0.5, and results were virtually identical. (See Table 1.) The difference appears to be greatest for the most extreme cases, i.e., where both estimators are severe underestimates of error. (See Figure 1.) This study was only done for gamma = 0.5, but results for these two estimators often appear to be nearly equal at other values of gamma also. (For a formulation of V_D at various values of gamma, I referred to Kirkendall (1991).)

Consider the case marked "MO" in Table 1. The census result was 53,887. From the sample, the 'best' estimate of gamma was 0.78, but when looking at census data (for the "y's," not just the "x's") gamma was calculated as 0.87. Since 0.78 would have been our 'best guess' for gamma here, I calculated the estimate, and cv estimates at that value of gamma, and obtained 53,567 (an improvement over the result for gamma=0.5, as can be seen from the table). I also calculated a 'robust' cv estimate (CV_D) of 0.26%, and a WLS cv estimate of 0.29%, which are still underestimates of the error, but are closer. Note that the set of graphs marked "Figure 2" pertains to this case. Figure 2a is of the x and y values from the sample, and similarly Figure 2c is for the census. Figures 2b and 2d are the corresponding plots for the absolute value of the 'slope' used to find the 'best' gamma (i.e., when the absolute value of the 'slope' is smallest). For such small sample sizes, however, it is not always the case that the sample based graphs will be such good indicators of the entire census-based graphs. See the case found in (set of graphs) Figure 3. In the case of Figure 3, unlike some other cases not shown, there is some indication of the 'best' gamma value to use. However, estimation was not improved over the use of gamma = 0.5 for that case. It seems that for establishment survey data typically found in the electric power industry, the use of gamma = 0.5 is often satisfactory.

Estimate of fuel costs per million BTU when total cost and total BTU are not well known:

Let Y represent the total fuel costs in a given category, and let X represent the total millions of BTU that could be obtained in that category. Thus, Y/X is the fuel cost per million BTU to be estimated. Let b_c be this estimate. Then,

$$b_c = \hat{Y} / \hat{X} = \left[b(\gamma)\, \hat{X}_u + Y_s \right] / \left[\hat{X}_u + X_s \right],$$

where: "s" subscripts are for sampled observations, and "u" represents the remainder of the population.

Now,

$$b_c = \left[b(\gamma)\, \hat{X}_u + (Y_s / X_s) \cdot X_s \right] / \left[\hat{X}_u + X_s \right]$$
$$= \left[b(\gamma)\, \hat{X}_u + b(1/2)\, X_s \right] / \left[\hat{X}_u + X_s \right]$$

If

$$\hat{X}_u = t X_s, \quad \textit{so (approximately)}\ X_s \textit{ is } \left[1 / (t+1) \right] X,$$
$$\textit{then} \quad b_c = \frac{t}{t+1}\, b(\gamma) + \frac{1}{t+1}\, b(1/2)$$

To apply V_D to any ratio of variates (R = N/D), the CV estimate for this ratio is

$$C\hat{V}_{D_R} = \left| V_{D_N} / \hat{Y}_N^2 + V_{D_D} / \hat{Y}_D^2 - 2\, V_{D_{ND}} / (\hat{Y}_N \cdot \hat{Y}_D) \right|^{1/2}$$

Using a model sample estimate equivalent of a double ratio estimate (see Knaub (1989) and Cochran (1977)) to take advantage of previously collected census data for both total costs and total millions of BTU, and using the relationship between costs and BTU, some indication of costs/million BTU was found to be obtained in advance of the collection of the entire census of fuel costs for a given month. The preliminary data being worked with, however, gave the author a lot of trouble, as nonsampling error not yet discovered was a primary concern. I established rules for eliminating some preliminary data, and attempted to determine the 'best' regressor data sets (e.g., data for the previous two months, or the same month in the previous year). I also experimented with gamma values. The most that can be done here appears to be to 'predict,' from preliminary data for half or more of the population, whether a large change in cost per million BTU is occurring for any given fuel. It is interesting to note, however, that the relationship between past and current data sets for heavy oil is so weak that the use of the double ratio estimate is not recommended for that fuel.

One lesson learned here is that when data, say for different types of coal, are treated together, if the individual coal types result in substantially different estimated b values, then for modelling purposes, these different types of coal should be disaggregated. That is, in model sampling, a 'separate' method can be far superior to a 'combined' method.

