Model-Based Sampling, Inference and Imputation

James R. Knaub, Jr., Energy Information Administration, EI-53.1
James.Knaub@eia.doe.gov

Key Words:
Survey statistics, Randomization, Conditionality, Random sampling, Cutoff sampling

Abstract:
Picking a sample through some randomization mechanism, such as random sampling within groups (stratified random sampling), or, say, sampling every fifth item (systematic random sampling), may be familiar to a lot of people. These are design-based samples. Estimates of means and totals for an entire population may be inferred from such a sample, along with estimation of the amount of error that might be expected. However, inference based on a sample and its (modeled) relationship to other data may be less familiar. If there is enough information in the sample and the related data are very useful, randomization may not be needed, and sometimes not even be desirable. In that case, one has a model, and results are conditional upon the sample. Therefore the sample integrity is very important. Even though the sample may not be drawn at random, it can not be drawn by only keeping data that conform to a preset notion of what should be observed. That is, even when data are not drawn at random, there have to be rules for constructing the sample. Note that cutoff, model-based sampling, described later and illustrated in three figures, is an alternative to design-based sampling that often performs very well for surveys of establishments. Such survey data tend to be highly skewed. That is, establishment survey data are normally dominated by a few relatively large values, with many relatively small ones.

Background:

Survey statistics are collected either through a census of all members of the target population, or by sampling. In either situation, substituting for data in the case of nonresponse is referred to as imputation. Sampling depends either on the Principle of Randomization, or conditionality. Design-based sampling and inference depend upon the former principle, and model-based inference depends upon the latter. However, model-based sampling can make use of randomization, and, further, the form of a design-based sample can be guided by the modeling of data. This latter point is an important part of the material found in Cochran (1977). In Chaudhuri and Stenger (1992), we see treatment of both design-based and model-based sampling and inference. Estimates of statistics, typically means or totals, are inferred from the sample. Thus the word “inference” is used.

Inference from a design-based sample may be aided by a model (referred to as 'model assisted' methods). See Sarndal, C.-E., Swensson, B. and Wretman, J. (1992). In Brewer(1995) and Knaub(1991), we see combinations of design-based and model-based inference.

Imputation in design-based sampling can be accomplished a number of ways, but the impact on data quality may be somewhat complex to analyze. This can be done (see Steel and Shao (1997)), but such an exercise may seldom have been accomplished. However, imputation in model-based sampling is performed in the same manner as estimation for such a sample, and as such, the impact on the results is directly estimated by a model-based variance estimate, although this is somewhat susceptible to model failure. (No model hypothesized will describe the relationship between the variate of interest and the optimal set of regressors with complete accuracy.) The missing observations are treated fundamentally as any other part of the population that is not in the sample. As with other methods of imputation, this method can also be used when a census was intended, and is one of the most useful methods available. Its usefulness is enhanced when dealing with imputation for a census because the model-based variance estimate may still help one determine if one is imputing for too much of the data. See Knaub (1999) for a new method of using existing, prediction-oriented software to do this.

In establishment surveys, data may be very skewed (i.e., with a few large values and relatively many small values) so that a cutoff sample of the largest establishments may be quite practical. Treatment of heteroscedasticity (a phenomenon involving variance) for optimizing estimation may be substantially altered if imputation is done for one or more 'larger' establishments. (See Knaub(1997). Note that “heteroscedasticity” refers to the nonconstant variance of the data about the regression line. Notice in Figures 1, 2 and 3 that data closer to the origin may have less variance.)

Model-based variance estimation will yield some indication as to whether results should be released without spending more time trying to obtain missing responses. However, as in design-based sampling, benefits of employing randomization (here, to reduce the impact of model failure) are eroded when imputation is involved, because data observed are no longer strictly obtained through randomization.

The word "model" can mean many things. In this context, however, it refers to a special algebraic estimation of a statistic, using one or more variables (regressors) and a “residual” term to represent unknown information. The most common format for use with survey data may be that of a single regressor, linear model, with (often) a zero intercept. This can be very useful when there is high correlation between the regressor and the variate of interest, such as when they represent the same data element, with the former collected in a previous (perhaps annual) census, and the latter collected in the current (perhaps monthly) sample. This concept of a model links survey statistics to econometrics, although econometric models are usually more complex.

