응용통계연구 제 6권 1호
1993년 3월 pp. 29 - 40

# Error-Robust Model-Based Sampling in Accounting

Young-Il Kim[1)]

## Abstract

In a model-based sampling problem, it often happens that the functional form of variance of error terms in regression model cannot be specified in an exact form. The goal of error-robust sampling design will be to minimize the 'ill effects' resulting from a lack of knowledge of the error structure. A sampling criterion, which is optimal if it minimizes the average of an inefficiency measure when taken with respect to all candidate error structures, is proposed and a computer algorithm is developed for construction of optimal sampling plans. Auditing problem is of particular relevance because of the uncertainty that currently clouds specification of the error structure.

## 1. Introduction

Model-based samplers such as Brewer(1963) and Royall(1970) attempted to derive an inference from considering the values $y_1, \cdots, y_N$ associated with the $N$ units of the population as the realized (but still unknown) outcome of real random variables $Y_1, \cdots, Y_N$. The $N$-dimensional joint distribution of $Y_1, \cdots, Y_N$ is denoted as $\xi$. As implied, a "model" defines a class of distributions $\xi$. In this paper we consider only linear regression models and follow the same notation as Wynn (1977). A finite population S constitutes N units labeled $i = 1, \cdots, N$. Unit i has an attribute $x_i$ which may be a vector of auxiliary (known) variables in $R^k$ ($i = 1, \cdots, N$). Also unit $i$ has attribute $Y_i$ which is unknown before the sampling. We assume a superpopulation model in which $Y_i$ is a random variable. The assumptions are as follows:

1) Department of Industrial Information, ChungAng University, San 40-1, NaeRee, DaeDeokMyun, AhnSungKun, KyungGiDo , 456-756

$$\begin{aligned} &E_\xi(Y_i)=f(x_i)^T\theta \\ &Var(Y_i)=\sigma^2 v_i \\ &Cov(Y_i,Y_j)=0 \quad (i,j=1,\cdots,N \ ; \ i\neq j) \end{aligned} \tag{1}$$

where $f(x)=(f_1(x),\cdots f_k(x))^T$, $\theta=(\theta_1,\cdots,\theta_k)^T$ is a vector of unknown parameters, and $\sigma^2 v_i$ is assumed known.

Once the superpopulation model is determined, a sample size of $n$ units is chosen from $S$, purposely, and the corresponding values of $Y_i$ ; $i\in s$ are observed. We estimate some linear function $\tau=\sum c_i Y_i$ of the $Y_i$. We seek an estimator $T$ for $\tau$ which is unbiased in the sense of $E_\xi E(T)=E_\xi(\tau)$ and minimizes the $E(T-\tau)^2$, the mean square error. It can be shown by the Gauss Markov theorem that the estimator $T$ takes the form

$$T=\sum_{i\in s} c_i Y_i+\sum_{i\in s'} c_i \hat{Y}_i$$

where $s'$ is the complement of $s$ in $S$ (i.e., the non-sampled units), and $\hat{Y}_i$ is the weighted least squares estimate of $E(Y_i)$ based on $Y_i : i\in s$. Here we consider the case $\tau=\sum_{i\in S} Y_i$ only for simplicity.

One criterion for choosing a sampling design is to choose $s\subset S$ to minimize $E(T-\tau)^2$. Hereafter such a sampling design to minimize the mean square error will be called optimal. We note that, by definition of $E(T-\tau)^2$ and some algebra, the following is true,

$$E(T-\tau)^2=\sigma^2\sum_{i\in s'} v_i+\sigma^2\sum_{i,j}\sum_{\in s'} f(x_i)^T(X_s^T V^{-1}X_s)^{-1}f(x_j) \tag{2}$$

where $X_s=(f(x_1),\cdots,f(x_n))^T$ and $V$ is the diagonal matrix with entries $\sigma^2 v_i$, $\cdots$, $\sigma^2 v_n$. From (2) it is clear that $E(T-\tau)^2$ can be minimized by effective selection of s. Therefore, the process of selecting a sample is not necessarily random and is chosen on the basis of the model and knowledge of all the $x_i$, for

$i \in S$. A sampling design problem, specified by the triplet $(f,S,v)$, is solved by selection of $s$ for the model $f$, the population $S$ and the variance function $v$.