Table 1.

Note that many "Est. Gamma" values of "0.00" are the result of my programming logic which finds local minima. Better estimates can often be found by looking at plots, such as Figures 2b, 2d, 3b, and 3d. This table only shows results for gamma equal to 0.5. As Nancy Kirkendall also has found, here gamma = 0.5 is a consistently good choice for estimation based on modeling. From a sensitivity analysis that I have done, one reason for this appears to be that for small sample sizes, such as found in this table, the addition or deletion of a single observation can change the estimate of gamma based on the sample by a considerable margin. It sometimes makes the difference as to whether a gamma value that will satisfy the mathematical model even exists. (In plots that follow, this is indicated when 'ABSYRL' does not equal zero for any value of gamma.) In some cases, the estimate (here, for total sales) does not vary much with gamma, but in some cases it does. There appears to be some indication, however, that when the estimator does vary with gamma appreciably, the estimated gamma value can often be used to improve the accuracy of the estimated total, but the cv estimate is likely to deteriorate. This seems not only to be true when using the same variate for a different time period as the auxiliary or regressor variate, but also when using another variate entirely. (See my paper on applications to generation and generation expense in the 1990 ASA proceedings for the Survey Research Methods section. CV_L is shown there, but CV_D is comparable as in the table below.) CV_D is the robust estimator shown here and associated with Royall and Cumberland's (1978) V_D. CV_L is associated with their V_L, and is shown under "WLS" (weighted least squares) here.

Note that I have dropped cases where n=2 since gamma can not be estimated. (It always appears to be equal to 1, and, as a JOS referee noted, this does not estimate the 'true' gamma value.) I also dropped observations from the universe where the value of the regressor was 0, as 'new' members of the universe would best be treated separately.

