SOME APPLICATIONS OF MODEL SAMPLING TO ELECTRIC POWER DATA

James R. Knaub, Jr.
US Dept. of Energy, Energy Information Administration, EI-541, Washington DC 20585

Key words:
Sample size determination, Design-based sampling, Incompletely specified auxiliary (i.e., regressor) data

ABSTRACT:

The Energy Information Administration (EIA) collects a variety of data from the energy industry. The use of censuses and design-based sampling is common in this agency. This paper, however, presents the results of some research in model-based sampling for differing purposes as pertains to electric power data. Some nonstandard methodology is explored. The restrictive assumptions of Richard Royall's 1970 Biometrika paper are used under circumstances where the adequacy of this class of models is somewhat quantifiable, and/or other choices are not readily available. Model failure is a particular concern here. However, under these models, data collection may be more convenient, which may mean smaller nonsampling error, and less inconvenience to the utilities that can least afford to participate in a survey.

One method developed may be of particularly wide application. It comprehensively examines a class of incompletely specified auxiliary data. FORTRAN source code to accomplish this is available on request.

INTRODUCTION:

Let $y_i = bx_i + x_i^{\gamma} e_{0i}$,and

$$b(\gamma) = (\sum_{g=1}^{n} x_g^{1-2\gamma} y_g) / (\sum_{g=1}^{n} x_g^{2-2\gamma}).$$

In Knaub (1990), this class of models is compared, for gamma equal to 0, 0.5, and 1, to unequal probability sampling when estimating generation expense. Comparison to a census is also made when estimating kilowatthours generated. Most model results for figures shown below have been compared to results from censuses, unequal probability sampling, or stratified random sampling using the ratio estimate. The graphs depict a variety of circumstances, yet this form of model sampling/estimation, generally using only the respondents with the largest regressor data, demonstrated widespread usefulness. Linear transformations did not seem to be needed. Even the addition of a constant appeared to possibly overspecify the model. Estimates for gamma in the model shown above were typically between 0.5 and 1. Some attempts were made to split the data of Figure B into groups as if other, currently unknown variables could be used to distinguish between them, but the simplest application of Royall (1970) seemed to work best. Also, although these models seem to estimate totals well in a number of cases, estimates of mean square error did not appear to be as good. Underestimation may be expected when there is serious model failure. If this is a problem, help may possibly be obtained, for example, from Royall and Cumberland (1978), and/or Herbert and Kott (1988). However, the method developed below for estimating gamma also indicates whether the weighted least-squares variance estimator, which is used here, is adequate. (Note that it is assumed that relative error, i.e., coefficient of variation (cv), is identical to the square root of mean square error, since model-unbiasedness is assumed. Nonsampling error is ignored here, although it could be a problem.)

There are several specialized investigations in this paper also. First, the relationship between unequal probability sampling and model sampling of the type found in Royall (1970), which was discussed in Knaub (1990), was further investigated. The idea is to start with stratified unequal probability sampling, and apply both estimation methods to the resulting data. Based on a preliminary sample performed in this manner, if further sampling is required, and preliminary results under either method are nearly identical, sample selection may proceed under the model. Sample size requirements are discussed, and sample selection, considering stratification, is also addressed.

A second specialized investigation is for the case where a few utilities have extremely large responses in relation to the other utilities and where the estimation of the value to be used for gamma is obviously distorted. An expediant solution is provided. A third investigation is for the case where auxiliary data are available for each observed element, but only the sum of the values for the auxiliary variate is available for the unobserved elements. The estimate of totals is not affected, but in all cases except for gamma equal to 1/2, the cv estimate is affected.

Note that $\hat{cv}^2(u)$, $\hat{cv}^2(w)$, and $\hat{cv}(u,w)$ may still be calculated for the ratio u/w case.

SAMPLE SELECTION BY STRATUM FOR UNEQUAL PROBABILITY SAMPLING AND MODEL SAMPLING AS IN ROYALL (1970):

Unequal probability sampling -

Consider generation and generation expense as in Knaub (1990). For a given cost (fuel, operations and maintenance, capital costs, or total), let $\hat{Y}_{Oh}$ be the total of such costs estimated for all plants in stratum h.

