NBER WORKING PAPER SERIES

MONTE CARLO TECHNIQUES IN STUDYING ROBUST ESTIMATORS

David C. Hoaglin*

Working Paper No. 16

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE
National Bureau of Economic Research, Inc.
575 Technology Square
Cambridge, Massachusetts 02139

November 1973

*Harvard University and NBER Computer Research Center. Research supported in part by National Science Foundation Grant GJ-1154X2 to the National Bureau of Economic Research, Inc.

Abstract

Recent work on robust estimation has led to many procedures which are easy to formulate and straightforward to program but difficult to study analytically. In such circumstances experimental sampling is quite attractive, but the variety and complexity of both estimators and sampling situations make effective Monte Carlo techniques essential. This discussion examines problems, techniques, and results and draws on examples in studies of robust location and robust regression.

Contents

Exhibits

The past several years have witnessed considerable research on robustness [12, 13], with the problems of point and interval estimation of symmetric location [2, 6] receiving a major share of attention. The results of these efforts are providing a basis for further work on more complicated problems such as robust regression [8], robust factorials and robust estimation of scale. In all these problems many of the procedures proposed and studied are relatively easy to formulate and generally straightforward to program for a digital computer, but they are quite difficult to study analytically. Even if we can get hold of their asymptotic behavior, as we often can in the symmetric location problem, their behavior in small samples is almost sure to be analytically intractable. This state of affairs will tend to drive us inexorably to experimental sampling, and as a consequence effective Monte Carlo techniques will take on considerable importance. This is clearly evident, for example, in the Princeton study of robust estimators of location [2]: the first phase considered 65 estimators in 30 sampling situations, and a later phase pursued simple linear combinations of pairs of estimators, five linear combinations per pair (adding some 10400 more "estimators"). Of course, if we already have Monte Carlo estimates of the variance of each estimator and of the covariance of each pair (as was the case in that phase of the Princeton study), it isn't necessary to start afresh with each linear combination. In any case, thirty situations make for a lot of computing, and we should pay close attention to accuracy and computational labor.

In what follows I will examine some of the Monte Carlo techniques which have been effective in studying point estimators of symmetric location and explore their generalizations in a study of regression procedures.

Building Blocks

Before I start on location and regression estimators, however, I should spend a little time on some of the essential building blocks of any experimental sampling process. They are easy to identify, and they are now reasonably well understood, but there still seem to be a few problems in making them as reliable and accessible as they should be.

The simplest and most important of these building blocks is the source of uniform pseudo-random numbers. Much has been written on the theory of uniform random number generators, and George Marsaglia discussed some interesting positive results [20] at last year's Interface, but practice still seems to lag a bit. As an example, one of our students recently checked at the Computer Center to see what generators were available. We use an IBM 370, and he found the three routines in the new SL-MATH [14]; he also found the two routines GGUl and GGU2 in the IMSL library [15]. Initially he had difficulty deciding which generator he should use, but the problem was easy to solve. The SL-MATH generators are versions of the one described by Lewis, Goodman, and Miller [17, 18] and their paper reports the results of extensive testing. The IMSL routines, on the other hand, are so sketchily documented that one cannot determine precisely what generator or generators they implement, and there is no reference to results of testing. The basic point which emerges here is hardly new, but apparently

it needs to be said once more. A user's minimum requirements for a uniform generator are that it be (1) of high quality, (2) extensively tested with published results, (3) fully and accurately documented, and (4) efficiently implemented. These amount to little more than "truth in packaging", but I'm afraid far too many of our consumers still settle for much less -- a black box.

Now that we have uniform random numbers, we must still turn them into a sample from the particular distribution we're using. This step in the process often consumes a major share of computer time; the important thing is to do it accurately and efficiently. See, for example, the books by Fishman [4] and Knuth [16] and the paper by Ahrens and Dieter [1]. For a bizarre example, see the paper by Neave [21].

If a number of distributions are involved, it may be possible to unify this part of the process by capitalizing on common features. The Princeton study provides two related examples. First, all the distributions were represented in the form

Gaussian/independent,

the ratio of a standard Gaussian numerator to an independent denominator. As Exhibit 1 shows, this class of distributions [23] is quite broad. One member, Gauss/uniform, known more briefly as the "slash" distribution, is an alternative to the Cauchy having Cauchy tails and Gaussian center. Second, since contaminated Gaussian mixtures belong to this class, the efficient way to handle them is in the form of contamponents, which take a fixed number of observations from the contaminating distribution instead of the varying number determined by the mixture probability. Then we can

Exhibit 1

Some Distributions Represented as Gauss/independent

prototype: $Y = Z/V$ $\quad Z \sim \text{Gau}(0,1)$, V independent

denominator (V)	distribution of Y
$V \equiv 1$	Gaussian (0,1)
$P\{V = 1/k\} = \alpha$ $P\{V = 1\} = 1 - \alpha$	contaminated Gaussian $100\alpha\%$ of Gau $(0,k^2)$
$\sqrt{\chi^2_n/n}$	t_n
half-normal	Cauchy
uniform [0,1]	slash
$V = 1/\sqrt{-2\times\ln(U)}$ $U \sim$ uniform [0,1]	double exponential

combine the contamponent results, using binomial weights, to get the result for any desired mixture.