					ROBUST			WLS						
ST	GAM	Est. Total = SHAT	Total = S	SHAT-S	cvHAT(%)	sigmaHAT	zHAT	cvHAT(%)	sigmaHAT	zHAT	Est. Gamma	n-Sample	N-Universe	"Coverage"
AK	0.5	4223	4244	-20.9	0.58	24.5	-0.85	0.64	26.8	-0.78	0.00	9	65	0.8319
AL	0.5	59424	59926	-502.0	0.60	357.6	-1.40	0.66	391.8	-1.28	0.36	6	62	0.8128
AR	0.5	26604	27365	-761.7	0.40	107.3	-7.10	0.28	74.7	-10.19	0.31	5	38	0.7425
AZ	0.5	41137	41376	-238.3	0.47	193.1	-1.23	1.06	436.2	-0.55	0.00	4	40	0.8808
CA	0.5	213170	211093	2077.2	0.24	518.3	4.01	0.38	811.4	2.56	0.00	5	50	0.8616
CO	0.5	30731	30795	-63.6	0.10	31.0	-2.05	0.22	67.9	-0.94	0.00	5	58	0.7584
FL	0.5	142986	143535	-548.9	0.20	284.9	-1.93	0.17	242.7	-2.26	0.31	4	55	0.7778
GA	0.5	80264	80440	-176.6	0.77	617.5	-0.29	0.37	294.7	-0.60	0.45	9	98	0.8097
IA	0.5	29514	29437	77.1	0.92	270.3	0.29	0.80	236.7	0.33	1.86	9	200	0.7882
ID	0.5	18014	18003	10.5	0.22	40.3	0.26	0.28	50.6	0.21	0.00	5	34	0.9138
IN	0.5	74429	73977	452.2	0.82	607.7	0.74	0.75	558.2	0.81	0.00	4	122	0.7298
KS	0.5	27213	27115	97.9	0.47	128.8	0.76	0.65	178.1	0.55	0.00	4	162	0.7433
KY	0.5	62330	61097	1232.8	5.03	3133.3	0.39	5.06	3153.2	0.39	0.87	8	64	0.7823
LA	0.5	63770	63826	-55.6	0.22	140.2	-0.40	0.25	156.3	-0.36	0.06	5	42	0.8716
MA	0.5	45404	45408	-4.5	0.32	145.2	-0.03	0.30	134.2	-0.03	0.46	4	50	0.7742
MN	0.5	47337	47167	169.9	0.66	312.7	0.54	0.43	201.9	0.84	0.75	8	183	0.7865
MO	0.5	53478	53887	-409.3	0.33	174.2	-2.35	0.36	191.8	-2.13	0.78	7	141	0.7627
MS	0.5	32368	31831	537.1	2.73	883.5	0.61	2.92	944.2	0.57	0.06	9	50	0.7295
MT	0.5	13209	13111	97.8	0.56	73.9	1.32	0.52	68.3	1.43	0.35	5	39	0.8641
NC	0.5	89691	89925	-233.7	0.17	156.0	-1.50	0.18	157.9	-1.48	0.45	7	110	0.8157
ND	0.5	7110	7014	96.0	0.63	45.0	2.14	0.78	55.5	1.73	0.00	8	42	0.7546
NE	0.5	17971	17868	102.7	0.64	114.4	0.90	1.09	196.1	0.52	0.02	13	166	0.7690
NH	0.5	8997	8980	17.0	0.26	23.6	0.72	0.21	18.5	0.92	0.70	3	13	0.8158
NM	0.5	13632	13821	-188.5	0.62	84.5	-2.23	0.76	103.3	-1.82	0.05	5	33	0.7525
NY	0.5	129516	129324	191.9	0.15	195.9	0.98	0.13	164.2	1.17	0.91	8	63	0.9722
OH	0.5	141798	141874	-76.2	1.72	2439.4	-0.03	2.17	3074.9	-0.02	0.60	8	121	0.9212
OR	0.5	43043	42977	66.0	1.07	460.2	0.14	1.36	584.6	0.11	0.00	4	40	0.7848
PA	0.5	114919	114751	168.3	0.40	454.9	0.37	0.34	390.8	0.43	1.33	4	58	0.7504
SC	0.5	55247	55652	-404.9	0.28	157.2	-2.58	0.30	165.5	-2.45	0.42	5	47	0.8363
SD	0.5	6390	6334	56.2	1.61	102.8	0.55	2.08	132.8	0.42	0.00	6	77	0.5837
TN	0.5	76854	77069	-215.7	0.42	325.7	-0.66	0.34	264.2	-0.82	0.81	25	93	0.7766
TX	0.5	237064	237335	-271.1	0.29	694.3	-0.39	0.30	719.7	-0.38	0.27	5	162	0.7373
VA	0.5	72549	72696	-147.6	0.17	123.3	-1.20	0.24	172.7	-0.85	0.00	5	34	0.9096
VT	0.5	4714	4716	-2.2	0.67	31.4	-0.07	0.59	28.0	-0.08	0.50	4	25	0.8524
WA	0.5	91074	91046	28.0	1.17	1067.7	0.03	1.38	1260.0	0.02	0.44	9	67	0.8116
WI	0.5	48979	49198	-219.4	0.32	158.1	-1.39	0.23	110.8	-1.98	0.00	6	123	0.8324
WY	0.5	12025	11769	256.0	0.62	74.5	3.44	0.46	55.7	4.59	0.71	3	36	0.8177

Note that one or two 'odd' observations can change results substantially. For example, in the case of California, if we eliminate a single observation from the universe which has an auxiliary value, in the units above (millions of kilowatthours), of 1550.4 when the actual value in the 'current' data set is 0, and this data point was not in the sample, then the following is obtained:

CA	0.5	211523	211093	429.9	0.24	502.5	0.86	0.37	786.7	0.55	0.00	5	49	0.8616

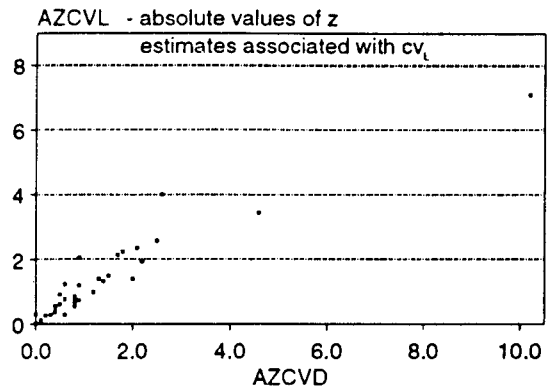
1) Gamma=0.5 still looks good for at least most of these cases.
2) V_D, one of Royall and Cumberland's 'robust' variance estimators, seems to work reasonably well, although variance estimation is not as impressive as the estimate of total (or a ratio), and may possibly be improved upon.
3) 'New' utilities, i.e., those having a zero or no data for the auxiliary variate, i.e., those who had no retail sales reported in the year used as the auxiliary prior to the period of current interest, should be required to respond and their data treated separately as a 'certainty' stratum.