$\hat{Y}_{Oh} = \sum_{g=1}^{n_h} \frac{y_{hg}}{\pi_{hg}}$, where π_{hg} is calculated from the Van Beeck and Vermetten Method. (See Konijn (1973) for a description of this method.) If G represents total generation, then the cost per kilowatthour is estimated by $\hat{Y}_D/G = \sum_{h=1}^{L} \hat{Y}_{Dh}/G$. G is currently obtained by a census and will be considered a constant here. Therefore, the estimated cv is the same for $\hat{Y}_D$ as for $\hat{Y}_D/G$.

$$\hat{cv}(\hat{Y}_D) = \left[\sum_{h=1}^{L} \hat{cv}^2(\hat{Y}_{Dh})\hat{Y}_{Dh}^{\,2}\right]^{1/2} / \hat{Y}_D$$

$$\hat{cv}(\hat{Y}_{Dh}) \cong \frac{s_h}{\hat{Y}_{Dh}} \left[\frac{N_h - n_h}{n_h(N_h - 1)}\right]^{1/2} \quad \text{, where}$$

N_h is the size of the population in stratum h,
n_h is the size of the preliminary sample in stratum h,

$$\text{and } s_h^2 = \sum_{g=1}^{n_h} \left[\frac{y_{hg}}{\pi_{hg/n_h}} - \hat{Y}_{Dh}\right] / \left[n_h(n_h - 1)\right].$$

$$\frac{d\hat{cv}(\hat{Y}_{Dh})}{dn_h} \cong \frac{-s_h N_h}{2\hat{Y}_{Dh}\left[n_h^3 (N_h - n_h)(N_h - 1\right]^{1/2}}.$$

Now, using unequal probability sampling, the appropriate estimates from the preliminary sample are

$\hat{Y}_D/G$, and $\hat{cv}(\hat{Y}_D) = \hat{cv}(\hat{Y}_D/G)$.

Comparing $\left|\frac{d\hat{cv}(\hat{Y}_{Dh})}{dn_h}\right| \hat{Y}_{Dh}$ for each h, we may add to the sample size for the stratum with the largest such value, recalculate a revised $\hat{cv}(\hat{Y}_D)$, and repeat until $\hat{cv}(\hat{Y}_D)$ is below a preselected value.

Note that if $N_h >> n_h$ for all h, comparisons may be based on $s_h/n_h^{3/2}$.

Model sampling -

One could find the value of gamma that most nearly makes $\hat{Y}(\gamma)$ equal to $\hat{Y}_D$, but in this paper, gamma is always set to the value that the data most nearly indicate based on the class of models found in Royall (1970). The iterated reweighted least squares method could have been used, but the computer programming done for this study found the gamma value such that the slope of the ordinary least squares regression line through the absolute value of the residuals divided by x_i^{γ}, approached 0. Note that when 0 is achieved, as opposed to always having a negative, or always a positive result, the model appears most reasonable. This may be the only time that the variance estimate shown here should be used.

A number of sets of results were displayed for iteratively increasing gamma values. The sensitivity of the model to differences in gamma was thus, easily apparent. (Relevant FORTRAN code is available from the author.)

Now,

$$\hat{Y}_{\gamma_h} = \left[b(\gamma_h)\right]X_{uh} + Y_{sh}.$$

$$b(\gamma_h) = \left(\sum_{g=1}^{m_h} x_{hg}^{1-2\gamma_h} y_{hg}\right) / \left(\sum_{g=1}^{m_h} x_{hg}^{2-2\gamma_h}\right),$$

the m_h values are sample sizes,
X_u = total of the auxiliary variate values for the unobserved portion of the population, and
Y_s = total of the observed y values ($Y_s = \sum_s y_i$).

$$\hat{Y}_\gamma = \sum_h \hat{Y}_{\gamma_h} = \hat{Y}_\gamma.$$

$$\text{Let} \quad \hat{M}(\gamma_h) = \sum^{N_h} x_{hg}^{2\gamma_h} - \sum^{m_h} x_{hg}^{2\gamma_h} + X_u^{2} / \left(\sum^{m_h} x_{hg}^{2-2\gamma_h}\right)$$

$$\hat{\sigma}^2(\gamma_h) = \sum^{m_h} (1/x_{hg}^{2\gamma_h})(y_{hg} - b(\gamma_h)x_{hg})^2/(m_h - 1)$$

$$\hat{cv}(\hat{Y}_{\gamma_h}) = \left[\hat{M}(\gamma_h)\right]^{1/2} \hat{\sigma}(\gamma_h)/\hat{Y}_{\gamma_h}$$

$$\text{Therefore,} \quad \hat{cv}(\hat{Y}_\gamma) = \left[\sum_h \hat{cv}^2(\hat{Y}_{\gamma_h})\,\hat{Y}_{\gamma_h}^2\right]^{1/2} / \hat{Y}_\gamma.$$

Thus, $\hat{cv}(\hat{Y}_\gamma)$ may be compared to $\hat{cv}(\hat{Y}_D)$.