In model-based sampling, we must decide on a strategy (s,T), not only on a predictor T. Since an optimal predictor $T^*$ minimizes $E(T-\tau)^2$ for every fixed sample, s, the question of choice of sampling design enters at the stage of pre-sampling. That is to say, once $T^*$ has been determined, we should minimize the mean square error via choice of s. If the result is $s^*$, then $(s^*,T^*)$ is an optimal strategy for the criterion to minimize the $E(T-\tau)^2$. Often $s^*$ turns out to be purposive in the sense of selecting one particular set s with probability one.

The thought of using such non-randomized selection is hard to accept among design-based samplers. In an effort to counter the robustness criticisms, Royall and Herson (1973) presented the concept of "balanced sampling ". They showed that when we have $Var(Y_i)=\sigma^2$under (1), then a balanced sample, one for which

$$\frac{1}{n}\sum_{i\in s} f_j(x_i) = \frac{1}{N}\sum_{i\in S} f_j(x_i), \quad (j=1,\cdots,k)$$

minimizes $E(T-\tau)^2$.

On the other hand, Wynn (1977) presented an alternative sampling design criterion closely related to the D- and G-criteria in experimental design. A new "continuous theory" for finite population sampling was given by Wynn (1976). This approach would be very appealing if sampler is interested in inference concerning $\theta$ or prediction of non-sampled units.

In auditing sampling, we have a rather firm knowledge of superpopulation model. Knowledge concerning the error structure, however, is fairly weak. Thus neither optimal model-based sampling nor model-based balanced sampling meet our needs for robustness. In the following two sections, we consider the selection and construction of samples that are robust with respect to the misspecification of error terms.

## 2. Error-Robust Sampling in Accounting

Recently, model-based sampling has been advocated by various auditors in the accounting setting (See Ko, 1986, and Gofrey, et. al., 1984). Knowledge about the

accounting population can be summarized in a model that describes the joint distribution of $Y_1, \cdots, Y_N$. The primary variable of interest ($Y_i$) in auditing is the audited account value which is usually accompanied by one auxiliary variable ($x_i$), the book values. Ideally, the company's accounting and internal control systems should operate in such a way that the reported book value of a unit in the accounting population should be its true audit value. However, due to the errors introduced into the accounting system, deviations between book value and audit value occur.

Empirical studies have repeatedly suggested that observed account values tend to be linearly related to the corresponding book values, as one might expect. Intuitively, one might also expect to find larger errors associated with larger accounts. As a result the following model has been proposed for most accounting populations (Johnson, et al.,1981).

$$\begin{aligned} &E_\xi(Y_i) = \theta_0 + \theta_1 X_i \\ &Var(Y_i) = \sigma^2 X_i^\alpha, \quad \text{for some value of } \alpha \\ &Cov(Y_i, Y_j) = 0, \quad (i \text{ and } j = 1, \cdots, N, \ i \neq j) \end{aligned} \tag{3}$$

It is generally admitted however that plausible values of $\alpha$ are in the range [0,1]. In order to understand the implication of uncertainty with regard to the error structure, optimal sampling designs were obtained for different $\alpha$'s in the model (3) for a hypothetical data set consisting of the integer values from 1 to 15. As $\alpha$ varies from 0 to 1 we note a resultant shift in the sample mean (See Table 1). Evidently, the sample mean is increasing as the value of $\alpha$ increases. The algorithm necessary to find the optimum sampling design is discussed later.

To compare samples, an efficiency measure is needed. We give the following.