Combination of estimators with different regressors:

There is a lot of literature available that applies single regressor models to survey sampling, but this is not true for multiple regression. (Note as an exception, however, Little (1988). Also, relatedly, see Rao and Mudholkar (1967).) To directly apply work such as that done by Royall (1970), or Royall and Cumberland (1978,1981), when more than one regressor may be available could be difficult. However, I noted that I did have two possible regressors for estimating generation expense. These are nameplate capacity, and net generation (with perhaps a small arbitrary constant added to avoid zero and negative net generation values). It seemed reasonable that a better estimate might be to 'average' the estimates resulting from these two single regressor models, and

Note that a "~" preceding a variate indicates an estimate, just as if a "hat" were placed above it.

Figure 1.
The graph below compares absolute values of z estimates associated with CV_D (the robust estimator used here) to those associated with CV_L. It appears that there is virtually no difference here except perhaps for cases with extreme values for z estimates. (Note $\sim z = (\sim S - S)/\sim\sigma$, where in this case, "S" stands for "total sales.")



absolute values of z estimates associated with CV_D

that the final estimate would have to have a cv estimate no greater than the largest of the two individual estimates. Further, this would appear to provide some limitation on bias. When I brought this to Nancy Kirkendall's attention, she suggested using a weighted average determined by the single regressor variance estimates and referred me to Granger and Newbold (1977), where a method for combining forecasts attributable to Bates and Granger (1969) was discussed. When applied to generation expense data, it appeared, as shown in Table 2, that just as using gamma = 0.5 for these small sample establishment surveys often appears to be practical, so to does it seem prudent to assume no correlation between errors for the two estimators even when rho is estimated to be far from zero. Perhaps the reason lies partly in the variance of the estimate of rho, or in nonsampling error, or bias which is ignored by Bates and Granger, or some other factor or combination of factors. Perhaps further case studies will reveal that rho should not be set equal to zero, but for now, for small sample establishment surveys, I would recommend it.

Note that another possible use of the Bates-Granger method that could not be replaced by multiple regression, like the original use, would be to combine estimates from completely different model forms, or a model based estimate and a design based estimate, etc.

The author thanks Prof.Poduri S.R.S. Rao for notes and a model covariance estimator used in generating Table 2.

Conclusions:

Model sampling is proving to be very useful for our establishment surveys. Test results show this. As I have not been using 'balanced' sampling, but instead have recommended use of those possible respondents with the largest regressor values (when available), the worst scenario seems to be one in which the relationship between the variant of interest and the regressor is very different for the 'smaller' members of the population than for the 'larger' ones. Fortunately, for highly skewed establishment surveys such as these, the threat is diluted.

The advantage in using a 'robust' variance estimator seems to have been exaggerated in the literature, at least as applied to the data analyzed here. However, it is still recommended that one be used, and further recommended that, barring good information to the contrary, gamma = 0.5 be used. (When gamma=0.5 is used, the problem solved by the section "Incompletely Specified Auxiliary Data" in Knaub (1991), becomes moot.) Also, for cases where more than one regressor is available, I recommend consecutive applications of the Bates-Granger method, with rho=0, unless another value of rho is proven to be better, or this method is shown to be inferior to that of Little (1988), or some other method, in some way important to your application. (See notes at the bottom of Table 2.)

References:

Bates, J.M., and C.W.J. Granger (1969), "*The Combination of Forecasts,*" Oper. Res. Q 20, 451-468.

Cochran, W.G. (1977), *Sampling Techniques*, 3rd ed. John Wiley & Sons.

Granger, C.W.J., and P. Newbold (1977), Forecasting Economic Time Series, Academic Press.

Kirkendall, N.J. (1991). Presentation to ASA Committee on Energy Statistics, Oct. 1991.