If this model sampling methodology is to be used following a preliminary sample, as discussed earlier, select only the units in each stratum with the largest x values which have not already been selected. As an estimate of the required sample size under model-based sampling, one could use the following approximation (exact when $\gamma = 0.5$):

$$m_h = \left[N_h\,(\bar{X}_h/a_h)^{\gamma_h}\,\omega_{\gamma_h}\right]^2 \Big/ \left[c_{\gamma_h}^2 + N_h\,\bar{X}_h^{2\gamma_h}\,\omega_{\gamma_h}^2\right],$$

where a_h should be chosen so that

$$\bar{x}_h = a_h\,\bar{X}_h \quad .$$

$$\omega_{\gamma_h} = \hat{\sigma}(\gamma_h)/\hat{Y}_{\gamma_h}$$

and c_{γ_h} is a value for $\hat{cv}(\hat{Y}_{\gamma_h})$ such that

if c is the desired overall cv,

$$c = \left[\textstyle\sum_n c_{\gamma_h}^2\,\hat{Y}_{\gamma_h}^2\right]^{1/2} / \hat{Y}_\gamma \,.$$

Now, redefine m_h as a preliminary sample size, solve for $\hat{cv}(\hat{Y}\gamma_h)$, and take the absolute value of the derivative with respect to sample size. Multiply by $\hat{Y}\gamma_h$.

Recalculating c_{γ_h} as the value of $\hat{cv}(\hat{Y}\gamma_h)$ for each new m_h, the m_h value will be increased, according to which h yields the

largest value for $\hat{Y}\gamma_h \dfrac{1}{2c_{\gamma_h}}\left[N_h\,(\bar{X}_h/a_h)^{\gamma_h}\,\omega_{\gamma_h}\right]^2 / m_h^2$.

c_{γ_h} will then be recalculated for each h as $c_{\gamma_h} = \left[\hat{M}'(\gamma_h)\right]^{1/2}\omega_{\gamma_h}$ where $\hat{M}'(\gamma_h)$ is $\hat{M}(\gamma_h)$ recalculated for the enlarged sample. $\hat{M}'(\gamma_h)$

would then change each time we add to m_h

Note that there may sometimes be other considerations, such as tolerance limits, when determining sample size requirements, but that here we are concentrating on the finite nature of the population.

Note that instead of determining the largest $\left|\dfrac{d\hat{cv}(\hat{Y}_{Dh})}{dn_h}\right| \hat{Y}_{Dh}$ under the design, or similarly for the model, to identify the stratum which should contain the next observation, we could simply calculate an estimate of the cv for each situation to see how to develope the goal low cv estimate with the fewest total number of observations. This would, however, require more programming and computer time.

Also note the following for the model-based sampling approach:

Let $N_h\,(\bar{X}_h/a_h)^{\gamma_h}\,\omega_{\gamma_h} = H_h$, then the quantity of interest for each stratum (h=1,L), is

$$\hat{Y}_{\gamma_h}\left|\frac{d\,\hat{cv}(\hat{Y}\gamma_h)}{d\,m_h}\right| = \hat{Y}_{\gamma_h}\,H_h^2/(\,2c_{\gamma_h}m_h^2).$$ Then also,

$$\hat{cv}(\hat{Y}\gamma_h) = \left[\frac{H_h^2}{m_h} - N_h\,\bar{X}_h^{2\gamma_h}\,\omega_{\gamma_h}^2\right]^{1/2},$$ so

$$\hat{Y}_{\gamma_h}\left|\frac{d\,\hat{cv}(\hat{Y}\gamma_h)}{d\,m_h}\right| = \hat{Y}_{\gamma_h}H_h^2\Big/\left[2m_h^2\left(\frac{H_h^2}{m_h} - N_h\,\bar{X}_h^{2\gamma_h}\,\omega_{\gamma_h}^2\right)^{1/2}\right].$$