Model-Based Sampling and Inference:

The 'modern' origins of model-based sampling are traceable to Brewer (1963) and Royall (1970), and according to P.S. Kott, Cochran (1953), pages 210-212. Similarity to material found in econometrics texts, such as Maddala (1977), is readily observable. Model-based

variance estimation is directly derived from the theory of least squares estimation found in econometrics, and as in econometrics, heteroscedasticity is a primary concern. However, in econometrics, often with more complex regression, heteroscedasticity seems to have been treated somewhat as a nuisance, something to be circumvented, whereas in survey statistics, heteroscedasticity has been explored perhaps more as a means toward greater accuracy in estimation. (See Knaub(1997).) Heteroscedasticity is very important in analytical data studies also. (See Carroll and Ruppert (1988).) Alternative variance estimates, seeking greater robustness can be found in Royall and Cumberland (1981), but they may be of limited added value. (See the corresponding figure in Knaub (1992) concerning one of these alternative variance estimators and its limited usefulness observed there.)

Using model-based inference with a cutoff sample may, at times, have advantages over design-based sampling. This could occur, and has occurred, for instance, when the data are highly skewed, and nonsampling error, time and cost considerations indicate that the data from the smallest entities are not efficiently obtainable. (In the case of an annual census of electric utilities, and a monthly sample, a small utility may not, for example, read its customers' meters more often than once every two or three months. This has happened, and this sort of thing can complicate the collection of accurate monthly observations!) A model may sometimes provide for better estimation than a design-based sample could, given these data quality considerations. Such a cutoff model sample will now be described with the help of three figures.

In Figure 1, we see a graphical illustration of a linear regression using hypothetical survey data, with one regressor. A two dimensional graph illustrates the one regressor case. An n+1 dimensional graph would illustrate an n regressor case. Figure 2 indicates what part of the data for the variate of interest would be 'cutoff' based on a cutoff level established on the single regressor shown. In Figure 3, we see an illustration as to how estimation under cutoff, model-based sampling may be understood. If one were to imagine that the regressor values are plotted on the x-axis for all members of the population, and the variate of interest is plotted on the y-axis for the observed sample, then the regression line through those points is used to estimate for the data not directly obtained due to sampling. For each value of the regressor (X) not having a corresponding value (O) in the sample, the regression line can be used to estimate the latter value, as in the two "X" to "O" data point illustrations shown.

This illustrates the usual situation, that is, where all regressor values are known. However, under a particular, but quite reasonable model specification (see Knaub(1992)), regressor values for members of the population not in the sample, may be presented as a subtotal.

Captions on Figures 1 through 3 can be used to guide one through the process, especially if model-based sampling and inference are not familiar to the reader.

References:

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

Brewer, K.R.W. (1995), "Combining Design-Based and Model-Based Inference," **Business Survey Methods**, ed. by B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott, John Wiley & Sons, pp. 589-606.

Carroll, R.J., and Ruppert, D. (1988), **Transformation and Weighting in Regression**, Chapman & Hall.

Chaudhuri, A. and Stenger, H. (1992), **Survey Sampling: Theory and Methods**, Marcel Dekker, Inc.

Cochran, W.G.(1953), **Sampling Techniques**, 1st ed., John Wiley & Sons.

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

Knaub, J.R.., Jr. (1991), "Some Applications of Model Sampling to Electric Power Data," **Proceedings of the Section on Survey Research Methods**, American Statistical Association, pp. 773-778.

Knaub, J.R., Jr. (1992), "More Model Sampling and Analyses Applied to Electric Power Data," **Proceedings of the Section on Survey Research Methods**, American Statistical Association, pp. 876-881.

Knaub, J.R., Jr. (1997), "Weighting in Regression for Use in Survey Methodology," **InterStat**, April 1997, electronic journal at URL http://interstat.stat.vt.edu/InterStat. (Note shorter, but improved version in the 1997 **Proceedings of the Section on Survey Research Methods**, American Statistical Association, pp. 153-157.)

Knaub, J.R., Jr. (1999), "Using Prediction-Oriented Software for Survey Estimation," **InterStat**, August 1999, electronic journal at URL http://interstat.stat.vt.edu/InterStat.