**Definition 1.** The MSE efficiency of a sampling design $s$ for $(f,S,v)$ with respect to $s^*$ is

$$\text{MSEE}(s, s^*, (f,S,v)) = \text{MSE}(s^*)/\text{MSE}(s)$$

For brevity, we will take the dependence on the triplet $(f,S,v)$ as implied, and simply use $\text{MSEE}(s,s^*) = \text{MSE}(s^*)/\text{MSE}(s)$. For $s^*$ = optimal sampling design,

$$MSEE(s,s^*)=MSEE(s)$$

Table 2 shows the MSE efficiencies with respect to the optimal sampling design under varying assumptions about $\alpha$. For example, the first line of the Table 2 summarizes how badly the balanced sample can do if the sampler assumes constant error terms will be necessary ($\alpha$ = 0). From this Table it is not unreasonable to suggest that the sample which is obtained by setting $\alpha$ equal to a middle value between 0 and 1, say 0.4 or 0.6, will be robust against the two possible extreme cases.

Suppose we assumed a class, A, of competing $\alpha$ values that could be reasonably expected to describe the accounting population from past data. Let A = $\{\alpha \mid 0 \leq \alpha \leq 1\}$. To evaluate the efficiency of a particular sampling design for $\alpha \in A$, we rely on in Definition 1. $MSEE_\alpha^{-1}(s)$ can be called the inefficiency measure with respect to optimal sampling design for a particular value of $\alpha$.

Definition 2. For a class of competing $\alpha$ values, a sampling design $s$ is said to be $\overline{M}$-optimal if it minimizes the average inefficiency over A

$$\overline{M}(s)=\min_{s\subset S}\int_A MSEE_\alpha^{-1}(s)d\beta(\alpha)$$

where $\beta$ is a user specified probability measure on A which could be provided by the auditor. This criterion provides a measure for optimality of a design with respect to several competing $\alpha$'s. This approach is a simple modification of $\overline{L}$ -optimality suggested by Cook and Nachtsheim (1982). $\overline{M}$-optimal sampling approach was applied to the simple linear regression superpopulation model with the same hypothetical population as before with A = {0, 0.2, 0.4, 0.6, 0.8, 1.0} (Table 3). Ideally, A = [0,1]. However, numerical construction of the (nearly) optimal sample required discretization of A. A simple modification of Johnson and Nachtsheim's k-exchange algorithm (1983) for experimental design was needed to construct the error-robust sampling design. Similar algorithm can be applied to find the optimum model-based sampling. A practical routine has not been available in sampling literatures.

Algorithm

1. Form a sample of size $n$ for which $X_{s_n}^T V^{-1} X_{s_n}$ is nonsingular, $s_n$.
2. For each $i$, $i=1,2,\cdots,n$,
   (2.1) Remove the i-th sampled unit from $s_n$ giving $s_{n-1}^{-i}$
   (2.2) Compute $\overline{M}(s_{n-1}^{-i})$.
3. Let

$$\overline{M}(s_{n-1}^{-j}) = \min_{i \in I_n} \overline{M}(s_{n-1}^{-1}),\ I_n = (1, \cdots, n)$$

   Delete the $j$-th unit from $s_n$ giving $s_{n-1}^{-j}$
4. For each $i \in I_{N-(n-1)}$, $I_{N-(n-1)} = (1, \cdots, N-(n-1))$
   (4.1) Add the i-th non sampled unit to $s_{n-1}^{-j}$, giving $s_n^{+i}$.
   (4.2) Compute $\overline{M}(s_n^{+i})$
5. Let

$$\overline{M}(s_n^{+k}) = \min_{i \in I_{N-(n-1)}} \overline{M}(s_n^{+i})$$

   Add the $k$-th unit to $s_{n-1}^{-j}$ giving $s_n^{+}$.
6. Repeat steps 2-5 until $\overline{M}(s_n^{+})$ cannot be improved. This sample will be regarded as nearly optimal.

Note this algorithm requires the input of optimum MSEs of samples size of $n$ and $n-1$ for each $\alpha \in A$. Further modification was made to avoid the local minimum in such a way that the best k units (not one unit) are saved in Step 3 and sequentially exchanged for k non-sampled units in Steps 4 and 5. If there is an improvement, then repeat Steps 2-5.