Therefore, if $N_h >> a_h^{2\gamma_h}\,m_h$, then

$$\hat{Y}_{\gamma_h}\left|\frac{d\,\hat{cv}(\hat{Y}\gamma_h)}{d\,m_h}\right| \approx \hat{Y}_{\gamma_h}\,H_h/2m_h^{3/2}.$$

Note that $N_h >> a_h^{2\gamma_h}\,m_h$ does not need to hold true for this process to be viable, as long as $a_h^{2\gamma_h}\,m_h$ is proportional to N_h.

MODEL SAMPLING WHEN A FEW, VERY LARGE OBSERVATIONS DOMINATE, AND DISTORT GAMMA:

Sometimes, as in Figures C and D, a few observations contain a very large portion of the total to be estimated. When we try to estimate gamma, instead of obtaining something in the usual range of values (for these data, 0.5 to 1.0, perhaps 0.8 to 0.9), an unusual estimate may be obtained due to the special nature of these few observations. (This was the case with Figure D, but not C.) A case such as that found in Figure E may yield an estimate of gamma, perhaps partly due to nonsampling error, that is negative, which may best be treated as zero, or perhaps a different methodology should be used. This is not something one wants to discover near a data publication deadline if gamma values have not yet been estimated, unless it is being used to help identify nonsampling error. If we treat data with the largest regressor values as belonging to a certainty stratum, then the remaining portion of the population may be estimated in almost any way, and although errors may be relatively large for this remainder portion, cvs which include the certainty portion may be very small. One possibility is to observe as many of the largest establishments as practical, and from among that set of observations, consider all but the two of them with the smallest regressor values as certainties. The remaining two observations may be applied to the modeling methodology used in this paper, as presented in (Royall (1970)). This may provide nominal estimates without the need for a last minute review that could delay publication. Test data could be used to demonstrate this for any given situation.

If two observations are used for the sampled portion, gamma will likely be approximately 1, which may be nearly correct in the cases of Figures C and D. The derivation below shows why the estimate of gamma is likely to be near unity. Also, note that if data for more than one variate are being collected, we may be flexible when noting which observations are to be treated as certainties, and which two are to be used to estimate the remainder. That is, data from a given establishment may be treated as part of the certainty stratum for one variate, and part of the sampled stratum for another variate.

Derivation -

The model sampling in this paper is based on the following:

$$y_i = bx_i + x_i^{\gamma} e_{0i} \text{ .}$$

As indicated earlier, 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.

Now, when we only have two observations, we estimate gamma from $|e_{01}| = |e_{02}|$. Thus, one possibility is $e_{01} = e_{02}$, and the other is $e_{01} = -e_{02}$. In the first case, after using some algebra to 1) solve for b, 2) set the result equal to the formula for b which provides minimum error under the model, and 3) collect some terms, it can be seen that if $y_1/x_1 = y_2/x_2$, then $\gamma = 1$ is a practical solution. In the second case, which is the situation to be expected in practice, $\gamma = 1$ is, without restriction on y or x, seen to be an obvious solution. It is worth noting, however, that the author did not first notice this as a result of considering the algebra, but instead found that in a practical example, gamma was approximately unity.

(Also, note that the problem of nonlinearity in these data may best be solved by treating several elements as certainties. This would have lead to slightly different results than those shown below, associated with Figure G. For that case, note also that, as can be seen below, a transformation is not an option.)

INCOMPLETELY SPECIFIED AUXILIARY DATA:

Suppose that we have auxiliary data for each observed element, with a suitably high correlation between these data and the data of interest, but suppose that for the unobserved elements we only know the total for the auxiliary data, and the number, or approximate number of elements in the universe. As a practical matter, this is the case in an instance encountered by the author inwhich data of interest and auxiliary data were becoming available for a number of facilities on one survey, and the universe total for the auxiliary data is available from another survey. This does not affect the estimate of the total for the data of interest, but does affect cv estimates, unless $\gamma = 0.5$, i.e., unless the ratio estimate is apropos. In all cases other than $\gamma = 0.5$, only upper and lower bounds on cv estimates are obtainable.