In the location problem many estimators are calculated from the ordered sample, so we require a procedure for generating the necessary order statistics. For complete samples of any reasonable size the efficient way is to sort; we just need to be sure we use an efficient sorting algorithm, whose labor will be proportional to $n \times \log(n)$. If the procedures we are studying depend on only a few order statistics at one end of a sample, the approaches of Lurie and Hartley [19] or Schucany [24] may be better. These are not designed for complete samples. For example, on an IBM 360 Model 65 the first algorithm of Lurie and Hartley is slower than sorting for complete samples smaller than about one million.

These are by no means all the useful building blocks, but handling these operations efficiently will provide a solid basis for any empirical sampling study. These components should be part of any reasonable statistically oriented subroutine library.

Estimators and Invariance

In both the location problem and the regression problem all the estimators under study will share some basic invariance properties which the Monte Carlo techniques may be able to exploit. Any reasonable location estimator $T(\underline{y})$ should be either location-invariant,

$$T(\underline{y} + b\underline{1}) = T(\underline{y}) + b,$$

or location-scale-invariant,

$$T(a\underline{y} + b\underline{1}) = aT(\underline{y}) + b,$$

depending on the context in which it is used. Usually we would demand location-scale invariance.

In the regression problem

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

one can require more invariance properties of an estimator $\underline{B}(\underline{y})$, such as invariance under multiplication of the data by a scalar, or under non-singular linear transformation of the carriers (the columns of X). The important ones seem to be regression invariance,

$$\underline{B}(\underline{y} + X\underline{\gamma}) = \underline{B}(\underline{y}) + \underline{\gamma} \ ,$$

and regression-and-scale invariance,

$$B(a\underline{y} + X\underline{\gamma}) = aB(\underline{y}) + \underline{\gamma} \ .$$

The first of these may be described less formally as transparency -- if the data is perturbed by an exactly fittable change, all the change goes into the fit. This condition is stronger in the regression problem than the corresponding condition in the location problem: such procedures as step-wise regression and ridge regression [9, 11] don't satisfy the transparency condition. If good Monte Carlo techniques are available, however, we may want to restrict our attention to such procedures, at least for the present.

The Location Problem

Once we have agreed to study only invariant estimators of location, the natural thing to do (for each finite sample size) is to find the best invariant estimator (in the sense of having the smallest mean squared error). This estimator is the Pitman estimator, and the primary object in studying it is to determine its variance in a variety of sampling situations so that we can use it as a standard in assessing the performance of other, more robust, invariant estimators.

Each sample of n "belongs to" a configuration, the set of all samples to which it is related by changes in only location and scale:

$$\underline{c}(\underline{y}) = \{a\underline{y} + b\underline{1} \mid a>0,\ -\infty < b < +\infty\} .$$

We usually specify the configuration $\underline{c}$ by giving a standard member of the set, defined in terms of a particular location statistic $\hat{y}$ and a particular scale statistic $\hat{s}$:

$$\underline{c} = (\underline{y} - \hat{y}\underline{1})/\hat{s},\quad \underline{y} = \hat{y}\underline{1} + \hat{s}\underline{c}$$

Thus $$T(\underline{y}) = \hat{y} + \hat{s}T(\underline{c}) .$$

The behavior of an invariant estimator is essentially determined by what it does for each configuration, and the Pitman estimator may be derived by minimizing mean squared error, configuration by configuration. For convenience we take the true location parameter to be zero and the true scale parameter to be one. Then it is a simple matter to find $T_0(\underline{c})$, the value of the Pitman estimator for configuration $\underline{c}$, by differentiating

$$mse(T \mid \underline{c}) = E\{[\hat{y} + \hat{s}T(\underline{c})]^2 \mid \underline{c}\}$$

with respect to T:

$$T_0(\underline{c}) = -E\{\hat{y}\hat{s} \mid \underline{c}\}/E\{\hat{s}^2 \mid \underline{c}\}$$