The third row of Table 3 indicates an error-robust sampling design when we have equal probability for each $\alpha$. For illustrative purposes, the average efficiency is computed for the sampling design of the third row of Table 3. The efficiencies are 91.8, 96.9, 99.2, 98.1, 90.3 and 84.3% for each $\alpha$ in increasing order respectively. Therefore, the average efficiency from this data set is 94.3%. Notice that the first

and the fifth designs are reflecting the extreme differences in probabilities. The main advantage of this approach is its ability to reflect the sampler's prior belief about the likelihood of each $\alpha \in A$.

## 3. Higher Order Model

An alternative method for balanced sampling design, optimal sampling design under more general model, was said to be under investigation by Royall and Herson (1973) when the model is not exactly known (They were only concerned about the $\alpha = 1$ case). Suppose we have the following model

$$E_\xi(Y_i) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2$$
$$Var(Y_i) = \sigma^2 x_i^\alpha \quad \text{for some fixed } \alpha$$
$$Cov(Y_i, Y_j) = 0 \quad (i,j = 1,2,\cdots,N \ : \ i \neq j)$$

Table 4 shows optimal sampling designs for various $\alpha$'s. The sample mean has a slower pace of increase than exhibited in Table 1. Overall, the MSE efficiencies are extremely high. This can be explained in part by the more uniform spread of sampled units. The extreme error-robustness exhibited by the optimal design for $\alpha = .6$ led to the question: Is this design similarly model-robust? To study this question, we computed the MSE efficiencies of the optimal sampling designs for quadratic regression under the assumption that the simple linear regression is true for varying $\alpha$.

Results are summarized in Table 6. Surprisingly, the worst case efficiency (67%) occurs when we assumed $\alpha = 0.0$ is appropriate and the true $\alpha = 1.0$. As we notice in Table 6, for a given assumption about $\alpha$, the MSE efficiency decreases with increasing true $\alpha$. This is because when the simple linear regression is true, the spread of the sample under the assumption of quadratic regression is quite different from that under simple linear regression with $\alpha = 1.0$. This low efficiency is not significantly increased even when we assume $\alpha = 1.0$ (76%). Overall, the MSE Efficiencies are fairly high. When the simple linear regression is assumed and quadratic regression is true, the MSE efficiencies will be significantly worse and therefore omitted. If the auditor were to assume $\alpha = .6$ and quadratic regression for purposes of sample construction, his expected MSE efficiency will be 88.6% if the model turns out to be linear. This is 5.76% lower

than the one we achieved using $\overline{M}$-optimal sampling design for linear regression, 94.3%, the simple average of six efficiencies shown in section 2, 91.8, 96.9, 99.2, 98.1, 90.3, and 84.3%. But the former plan will be extremely efficient if the true superpopulation model is indeed quadratic (Table 5). Similar arguments were discussed in many situations of sampling and optimal experimental design (See Royall and Herson, 1973, Kussmaul, 1969 and Cook and Nachtsheim, 1982).

## 4. Concluding Remarks

In this paper we have conducted research directed toward the characterization of model-based sampling designs that are insensitive to the specification of error structure. A number of implications of the findings for further research, though limited, are possible for audit samplers as summarized in the following:

1. For the case when the superpopulation model as well as the variance function is known, algorithms have been given for constructing optimal samples for arbitrary regression models. Unfortunately the efficiency of designs produced can be highly sensitive to changes in assumptions about the error structure.
2. Often in accounting, the model is fairly well determined while the error structure is not. In the formulation of the variance function, the choice of $\alpha$ is not always clear. in such cases we recommend that the sampler, at a minimum, define the space of collection of possible $\alpha$'s and take the optimal sampling design associated with an intermediate value of $\alpha$. If computing resources permit, we recommend construction of the $\overline{M}$ -optimal sampling design.
3. For the case in which the model as well as variance function is unknown, our results suggest the optimal sampling design for the higher degree polynomial regression model (i.e., use quadratic regression model when the simple linear regression model is in doubt and the value of $\alpha$ is not determined) with an intermediate value of $\alpha$. This sampling design will cover robustness in both areas as illustrated partially in Table 5 and 6.