Maddala, G.S. (1977), **Econometrics**, John Wiley & Sons.

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

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

Sarndal, C.-E., Swensson, B. and Wretman, J. (1992), **Model Assisted Survey Sampling**, Springer-Verlag.

Steel, P.M. and Shao, J. (1997), "Estimation of Variance Due to Imputation in the Transportation Annual Survey (TAS)," to appear in the 1997 **Proceedings of the Section on Survey Research Methods**, American Statistical Association.

Figure 1

One Regressor Case

Illustration where not only are all regressor values known, but also, all values for the variate of interest are known. This would be useful as test data.

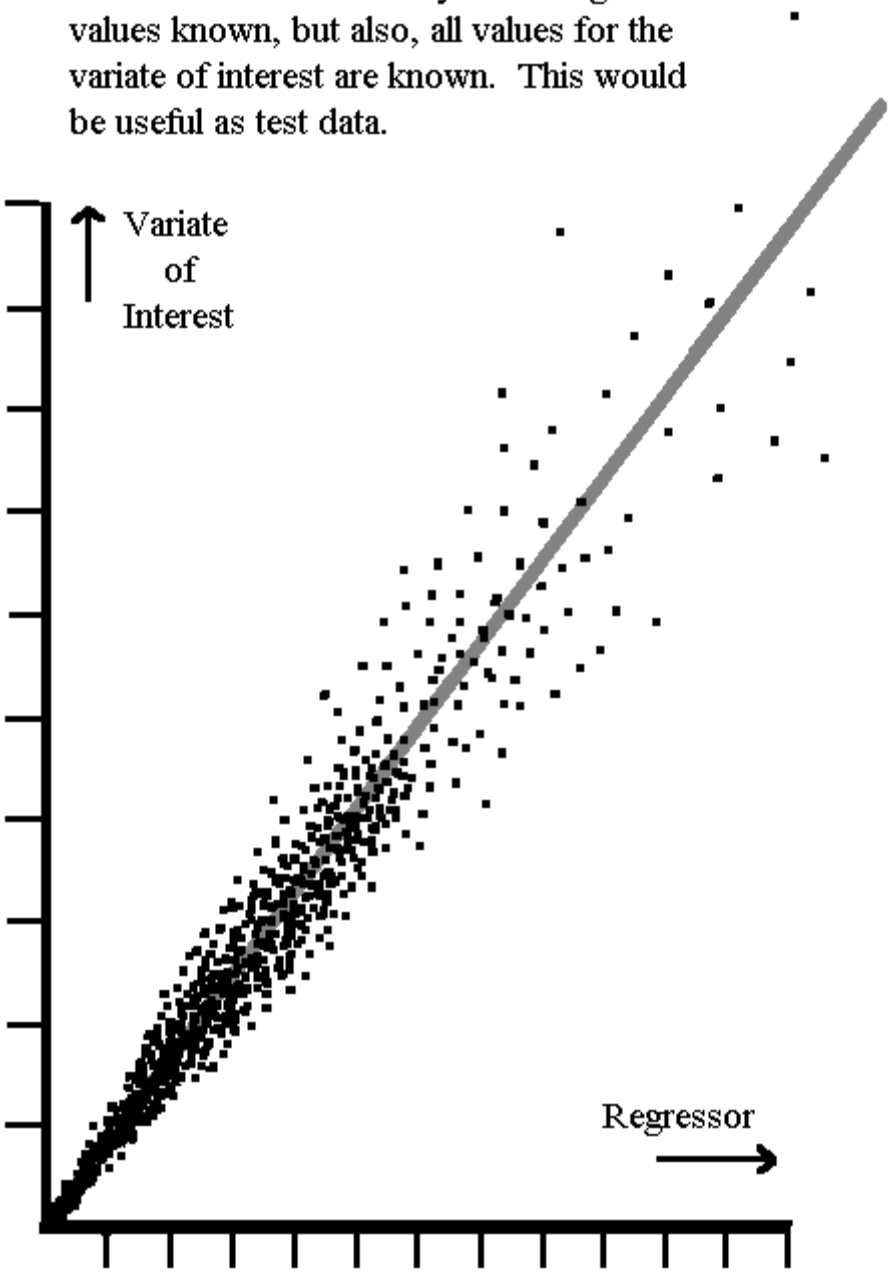

Figure 2

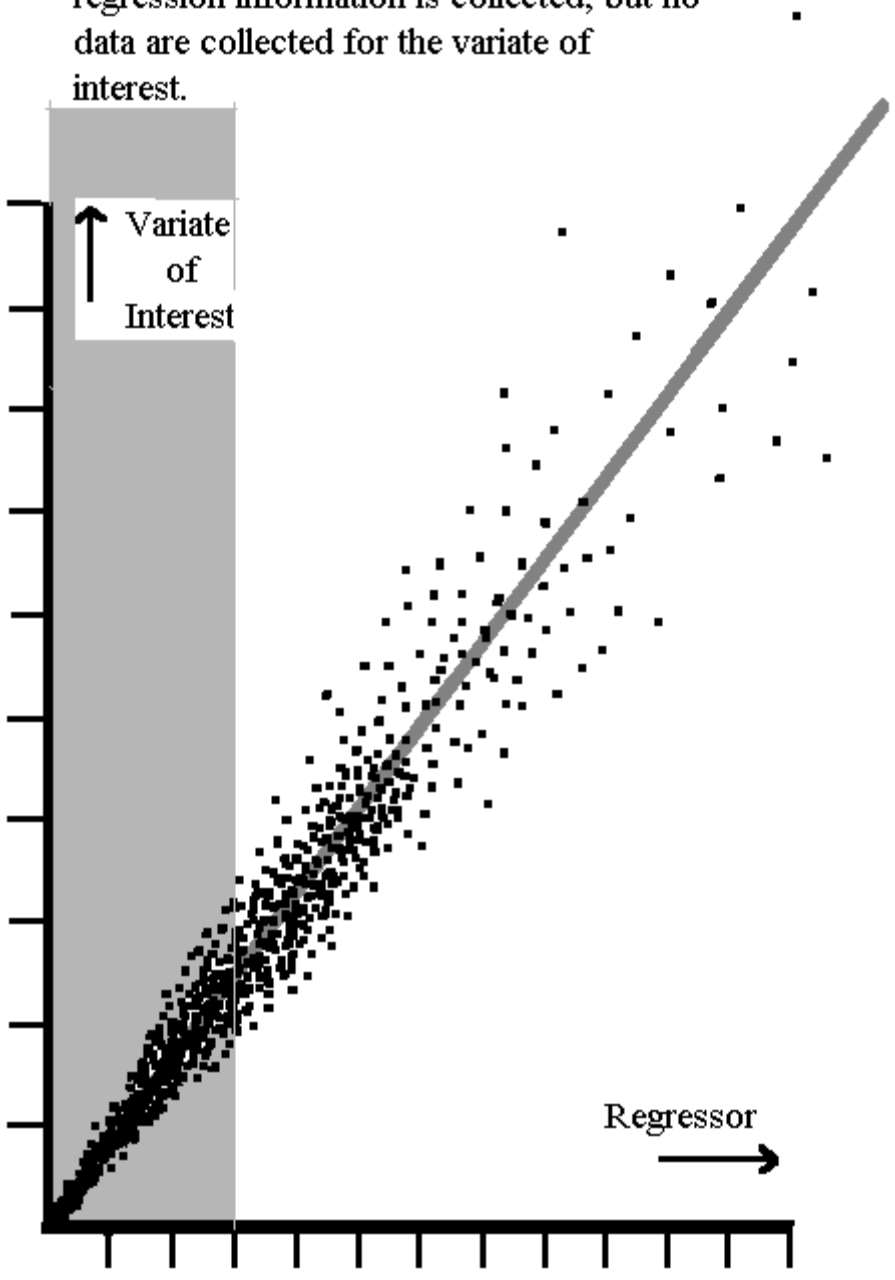
Shaded area represents "cutoff" region where regression information is collected, but no data are collected for the variate of interest.
Variate of Interest
Regressor

Figure 3

Example "X" values of the regressor translate to estimated "O" values for the variate of interest