A COMPARISON OF THREE METHODS OF MODELING INPUT DISTRIBUTIONS

Stephen C. Hora
Associate Professor
College of Business Administration
Texas Tech University
Lubbock, Texas 79409

ABSTRACT

Three methods of estimating the inverse of a continuous cumulative distribution function for the purpose of random deviate generation are discussed. These methods are 1) the empirical approach, 2) the maximum likelihood approach, 3) a newly developed regression based estimation procedure.

Analytic results are obtained which permit comparisons of the accuracies of each of these methods under alternative assumptions about the underlying distribution. Expressions for the variance of each estimate at any given quantile of the random variable are provided.

A demonstration of the procedures is given using data from the outer continental shelf oil and gas lease program.

1. Introduction

Our aim is to provide an analytic comparison among three methods of estimating the inverse function of a continuous random variable. The estimated inverse function is intended for use in the generation of random deviates using the well known inverse function method [Kennedy and Gentle (1980)]. The three methods of estimation to be compared are the maximum likelihood estimation method (MLE), the empirical method, and a recently developed regression method. The comparisons are made on the basis of accuracy of the estimates rather than on the computational efficiencies and ease of generation of random deviates from the estimated inverse function.

In the following section, a description of each of the three methods is given and some of their salient properties are discussed. Section 3 contains the analytic comparisons. The variance of the quantile estimator for each of the three methods is derived under general conditions as the sample size becomes infinite. These variances are used to make comparisons on the basis of asymptotic relative efficiency (ARE). Some examples are given. In the final section a demonstration of the regression method is provided.

2. Description of the Methods

Let $X_1, \ldots, X_n$ be an independent random sample on a continuous random variable X having the distribution function F. When F is completely specified we write F(x) for $P(X \leq x)$ and when F is known up to a vector of parameters θ we write F(x;θ) for $P(X \leq x)$. Since F is assumed to be continous the inverse function of F given by

$$F^{-1}(p) = \inf \{x: F(x) = p\}$$

exists. If F^{-1} is estimated by the function $\hat{F}^{-1}$ then random deviate generation is easily accomplished by the generation of uniform random deviate, U, and the subsequent evaluation of $\hat{F}^{-1}(U)$.

The estimation of F^{-1} is of central concern here. There are two basic approaches to the solution of this problem. First we may estimate θ with, say, t and then use $F^{-1}(p;t)$ as an estimator of F^{-1}. The MLE procedure follows this approach. The second basic approach is to estimate the entire inverse function without directly estimating θ. The empirical procedure follows this second approach.

The advantage to the first approach, parametric estimation, is that if the correct family of distributions is chosen, the information in the sample can be used efficiently. The danger exists that the incorrect family will be chosen and thus, no matter how large the sample, the estimate of F^{-1} may differ substantially from the true function.

The particular form of parametric estimation examined here is MLE. In order to obtain an MLE estimator we consider the likelihood function

$$L = \prod_{i=1}^{n} f(X_i;\ \theta)$$

where $f(x;\theta) = dF(x;\theta)/dx$ is the density function of X. Under certain assumptions concerning the

81CH1709-5/81/0000-0309$00.75 © 1981 IEEE

properties of the derivatives of the logarithm of f, L will yield a single consistent solution, say, t. Furthermore, the limiting variance of the estimator t is

$V = ||v_{ij}||$ where $V^{-1} = ||v^{ij}||$ has the elements

$$v^{ij} = -E(\partial^2 \log L / \partial\theta_i \; \partial\theta_j).$$

This result will be useful in making comparisons of the MLE to other methods. [See Kendall and Stuart (1967) for a full discussion of MLE.]

The empirical method avoids the problem of specifying a family of distributions. Instead, the sample or empirical distribution function defined by

$$F_n(x) = \sum_{i=1}^{n} u(x-X_i)$$

where u(x) is the indicator function u(x)=1 if $x \geq 0$, and u(x)=0 otherwise. F_n (x) is itself a random variable. By the Glivenko theorem [Rao (1973)] F_n converges almost surely to F.