It follows that

$$mse_0(\underline{c}) \equiv mse(T_0 \mid \underline{c}) = E\{\hat{s}\hat{y} \mid \underline{c}\}\ T_0(\underline{c}) + E\{\hat{y}^2 \mid \underline{c}\}$$

and

$$mse(T \mid \underline{c}) = mse_0(\underline{c}) + E\{\hat{s}^2 \mid \underline{c}\}\ [T(\underline{c}) - T_0(\underline{c})]^2 .$$

Now we should ask where Monte Carlo enters. What we want is the variance of T_0. To get it, or more precisely, to estimate it, we obtain the configurations by simple experimental sampling. We then calculate $T_0(\underline{c})$ and $mse_0(\underline{c})$ by numerical quadrature. We also calculate $E\{\hat{s}^2 \mid \underline{c}\}$, because we'll be interested in the performance of other invariant

estimators [7]. We can estimate $\mathrm{var}(T_0)$ by simply averaging the values of $\mathrm{mse}_0(\underline{c})$ over the sample of configurations, but in general we can do substantially better by carrying along some other estimators whose variances we know, such as linear combinations of order statistics. This will put us in a position to estimate $\mathrm{var}(T_0)$ more accurately by regression estimation [3], taking advantage of the correlation between $\mathrm{mse}_0(\underline{c})$ and $\mathrm{mse}(T|\underline{c})$ for some appropriate estimator T (such as the BLUE, whose variance can be calculated from order-statistic covariances). Using the values of $\mathrm{mse}_0(\underline{c})$, we fit the regression line

$$\mathrm{mse}_0(\underline{c}) \sim a + b\ \mathrm{mse}(T|\underline{c})$$

and estimate $\mathrm{var}(T_0)$ by

$$\overline{\mathrm{mse}_0} + b[\mathrm{var}(T) - \overline{\mathrm{mse}}(T)]$$

(the bars indicate averages over the sampled configurations, and var(T) is known). In principle we could use several estimators as carriers in this regression, but as a practical matter one will usually suffice, and for some sampling situations there may not be many estimators whose variance we know.

Taking a larger view for a moment, we should consider using regression estimation in a wide variety of problems. Not only should we plan to provide a basis for calibrating the sampling results, but the ability to improve the accuracy of Monte Carlo estimates in this fashion will also be broadly useful.

Now let's turn to another set of techniques applied in studying location estimators. As I mentioned earlier, the sampling situations in the Princeton study involve distributions which can be represented as Gaussian/independent: $(y_1, \ldots, y_n) = (z_1/v_1, \ldots, z_n/v_n)$. For invariant

estimators this leads to a neat swindle and some quite efficient Monte Carlo. The basic observation is that we can take the sample of v's (the denominators) and then, by conditioning on these v's, take advantage of the Gaussian distributions of the z's. Thus we use

$$\hat{y} = (\sum_{i=1}^{n} v_i^2 y_i)/(\sum_{i=1}^{n} v_i^2)$$

and

$$\hat{s}^2 = (n-1)^{-1} \sum_{i=1}^{n} v_i^2 (y_i - \hat{y})^2 .$$

Now $\hat{y}$ and $\hat{s}$ are conditionally independent so that

$$E\{T^2(\underline{y})|\underline{c},\underline{v}\} = T^2(\underline{c}) + (\sum_{i=1}^{n} v_i^2)^{-1} .$$

One version of this approach has been described by Relles [22] for Student's t distributions, and W. H. Rogers independently rediscovered it and extended it for the Princeton study. The gains in efficiency, that is, the reductions in sampling variance, come from two sources: the conditional independence of $\hat{y}$ and $\hat{s}$, given $\underline{v}$, and the ability to evaluate one term of the conditional mean squared error analytically. Since all definitions of configuration are equivalent, we see in this case the benefit of choosing a convenient one.

In preparation for a later generalization to the regression problem we can pursue the matter of configurations a bit further, defining

> outer configuration: the y_i are fixed, up to location and scale, but neither the z_i nor the v_i are individually fixed; and
>
> inner configuration: the y_i are fixed, up to location and scale, and the v_i are fixed.

If we wanted more detailed information for each outer configuration, we could (in principle) sample the inner configurations. Often, as in the Princeton study, we have no particular interest in this information, and

we take only a single inner configuration for each outer configuration. This keeps the experimental sampling simple: we generate the sample of z's and the sample of v's, determine the outer configuration from the resulting $\underline{y}$, and condition on the v's. In this setting the formula for the conditional expectation of the product of two estimators is a simple generalization of the conditional mean square error formula above.