Kirkendall, N.J. (1992), Paper published in this volume; *Proceedings of the Section on Survey Research Methods*, Amer. Stat. Assoc.

Knaub, J.R., Jr. (1989), "*Ratio Estimation and Optimum Stratification in Electric Power Surveys,*" Proceedings of the Section on Survey Research Methods, Amer. Stat. Assoc., 848-853.

Knaub, J.R., Jr. (1990), "Some Theoretical and Applied Investigations of Model and Unequal Probability Sampling For Electric Power Generation and Cost," Proceedings of the Section on Survey Research Methods, Amer. Stat. Assoc., 748-753.

Knaub, J.R., Jr. (1991), "Some Applications of Model Sampling to Electric Power Data," Proceedings of the Section on Survey Research Methods, Amer. Stat. Assoc., 773-778.

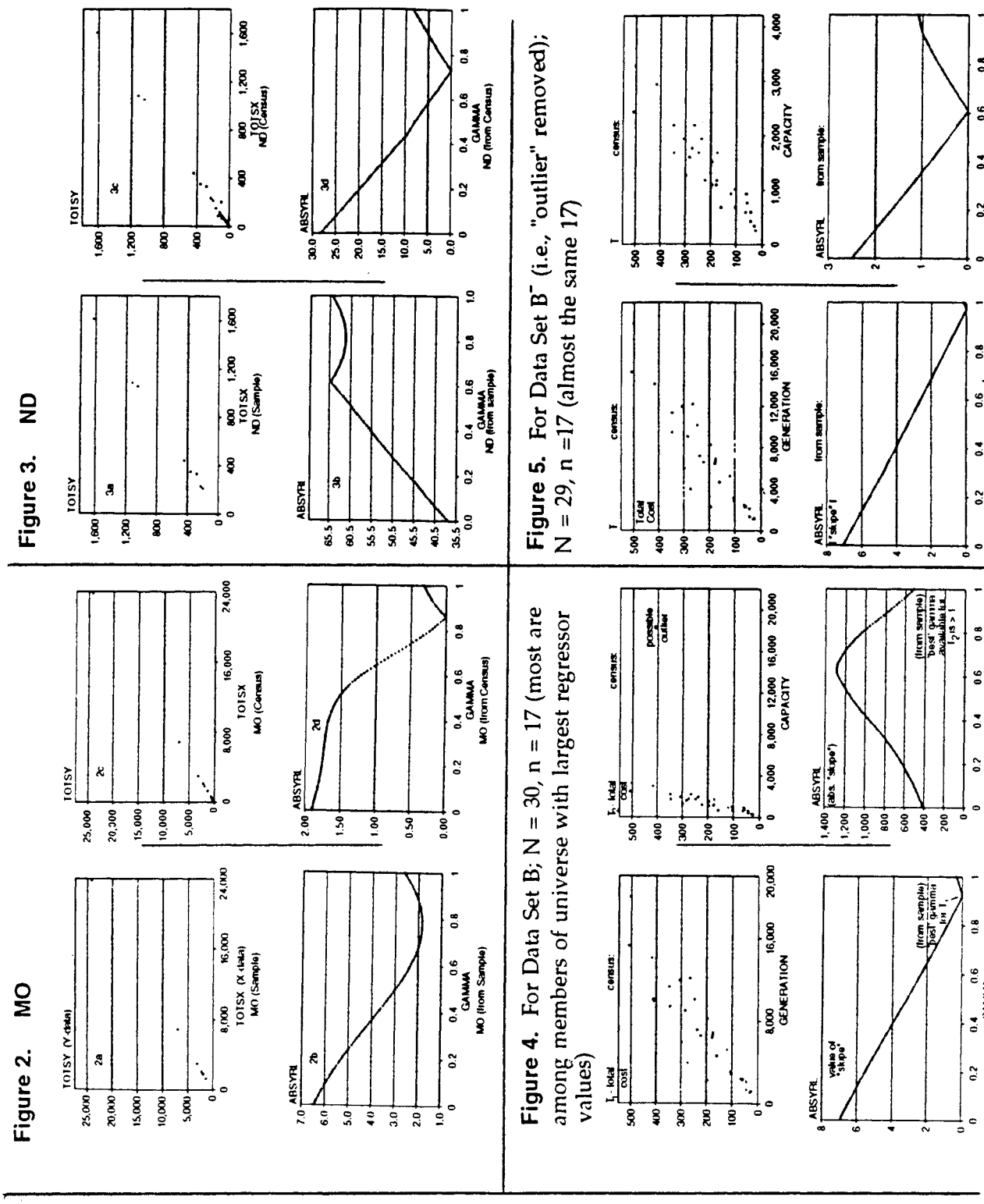
Little, R.J.A. (1988), "Some Statistical Analysis Issues at the World Fertility Survey," *The American Statistician*, 42, 31-36.