Therefore we may estimate F by F_n and F^{-1} by F_n^{-1} where (2.1) is used to define F_n^{-1}. Ogawa (1962) has shown that the limiting variance of F_n^{-1} (p) is given by

$$\mathrm{var}[F_n^{-1}(p)] = p(1-p)n^{-1} \; [f(x_p)]^{-2} \qquad (2.1)$$

where x_p satisfies $F(x_p) = p$. This last result will be used to make comparisons of the empirical method to the other two methods.

Last consider the regression method presented by Hora (1981). We give without the corresponding proofs the following results.

1. Define

$$\gamma[F(x)] = d \log[F_o(x)]/d \log[F(x)]$$

where $F_o(x)$ is a continuous distribution function having the same support as F(x). Assume that

$$\gamma(p) = \beta_o + \beta_1 \, p + \beta_2 \, p^2 + \ldots + \beta_r \, p^r$$

Then the inverse of F can be expressed in terms of the inverse of F_o as

$$F^{-1}(p)=F_o^{-1}[k \exp(\beta_o \log \, p+\beta_1 p+\beta_2 p^2/2+ \ldots+\beta_r p^r/r]$$

where $k=\exp \, (-\beta_1-\beta_2/2-,\ldots,-\beta_r/r)$.

2. Let $X_1,\ldots, X_n$ be the ascending order statistics in a sample of size n on X and let

$$W_j=j \; [\log F_o \; (X_{j+1}) - \log F_o \; (X_j)]$$

where $F_o(X_{n+1}) \equiv 1$. Then as $n \to \infty$ the W_j converge in probability to independent exponential random variables such that

$$E(W_j) = \gamma(p)$$

where $j \to np$ as $n \to \infty$.

3. From 2 and the polynomial representation of γ(p) we have

$$E(W_j) = \sum_{i=0}^{r} \beta_i \; [j/(n+1)]^i$$

as $n \to \infty$.

From the third result we see that the vector of coefficients $\beta' = (\beta_o,\ldots,\beta_r)$ can be estimated using weighted least squares. We construct the matrix of fixed regressors given by

$$A = ||a_{ij}|| \quad \text{where } a_{ij} = [j/(n+1)]^i$$

and obtain preliminary estimates of β by

$$\hat{\beta}^* = (A'A)^{-1}A'W$$

where $W' = (W_1,\ldots, W_n)$. The preliminary estimates are asymptotically unbiased for β but, noting the second result, not necessarily efficient. If β has some nonzero elements, other than β_o, the variances of the W_i will not be constant and the efficiency of the estimator can be improved by using generalized (here weighted) least squares. Let $\hat{W}_j$ be the jth element of $\hat{W}$ where

$$\hat{W} = A(A'A)^{-1}A'W.$$

Define $D = ||d_{ij}||$ so that $d_{jj} = W_j^{-2}$ and $d_{ij} = 0$ if $i \neq j$.

The improved estimator is then given by

$$\hat{\beta} = (A'DA)^{-1}A'DW.$$

The variance of $\hat{\beta}$ is asymptotically given by

$$\mathrm{var}(\hat{\beta}) = (A'\Delta A)^{-1}$$

where $\Delta \;\; ||\delta_{ij}|| \;\; , \;\; \delta_{jj} = \{\gamma[j/(n+1)]\}^{-2}$

and $\delta_{ij} = 0$ if $i \neq j$.

Now, backtracking somewhat, remember that the estimate of β was calculated using a reference distribution, F_o. We use our estimate of β, F_o, and the first result to estimate the inverse function F^{-1}. The selection of F_o and r, the order of the polynomial that connects F and F_o, should be made in such a manner that F_o is conducive to random deviate generation and r is as small as possible. When F_o is chosen to be close to F, the polynomial is required to do less work in bringing the inverse function into conformance with the data. A wise choice of both F_o and r can produce an estimated inverse function having substantially better efficiency (in statistical sense) than the empirical method without having to specify a family of distributions with a usable inverse as is required in parametric estimation.

3. Comparisons

Since it is our intention to provide comparisons of the accuracies and statistical efficiencies of the three competing methods of estimation, it is necessary that expressions similar to (2.1) be developed for the maximum likelihood approach and the regression approach. We begin with the MLE estimator.