In summary, a strategy derived in blind reliance on a model-based sampling argument could give misleading estimates when the model is in error. This is particularly true if the error structure is misspecified. In order to be of practical use, and not only of theoretical interest, a model-based sampling design should be

formulated to be robust for a suitably broad family of models. In general, robustness considerations can be expected to favor "representative" selection rather than "extreme" selection.

Further simulation study considering various modelling situations and data set is under investigation to derive more reliable conclusions.

Table 1.
Simple Linear Regression
Data : (1(1)15)
$n$=5, $N$=15, $Var(Y)$ is proportional to $x^{\alpha}$

| | Optimal sample | | | | | Sample Mean |
|---|---|---|---|---|---|---|
| $\alpha = 0$ | 1 | 4 | 6 | 14 | 15 | 8.0 |
| $\alpha = 0.2$ | 1 | 2 | 12 | 14 | 15 | 8.8 |
| $\alpha = 0.4$ | 1 | 2 | 13 | 14 | 15 | 9.0 |
| $\alpha = 0.6$ | 1 | 9 | 13 | 14 | 15 | 10.4 |
| $\alpha = 0.8$ | 1 | 12 | 13 | 14 | 15 | 11.0 |
| $\alpha = 1.0$ | 1 | 12 | 13 | 14 | 15 | 11.0 |

Table 2.
MSE Efficiencies for simple Linear Regression Model.

| Assumed $\alpha$ | True $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| 0 | 1.00 | 0.97 | 0.92 | 0.86 | 0.75 | 0.67 |
| 0.2 | 0.97 | 1.00 | 0.99 | 0.95 | 0.85 | 0.77 |
| 0.4 | 0.96 | 0.99 | 1.00 | 0.97 | 0.87 | 0.80 |
| 0.6 | 0.75 | 0.86 | 0.95 | 1.00 | 0.97 | 0.94 |
| 0.8 | 0.66 | 0.79 | 0.91 | 0.99 | 1.00 | 1.00 |
| 1.0 | 0.66 | 0.79 | 0.91 | 0.99 | 1.00 | 1.00 |

Table 3

| Priors for $\alpha$ = | | | | | | Error-Robust Sampling Design | | | | | Sample Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 | | | | | | |
| 0.8 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 3 | 4 | 5 | 14 | 15 | 8.2 |
| 0.5 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 1 | 2 | 13 | 14 | 15 | 9.0 |
| 0.16 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 | 1 | 4 | 13 | 14 | 15 | 9.4 |
| 0.1 | 0.1 | 0.3 | 0.3 | 0.1 | 0.1 | 1 | 4 | 13 | 14 | 15 | 9.4 |
| 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.5 | 1 | 12 | 13 | 14 | 15 | 11.0 |

Note: $6 \times 0.16$ is not summed up to 1 due to truncation error.

Table 4
Quadratic Regression
Data : (1(1)15)
$n$ = 5, $N$ = 15, $Var(Y)$ is proportional to $x^{\alpha}$

| | Optimal Sample | | | | | Sample Mean |
|---|---|---|---|---|---|---|
| $\alpha$ = 0 | 1 | 6 | 8 | 12 | 13 | 8.0 |
| $\alpha$ = 0.2 | 1 | 8 | 9 | 10 | 14 | 8.4 |
| $\alpha$ = 0.4 | 1 | 8 | 9 | 10 | 14 | 8.4 |
| $\alpha$ = 0.6 | 1 | 8 | 9 | 10 | 14 | 8.4 |
| $\alpha$ = 0.8 | 2 | 8 | 9 | 10 | 15 | 8.8 |
| $\alpha$ = 1.0 | 3 | 6 | 10 | 12 | 15 | 9.2 |

Table 5.
MSE Efficiencies for Quadratic Regression Model.