In studying robust location estimators there is no reason to put all the emphasis on variances -- percentage points deserve attention too. For percentage points the exact conditional calculations involve the non-central t distribution but are only slightly more complicated than the ones for covariances. Having derived these Monte Carlo techniques, we should naturally ask what they buy us; the reductions in sampling variance are quite encouraging, both for variances of estimators and for percentage points of estimators. Exhibit 2 has a few of the efficiencies (in the log scale) for the 2.5% point [5]. The estimators are the mean (M) and four trimmed means (the % value is trimmed from each end -- the limiting case of 50% is the median). Sampling situations are Gaussian at n=5, 10, 20, and 40, three contamponents ("10%3G n=20" means that exactly 2 of the 20 values are Gaussian with mean 0 and scale 3), and Cauchy at n=20. In this data and in other more extensive data a general pattern emerges: the Monte Carlo does better for more robust estimators and for distributions closer to Gaussian. Being able to reduce sampling variance by a factor of 10000 is a nice gain indeed, but a factor of 100 or even 5 is not to be overlooked.

This look at the efficiency of the Monte Carlo is one facet of the sort of analysis one should do on the results. When we have as much

Exhibit 2

Efficiency of Monte Carlo
in Estimating 2.5% Point of Estimators

(entries are $\log_{10}$(efficiency))

	estimator				
	M	5%	10%	25%	50%
G n=5	∞	4.4	3.9	2.9	2.1
G n=10	∞	4.2	3.6	2.7	2.1
G n=20	∞	3.7	3.7	2.5	1.9
G n=40	∞	3.9	3.3	2.7	1.9
10%3G n=10	1.9	2.5	2.9	2.5	2.0
10%3G n=20	1.9	2.3	2.7	2.2	1.9
10%10G n=20	1.1	1.2	2.5	2.3	1.9
Cauchy n=20	0.6	0.7	0.9	1.1	1.1

Source: A. M. Gross [5]

structure as in the Princeton Study, there are many other ways to approach the results. The book devotes a long and imaginative chapter to such matters, so I will only remark that several groups of the estimators were actually members of one-parameter families, and much has been learned from studying their behavior as families.

Regression Problems

When we venture into studying candidates for robust estimators in the linear regression problem, we face many more difficult design and Monte Carlo problems. For example, in the model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon} ,$$

where $\underline{y}$ and $\underline{\varepsilon}$ are $n\times 1$, X is $n\times p$, and $\underline{\beta}$ is $p\times 1$, we must choose values of n and p, give careful attention to the matrix X, and select a class of disturbance distributions for $\underline{\varepsilon}$. On this last point we can hope to use the same sort of swindle which proved so effective in the location problem [10]. As we shall see, the matter of configuration becomes somewhat more complicated: we now have

> outer configuration: the y_i are fixed, up to regression and scale, but neither the z_i nor the v_i are individually fixed;
>
> middle configuration: the y_i are fixed, up to regression and scale, and the set $\{v_i\}$ is fixed; and
>
> inner configuration: the y_i are fixed, up to regression and scale, and the v_i (including their permutation) are fixed.

The need here for a third level of configuration arises because we can no longer freely permute the denominators. A tendency for more variable errors to arise at a sensitive place in the design will affect the covariance matrix of an estimator of $\underline{\beta}$, and we will need to have control

of this so that we can investigate it.

To derive the swindle for the regression problem with disturbances $\varepsilon_i = z_i/v_i$, we let $w_i = v_i^2$ and $\langle \underline{w} \rangle = \text{diag}\ (w_1, \ldots, w_n)$. Then the outer configurations are based on

$$\underline{b}_* = (X^T \langle \underline{w} \rangle X)^{-1} X^T \langle \underline{w} \rangle \underline{y}$$

and, letting $\underline{y}_* = X\underline{b}_*$,

$$s_*^2 = (n - p)^{-1} \sum_{i=1}^{n} w_i (y_i - y_{*i})^2$$

In estimating the covariance matrix of a regression estimator $\underline{B}(\underline{y})$ the details parallel those in the location problem, but the gains don't seem to be nearly so dramatic. D. F. Andrews reports factors of about 2 to 10 for a small study at p=3. Our experience at the National Bureau of Economic Research has largely involved p=6, with roughly the same results.

At NBER we have recently begun a substantial Monte Carlo study of robust regression procedures [8]. To get started, one has to restrict his scope quite severely and plan to proceed in stages. As a result, spurred by suggestions from J. W. Tukey [25], we have concentrated on n=20, p=6, essentially two carefully structured X matrices, and a rather restrictive class of estimators. Our set of disturbance distributions is small, selected in part on the basis of the Princeton study. We expect to report further details and at least preliminary results in a month or two.