Let $\hat{x}_p$ be an estimate of the pth quantile of the inverse function. Then for the MLE estimator t of θ we have as $n \to \infty$

$$\hat{x}_p - x_p = F^{-1}(p;t) - F^{-1}(p;\theta)$$

and

$$\text{var}(\hat{x}_p) = d_\ell' V d_\ell + o(n^{-1}) \qquad (3.1)$$

where $V = \text{var}(t)$ and $d_\ell = \partial F^{-1}(p;\theta)/\partial\theta$ evaluated at the true value of θ.

Similarly for the regression method we have

$$\hat{x}_p - x_p = F_o^{-1}[g(p,\hat{\beta})] - F_o^{-1}[g(p,\beta)]$$

where $g(p,\beta)=k \exp(\beta_o \log p + \beta_1 p + \beta_2 p^2/2 + \ldots\ldots + \beta_r/r)$.

By imposing certain mild conditions on the behavior of $\gamma(p)$ it can be shown that $\hat{\beta}$ converges in distribution to a normal random vector and, further, that

$$\text{var}(\hat{x}_p) = [f_o(x_p)]^{-2} d_r' (A'\Delta A)^{-1} d_r + o(n^{-1}) \qquad (3.2)$$

where $d_r = \partial g(p,\beta)/\partial\beta$ evaluated at the true value of β.

Through equations (2.1), (3.1), and (3.2) comparisons of asymptotic relative efficiency (ARE) can be made. First let us compare the empirical with the regression method. The ratio of (3.2) to (2.1) gives the ARE of the regression method to the empirical method as

$$\text{ARE}(R,E)=p(1-p)n^{-1}\,[f_o(x_p)/f(x_p)]^2\,[d_r'(A'\Delta A)^{-1}\,d_r]^{-1}$$

Evaluation of the ARE(R,E) then depends upon p,r, the choice of f_o and the true density f. In order to simplify the comparison we have chosen the case where $f=f_o$ and we have evaluated the ARE at p=.1,.2,...,.9 and r=0, 1,..., 5. These results are presented in the table.

TABLE

ARE of Regression to Empirical Method

r\p	.1	.2	.3	.4	.5	.6	.7	.8	.9
0	1.70	1.54	1.61	1.79	2.08	2.55	3.37	5.02	10.01
1	1.49	1.54	1.49	1.39	1.31	1.30	1.40	1.74	2.93
2	1.47	1.30	1.22	1.23	1.28	1.29	1.25	1.27	1.67
3	1.32	1.19	1.21	1.19	1.16	1.18	1.23	1.23	1.31
4	1.21	1.19	1.15	1.13	1.15	1.14	1.14	1.21	1.27
5	1.17	1.15	1.12	1.13	1.11	1.13	1.13	1.15	1.21

Examination of the table shows that the ARE(R,E) is always greater than unity. We conjecture that this is true whenever $f=f_o$ and that as $r \to \infty$ the ARE(R,E) $\to$ 1. Remembering that ARE is to be interpreted as the ratio of sample sizes required to obtain the same level of precision, we see that the regression method can produce substantial reductions in sample size requirements. The gain is most striking in the upper tail of the inverse function.

Comparisons made between the regression and MLE procedures are more difficult. The difficulty arises from the difference between the set of alternative distributions considered by each of the procedures. The MLE procedure treats sets of parametric families while the regression procedure treats a particular parametric family formed through F_o and β.

It is possible, however, to compare the two procedures by estimating the vector of coefficients, β, using the alternative methods. When r=0 the MLE and regression estimates are identical regardless of f_o and β_o. Thus the ARE(R,MLE)=1 whenever the true distribution function is equal to F_o raised to an arbitrary power. When r is greater than zero it becomes difficult to express f in terms of F_o and β and thus the MLE procedure is not easily applied. Work is being conducted in this direction, however, since the MLE estimates will most likely have the minimum obtainable variance and therefore make an excellent standard for comparison.