Rao, Poduri, S.R.S., and G.S. Mudholkar (1967), "Generalized Multivariate Estimations for the Mean of Finite Populations," JASA, 62, 1008-1012.

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

Royall, R.M., and W.G. Cumberland (1978), "Variance Estimation in Finite Population Sampling," JASA, 73, 351-358.

Royall, R.M., and W.G. Cumberland (1981), "An Empirical Study of the Ratio Estimator and Estimators of Its Variance," JASA, 76, 66-88.



The model used here is of the form

$$y_i = bx_i + x_i^{\gamma} e_{o_i}.$$

From page 776, Knaub (1991), "...gamma may be obtained by fitting a homoscedastic linear regression to the result of dividing the absolute values of the residuals by x_i^{γ} to see when the slope approaches zero. The reasoning is that if each error is a multiple of a random error, then the absolute values of the errors, divided by the corresponding multipliers, should not be increasing or decreasing with increasing x."

Table 2.

Note that a "~" preceding a variate indicates an estimate, just as if a "hat" were placed above it.
(V_L) indicates variance (and covariance) estimates are based on WLS
(V_D) indicates variance estimates are of the 'robust' form so-named by Royall and Cumberland
$\sim T = k\sim T_1 + (1-k)\sim T_2$ = est. of T, where k is Bates-Granger factor for combining two estimators, and $\sim T_1$ and $\sim T_2$ are each estimators of T using separate single regressors;

$k = (\sigma_2^2 - \rho\sigma_1\sigma_2) / (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$

$\sim\gamma_1$ and $\sim\gamma_2$ are estimates of γ_1 and γ_2, the γ model parameters for estimators $\sim T_1$ and $\sim T_2$ (Note: γ_1 and γ_2 are 'unknown')
$\sim\rho$ is estimator of correlation coefficient between errors for two estimators of T
$\sigma^2 = (\sigma_1^2\sigma_2^2 [1-\rho^2]) / (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)$; $\sigma^2 \leq \min(\sigma_1^2, \sigma_2^2)$

DATA SET A:

N=32, n=17, $\sim\gamma_1$=2.80, γ_1=0.99, $\sim\gamma_2$=1.53, γ_2=0.77, and $\sim\rho(\gamma)$ is such that we have the following table for the WLS case:

γ_1,γ_2	0,0	0.5,0.5	0.8,0.8	1.0,1.0	1.0,0.8
$\sim\rho(\gamma_1,\gamma_2)$	0.867	0.822	0.782	0.753	0.839
$\sim SE_1/\sim SE_2$	0.884	0.793	0.728	0.685	0.634

Finally, T=5113 is 'unknown' total for variable of interest

EST. 'TYPE'	(γ_1,γ_2)	Set ρ Equal To:	$\sim T_1$	$\sim SE_1$	$\sim CV_1$	$\sim T_2$	$\sim SE_2$	$\sim CV_2$	$\sim T$	$\sim SE$	$\sim CV$
(V_L)	(0.5,0.5)	0.822	5046	144	2.9	5239	182	3.5	5032	144	2.9
(V_L)	(0.5,0.5)	0	5046	144	2.9	5239	182	3.5	5121	113	2.2
(V_D)	(0.5,0.5)	0	5046	166	3.3	5239	201	3.8	5124	128	2.5
(V_L)	(1.0,1.0)	0.753	5036	99	2.0	5250	144	2.7	5014	99	2.0
(V_L)	(1.0,1.0)	0	5036	99	2.0	5250	144	2.7	5104	82	1.6
(V_D)	(1.0,1.0)	0	5036	112	2.2	5250	154	2.9	5110	91	1.8
(V_L)	(1.0,0.8)	0	5036	99	2.0	5246	156	3.0	5096	83	1.6
(V_D)	(1.0,0.8)	0	5036	112	2.2	5246	170	3.2	5100	93	1.8