| Assumed $\alpha$ | True $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| 0 | 1.00 | 0.99 | 0.97 | 0.95 | 0.91 | 0.88 |
| 0.2 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| 0.4 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| 0.6 | 0.99 | 1.00 | 1.00 | 1.00 | 0.99 | 0.99 |
| 0.8 | 0.94 | 0.97 | 0.99 | 0.99 | 1.00 | 0.99 |
| 1.0 | 0.89 | 0.91 | 0.94 | 0.98 | 0.98 | 1.00 |

Table 6
MSE Efficiencies of Optimal Sampling Design for Quadratic when Simple Linear Regression is true model

| Assumed $\alpha$ | True $\alpha$ | | | | | |
|---|---|---|---|---|---|---|
| | 0 | 0.2 | 0.4 | 0.6 | 0.8 | 1.0 |
| 0 | 1.00 | 0.97 | 0.91 | 0.85 | 0.74 | 0.67 |
| 0.2 | 0.99 | 0.98 | 0.95 | 0.89 | 0.79 | 0.72 |
| 0.4 | 0.99 | 0.98 | 0.95 | 0.89 | 0.79 | 0.72 |
| 0.6 | 0.99 | 0.98 | 0.95 | 0.89 | 0.79 | 0.72 |
| 0.8 | 0.95 | 0.95 | 0.93 | 0.90 | 0.81 | 0.76 |
| 1.0 | 0.89 | 0.89 | 0.88 | 0.86 | 0.80 | 0.76 |

# References

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

[2] Cook, R. D., and Nachtsheim, C. J. (1982), "Model Robust, Linear-Optimal Designs", *Technometrics*, **24**, 49-54.

[3] Gofrey, J., Roshwalb, A., and Wright, R. L. (1984), "Model-Based Stratification in Inventory Cost Estimation", *Journal of Business and Economic Statistics*, **2**, 1-9.

[4] Johnson, J. R., Leitch, R. A., and Neter, J. (1981), "Characteristics of Errors in Accounts Receivables and Inventory Audits", *The Accounting Review*, April, 270-293.

[5] Johnson, M. E., and Nachtsheim, C. J. (1983), "Some Guidelines for Constructing Exact D-optimal Designs on Convex Spaces", *Technometrics*, **25**, 271-277.

[6] Ko, C. E. (1986), "Alternative Statistical Inference Frameworks in Auditing Sampling", *Ph.D. Dissertation*, University of Minesota, Dept. of Accounting.

[7] Kussmaul, K. (1969), "Protection Against Assuming the Wrong Degree in Polynomial Regression", *Technometrics*, **11**, 677-682.

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

[9] Royall, R. M., and Herson, J. (1973), "Robust Estimation in Finite Populations I, II", *Journal of the American Statistical Association*, **68**, 880-893.

[10] Wynn, H. P. (1976), "Optimum Designs for Finite Population Sampling", *Statistical Decision Theory and Related Topics*, Edited by S. S. Gupta, and D. S. Moore, New York, Academic Press.

[11] Wynn, H. P. (1977), "Minimax Purposive Survey Sampling Design", *Journal of the American Statistical Association*, **72**, 655-657.

# 회계감사예에 적용시켜본 오차로버스터적 모델표본론

김영일[1)]

## 요 약

모델을 이용한 표본론에서는 오차에 대한 함수식이 불확실한 경우가 종종 발생되는데 이러한 오차에 대한 지식이 결여 되었을때 발생되는 잘못된 효과를 줄일 수 있는 방법이 연구되었다. 제시된 표본방법론은 모든 가능한 오차함수식에 대한 비효율성에 대한 평균을 최소화하는 데 그 목적이 있다. 컴퓨터를 이용한 알고리즘이 제시되었고 회계감사에 관련된 특수한 경우의 예를 들어 이러한 방법의 효율성을 알아 보았다.

[1)] (456-756) 경기도 안성군 대덕면 내리 산 40-1 중앙대학교 산업대학 산업정보학과