4. Demonstration

The regression procedures presented in this paper were developed during the study of bidding activity in the oil and gas lease program on the outer continental shelf. As part of this study a simulation model was developed that facilitates comparisons of bids tendered by a single firm (solo bids) and bids tendered jointly by several firms (joint bids).

The purpose of the simulation model was to generate a randomly drawn solo bid for each of several companies participating in joint bid. Repeating this process many times allowed one to estimate the distribution of the largest of the several generated solo bids. The distribution of the largest solo bid was then to be compared to actual joint bid to determine if the policy that has disallowed major oil firms bidding together has resulted in an increase in revenues from the oil and gas lease program.

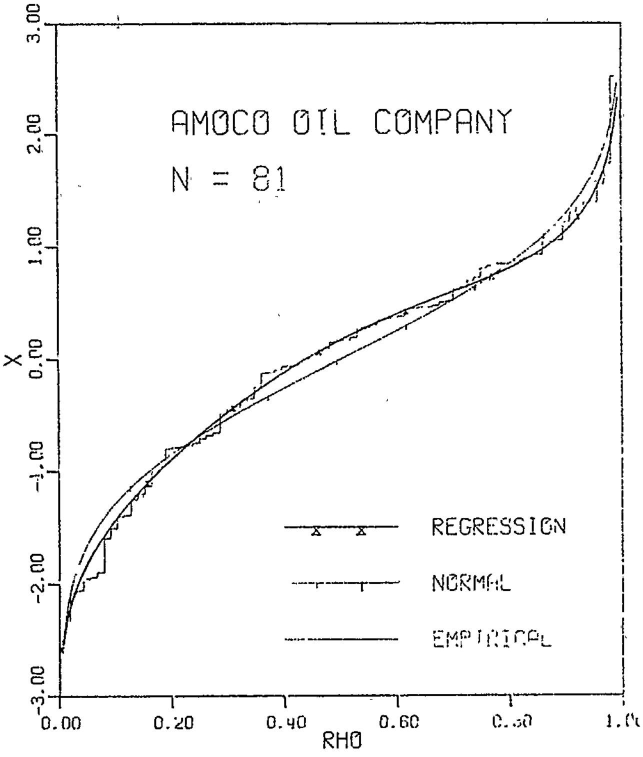
The model of individual bidding behavior for each firm was assumed to be of the form

$$\log\ (b_{ij}/\theta_j)=\alpha_j+\beta_j\theta_i+\varepsilon_{ij} \qquad (4.1)$$

Where b_{ij} is the bid of the jth firm on the ith lease, θ_i is a value measure for the ith lease, α_j and β_j are parameters to be estimated, and ε_{ij} is an error term with an unspecified distribution except for the condition E (ε_{ij})=0.

The model given in (4.1) was estimated by a two step process for each of many firms. The first step entailed the estimation of the parameters α and β using ordinary least squares. The residuals from the model were then used to estimate the probability distribution of ε_{ij} -- or more precisely the inverse of the distribution of the ε_{ij}.

The interest here is on the estimates of the inverse functions obtained using the procedures of the preceding section. One such estimate is presented in the figure. The inverse function was estimated using a fifth degree polynomial and normal distribution for $F_0(x)$. The normal distribution was chosen because of the suggestion in previous studies that the logs of bids follow a normal distribution function.



Figure

In the figure the empirical inverse function and normal inverse function are plotted with the inverse function obtained using the new procedure. The new procedure produces a smooth estimate that is closer to the empirical estimate than the normal estimate.

The ability to estimate the inverse of nearly any continuous distribution function and the ease of random deviate generation using the estimated inverse function make this procedure a useful tool for the building of discrete simulation models.

References

Hora, S.C. (1981), "Estimation of an Inverse Function for Random Deviate Generation", unpublished paper, Texas Tech University.

Kendall, M.G. and A. Stuart (1967), The Advanced Theory of Statistics, Hafner, New York.

Kennedy, W.J. and J.E. Gentle (1980), Statistical Computing, Marcel Dekker, New York.

Rao, C. R. (1973), Applied Linear Statistical Inference, John Wiley and Sons, New York.