Derivations -

Let $\hat{CV}_1^2 \equiv \hat{CV}^2$ for the usual case where we have knowledge of the x_i values for all N elements, and let $\hat{CV}_2^2 \equiv \hat{CV}^2$ for the case where X_u (the sum of the x_i for the elements not observed for y_i values) is known, but not the individual x_i that constitute X_u ,i.e., when i=n+1,N. Further, let $\hat{CV}_2^2$ be calculated using one artificial data point for all unobserved elements. (Note the poor notation on page 749 in Knaub (1990), where both X_N and X-n$\bar{x}$ were used to designate what is referred to as X_u here.)

Now, let $A(\gamma) = \hat{CV}_1(\gamma)/\hat{CV}_2(\gamma)$, so that we may think of $A(\gamma)$ as the factor which adjusts $\hat{cv}_2(\gamma)$ to obtain $\hat{cv}_1(\gamma)$.

$$A(\gamma)=\left[\frac{\sum_{i=n+1}^{N} x_i^{2\gamma} + (\sum_{i=n+1}^{N} x_i)^2/(\sum_{i=1}^{n} x_i^{2-2\gamma})}{(\sum_{i=n+1}^{N} x_i)^{2\gamma} + (\sum_{i=n+1}^{N} x_i)^2/(\sum_{i=1}^{n} x_i^{2-2\gamma})}\right]^{1/2},$$

where $(\sum_{i=n+1}^{N} x_i)^{2\gamma} = X_u^{2\gamma}$, summation over i=n+1 to N is for unobserved elements, and summation over i=1 to n is for the sample. $A(\gamma)$ is, therefore, the adjustment needed to convert $\hat{cv}_2$ to $\hat{cv}_1$. Note that $A(\gamma=0.5) = 1$. For $\gamma<0.5$, we have $A(\gamma)>1$, and, similarly, $A(\gamma>0.5)<1$.

If the ratio estimate (i.e., $\gamma=0.5$) can be used with some degree of comfort, then estimating the cv is not a particular problem. Also, when X_u is dominated by one x_i value, say x_k, such that $(\sum_{i=n+1}^{N} x_i)^{2\gamma} \approx x_k^{2\gamma}$, then $A(\gamma)\simeq 1$. But, what if the opposite extreme is the case? That is, consider $X_u = (N-n)x_u$, where $x_i=x_u$ for i=n+1,N. In that case,

$$A(\gamma) \text{ is } \left[\frac{x_u^{2\gamma-2} + (N-n)/(\sum_{i=1}^{n} x_i^{2-2\gamma})}{(N-n)^{2\gamma-1}\, x_u^{2\gamma-2} + (N-n)/(\sum x_i^{2-2\gamma})}\right]^{1/2}.$$

Let this be A_L $(\gamma>0.5)$ or A_{UP} $(\gamma<0.5)$ since, for $\gamma>0.5$, it is a lower bound on A, and for $\gamma<0.5$, it is an upper bound on A. For most cases of interest to electric power, it appears that $\gamma>0.5$, so $A_L < A < 1$. This can be a very wide interval. Note, however, that it is not likely that one element will dominate the unobserved portion of the population. The true value of A is likely to be much closer to the lower bound in most cases. As an attempt at a 'reasonable' estimate of A, consider A*, where we assume X_u is distributed over the N-n unobserved elements incrementally from (near) zero to $2X_u/(N-n+1)$. (FORTRAN code is available from the author for calculating $\hat{cv}_2 A^*$.)

$$A^*(\gamma)=\left[\frac{SX + (\sum_{i=n+1}^{N} x_i)^2/(\sum_{i=1}^{n} x_i^{2-2\gamma})}{X_u^{2\gamma} + (\sum_{i=n+1}^{N} x_i)^2/(\sum_{i=1}^{n} x_i^{2-2\gamma})}\right]^{1/2},$$

where $$SX = \sum_{j=1}^{N-n}\left[\frac{2jX_u}{(N-n+1)(N-n)}\right]^{2\gamma}.$$

EXAMPLES OF FORTRAN/MODEL OUTPUT:

A) From Figure C (nominal case - small n) -
N = 93; census result: Y = 751123;
n = 5 (i.e., 5 'largest'):

gamma	0.80	0.85	0.90
b	1.0632	1.0628	1.0623
est. of Y	751119	751119	751119
est. of cv	0.55%	0.65%	0.79%
b for $\lvert e_{0i}\rvert$	0.168	-0.030	-0.238

B) From Figure G (incompletely specified auxiliary data) -
(Note: This comes from a set of preliminary data which needs to be edited. Also, N is approximate.)
N = 925; n = 325:

gamma	0.85	0.90	0.95
b	0.000256	0.000267	0.000278
est. of Y	19243	19591	19956
est. of CV_2	20%	28%	38%
est. of A_L	0.16913	0.12610	0.09340
est. of CV_2A_L	3.4%	3.5%	3.5%
est. of A*	0.17569	0.13168	0.09810
est. of CV_2A^*	3.5%	3.6%	3.7%
b for $\lvert e_{0i}\rvert$	$1.78 *10^{-10}$	$-0.175 *10^{-10}$	$-1.99 *10^{-10}$

REFERENCES:

Herbert, J.H., Kott, P.S. (1988). "Robust Variance Estimation in Linear Regression," Journal of Applied Statistics, Vol. 15, pp. 341-345.

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, American Statistical Association, pp. 748-753.

Konijn, H.S. (1973). Statistical Theory of Sample Survey Design and Analysis, North-Holland Pub. Co., and American Elsevier Pub. Co., Inc.

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

Royall, R.M., Cumberland, W.G. (1978). "Variance Estimation in Finite Population Sampling," Journal of the American Statistical Association, June 1978, Vol. 73, pp. 351-358.

Thanks to Phil Kott for some helpful comments, and Nancy Kirkendall for an informative lecture.

GRAPHS:

A relationship between data of interest (on the y-axis), and auxiliary data (x-axis) is shown in each graph. Figures A and B show plant capacities used to estimate the total cost of operating. Figure C shows current and previous period sales volumes, and similarly for revenue in Figure D. In Figure E we see current and previous period fuel volume receipts. Figure F shows plant capacities used to estimate fuel volume receipts. In this case, as with surveying generation volumes, model sampling may be used to impute for the smallest plants, rather than insist on a complete census. In Figure G, sales to utilities by non-utilities are shown to be a help in estimating total capacity dedicated to the power grid.