DATA SET B:

N=30, n=17, $\sim\gamma_1$=0.93, $\sim\gamma_2$>1, T = 'unknown' total = 6393

EST. 'TYPE'	(γ_1,γ_2)	Set ρ Equal To	$\sim T_1$	$\sim SE_1$	$\sim CV_1$	$\sim T_2$	$\sim SE_2$	$\sim CV_2$	$\sim T$	$\sim SE$	$\sim CV$
(V_L)	(0.5,0.5)	-0.229	6305	187	3.0	5190	607	11.7	6156	164	2.7
(V_L)	(0.5,0.5)	0	6305	187	3.0	5190	607	11.7	6208	179	2.9
(V_D)	(0.5,0.5)	0	6305	227	3.6	5190	1613	31.1	6284	224	3.6
(V_L)	(1.0,0.8)	0	6184	215	3.5	5889	473	8.0	6133	196	3.2
(V_D)	(1.0,0.8)	0	6184	216	3.5	5889	1046	17.8	6172	212	3.4

Data Set B*, Data Set B with one possible 'OUTLIER' REMOVED:
N=29, n=17, $\sim\gamma_1$=0.98, $\sim\gamma_2$=0.61, T = 'unknown' total = 5986

EST. 'TYPE'	(γ_1,γ_2)	Set ρ Equal To	$\sim T_1$	$\sim SE1$	$\sim CV_1$	$\sim T_2$	$\sim SE_2$	$\sim CV_2$	$\sim T$	$\sim SE$	$\sim CV$
(V_L)	(0.5,0.5)	0.679	5754	151	2.6	5886	284	4.8	5736	148	2.6
(V_L)	(0.5,0.5)	0	5754	151	2.6	5886	284	4.8	5783	133	2.3
(V_D)	(0.5,0.5)	0	5754	184	3.2	5886	306	5.2	5789	158	2.7
(V_L)	(1.0,0.8)	0	5673	180	3.2	5767	317	5.5	5696	157	2.8
(V_D)	(1.0,0.8)	0	5673	178	3.1	5767	313	5.4	5696	155	2.7

NOTES: $\sigma^2(\rho)$ generally tends to zero as ρ approaches -1 or 1 (see Granger and Newbold (1977), pp. 270 and 271) - and, letting σ_L = MIN (σ_1,σ_2) and σ_u = MAX (σ_1,σ_2), $\sigma^2(\rho)$ is maximized at $\rho = \sigma_L/\sigma_u$. (See Granger and Newbold (1977), p. 270, and note also that $\partial\sigma^2(\rho)/\partial\rho = 0$ yields $\rho = \sigma_L/\sigma_u$.) Therefore, ρ=0 is not a maximizing or minimizing value for $\sigma^2(\rho)$. Also, note that if $\rho > \sigma_L/\sigma_U$, T will be outside of the (T_1,T_2) range(!!!).

Therefore, ρ = 0 is not an unreasonable value to use, especially if bias possibilities are considered and one may therefore want to use an estimator of T closer to $(\sim T_1 + \sim T_2)/2$ than is obtained in may cases when ρ is estimated.

Like $\sim\gamma$, perhaps $\sim\rho$ is too easily disturbed by an outlier, and for establishment surveys, ρ=0, like γ=0.5, may often be useful.

For m regressors, if all covariances are set equal to zero, $\sim T = k_1\sim T_1 + k_2\sim T_2 + k_3\sim T_3 + \ldots + k_i\sim T_i + \ldots k_m\sim T_m$

where

$$k_i = \left(\prod_{h\neq i} \sigma_h^2\right) / \left(\sum_{j=1}^{m} \prod_{h\neq j} \sigma_h^2\right)$$

$$\sigma^2 = \left(\prod_{h=1}^{m} \sigma_h^2\right) / \left(\sum_{j=1}^{m} \prod_{h\neq j} \sigma_h^2\right)$$