Concluding Remarks

In concluding I'd like to look at how the techniques I've discussed fit into the broader framework of Monte Carlo. We would like to have at our command a number of general techniques, broadly applicable and programmed ready for use. Regression estimation is one reasonable possibility; I

think it could profitably be used more often. Overall, however, I am somewhat skeptical about finding many general techniques which provide great gains in efficiency. It seems more likely that general techniques will offer only very limited gains and that the real improvements will continue to come from working hard to exploit specific features of the particular problem or of a class of closely similar problems. The needs for calibration and analysis of results will continue to demand careful attention.

References

[1] J. H. Ahrens and U. Dieter, Computer Methods for Sampling from the Exponential and Normal Distributions. Communications of the ACM 15 (1972), 873-882.

[2] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, Robust Estimates of Location: Survey and Advances. Princeton University Press, 1972.

[3] W. G. Cochran, Sampling Techniques, second edition. Wiley, 1963.

[4] G. S. Fishman, Concepts and Methods in Discrete Event Digital Simulation. Wiley-Interscience, 1973.

[5] A. M. Gross, Formulae for a Monte Carlo Swindle for Estimators of Location. Technical Report 14, Series 2, Department of Statistics, Princeton University, May 1972.

[6] A. M. Gross, A Robust Confidence Interval for Location for Symmetric, Long-Tailed Distributions. Proceedings of the National Academy of Sciences USA 70, 7 (July 1973), 1995-1997.

[7] D. C. Hoaglin, Optimal Invariant Estimation of Location for Three Distributions and the Invariant Efficiencies of Some Other Estimators. Ph.D. Dissertation, Department of Statistics, Princeton University, 1971.

[8] D. C. Hoaglin, P. W. Holland, and R. E. Welsch, Design for a Study of Robust Regression Procedures. Working Paper, Computer Research Center for Economics and Management Science, National Bureau of Economic Research, 1973.

[9] A. E. Hoerl and R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 12 (1970), 55-67.

[10] P. W. Holland, Monte Carlo for Robust Regression: The Swindle Unmasked. Working Paper 10, Computer Research Center for Economics and Management Science, National Bureau of Economic Research, September 1973.

[11] P. W. Holland, Weighted Ridge Regression: Combining Ridge and Robust Regression Methods. Working Paper 11, Computer Research Center for Economics and Management Science, National Bureau of Economic Research, September 1973.

[12] Peter J. Huber, Robust Statistics: A Review. Annals of Mathematical Statistics 43 (1972), 1041-1067.

[13] Peter J. Huber, Robust Regression: Asymptotics, Conjectures, and Monte Carlo. Annals of Statistics 1 (1973), 799-821.

[14] IBM Corporation, Subroutine Library -- Mathematics, User's Guide (SH12-5300-0). November 1971.

[15] International Mathematical and Statistical Libraries, Inc., The IMSL Library 1 Reference Manual, third edition (FORTRAN IV, S/370-360), July 1973.

[16] D. E. Knuth, The Art of Computer Programming, Volume 2: Semi-numerical Algorithms. Addison-Wesley, 1969.

[17] G. P. Learmonth and P. A. W. Lewis, Naval Postgraduate School Random Number Generator Package LLRANDOM. Naval Postgraduate School, June 1973.

[18] P. A. W. Lewis, A. S. Goodman, and J. M. Miller, A Pseudo-random Number Generator for the System/360. IBM Systems Journal 8 (1969), 136-146.

[19] D. Lurie and H. O. Hartley, Machine-Generation of Order Statistics for Monte Carlo Computations. The American Statistician 26, 1 (February 1972), 26-27.

[20] G. Marsaglia, The Structure of Linear Congruential Sequences. Applications of Number Theory to Numerical Analysis (S. K. Zaremba, editor), 249-285. Academic Press, 1972.

[21] H. R. Neave, On using the Box-Muller Transformation with Multiplicative Congruential Pseudo-random Number Generators. Applied Statistics 22 (1973), 92-97.

[22] D. A. Relles, Variance Reduction Techniques for Monte Carlo Sampling from Student Distributions. Technometrics 12 (1970), 499-515.

[23] W. H. Rogers and J. W. Tukey, Understanding some Long-tailed Symmetrical Distributions. Statistica Neerlandica 26 (1972), 211-226.

[24] W. R. Schucany, Order Statistics in Simulation. Journal of Statistical Computation and Simulation 1 (1972), 281-286.

[25] J. W. Tukey, A Way Forward for Robust Regression. Unpublished memorandum, Bell Laboratories (Murray Hill), July 1973.