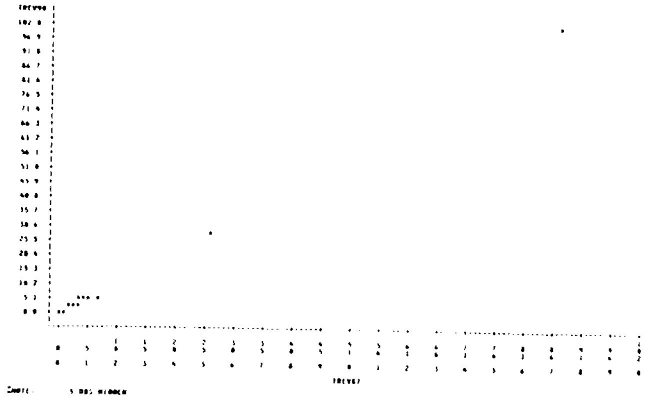

FIGURE D

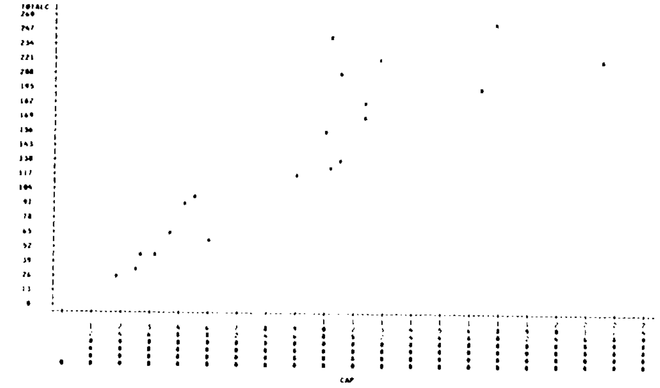

FIGURE A

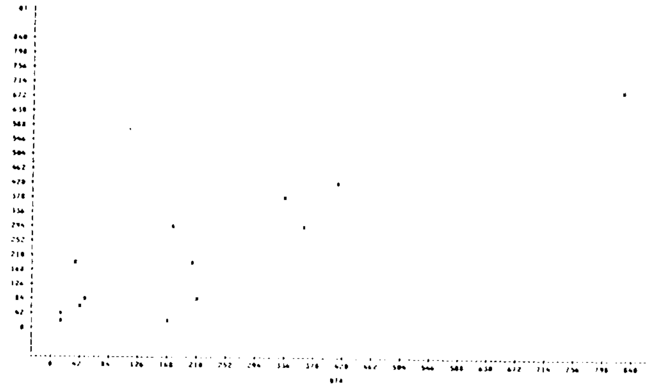

FIGURE E

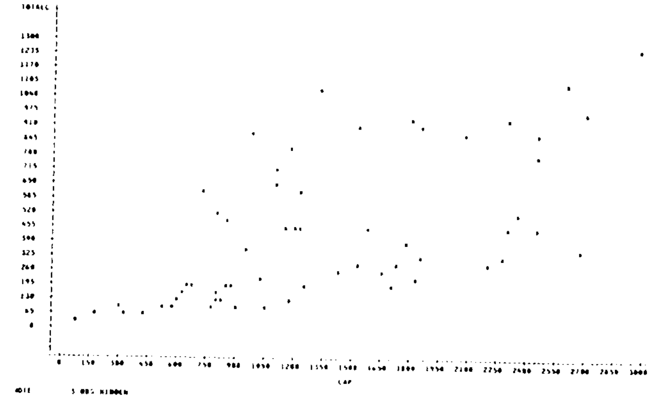

FIGURE B

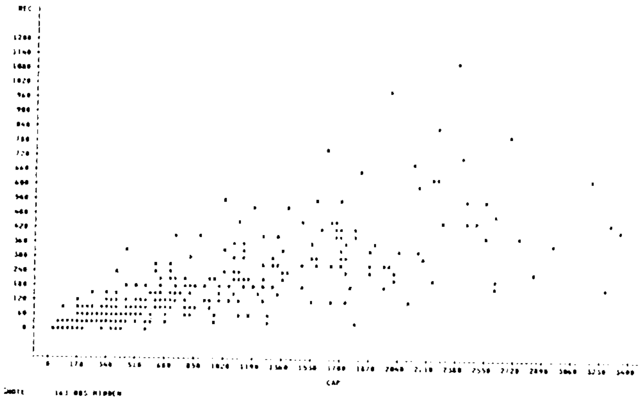

FIGURE F

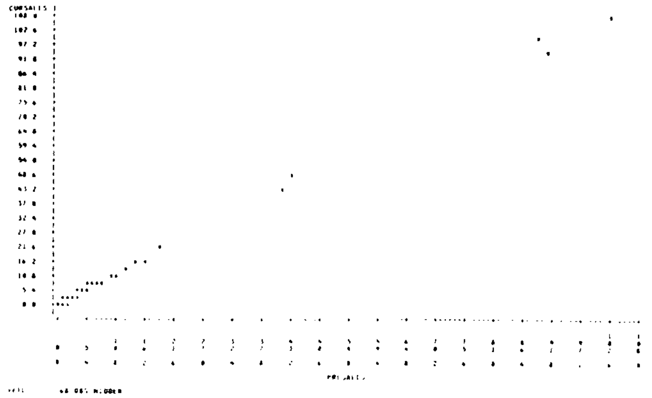

FIGURE C

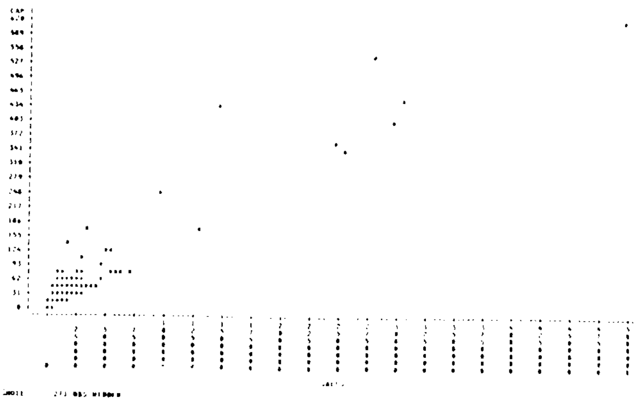

FIGURE G