NBER WORKING PAPER SERIES

MONTE CARLO FOR ROBUST REGRESSION: THE SWINDLE UNMASKED

Paul W. Holland*

Working Paper No. 10

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

September 1973

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

Abstract

Gives an alternative derivation of a Monte Carlo method that has been used to study robust estimators. Extensions of the technique to the regression case are also considered and some computational points are briefly mentioned.

1. Introduction

In this paper, we discuss a method for achieving more accuracy from a Monte Carlo study than is possible from simple random sampling. Such Monte Carlo "swindles" are important in the large scale use of Monte Carlo studies. The particular method we discuss here has been described before in Relles [1970] and Andrews *et al.* [1972], but their approaches are somewhat different from the one we employ. A deeper understanding of the method and its properties is gained by having alternative derivations available.

The particular problem we consider is the following. We begin with the familiar linear regression problem

$$y = X\beta + e \tag{1-1}$$

where y is N×1, X is N×p, β is p×1 and e is N×1. We furthermore assume that the components of e, e_i, are independent and identically distributed random variables with common density

$$\frac{1}{\sigma} f(\frac{e}{\sigma}) \tag{1-2}$$

where f is assumed to be symmetric about 0, i.e., $f(-x) = f(x)$. The linear regression problem is to estimate β using y and X. Let $\hat{\beta}$ denote a generic estimator of β. Sometimes its dependence on y will be denoted by $\hat{\beta}(y)$.

There are two notions of invariance that will be important in the rest of this paper.

Scale Invariance

An estimator $\hat{\beta}(y)$ is said to be scale invariant if

$$\hat{\beta}(c\ y) = c\ \hat{\beta}(y) \tag{1-3}$$

for any constant, c.

Regression Invariance

An estimator $\hat{\beta}(y)$ is said to be regression (on X) invariant if

$$\hat{\beta}(y + X\gamma) = \hat{\beta}(y) + \gamma \tag{1-4}$$

for any p×1 vector, γ.

We shall restrict our discussion to estimators, $\hat{\beta}$, which are both regression and scale invariant. The problem of main concern is to study

$$\mathrm{Cov}_{\beta}(\hat{\beta}) = E_{\beta}(\hat{\beta} - \beta)(\hat{\beta} - \beta)^{T} \tag{1-5}$$

However, because we have restricted attention to regression and scale invariant estimators we have

$$E_{\beta}(\hat{\beta}(y) - \beta)(\hat{\beta}(y - \beta))^{T} = \sigma^2 E_{\beta}(\hat{\beta}(\tfrac{y - X\beta}{\sigma})\hat{\beta}(\tfrac{y - X\beta}{\sigma})^{T}) = \sigma^2 \mathrm{Cov}_{0}(\hat{\beta}) \tag{1-6}$$

Thus we may assume without loss of generality that $\beta = 0$ and $\sigma^2 = 1$.

In order to compute $\mathrm{Cov}_{0}(\hat{\beta})$, we must often resort to a Monte Carlo study. The swindle we will consider is designed for such an investigation. When $p = 1$ and $x_i \equiv 1$, the regression problem reduces to the "location" problem in which we are estimating the center of a symmetric distribution. In the location case (1-5) reduces to the variance of $\hat{\beta}$.

We shall divide our discussion of the swindle into four cases: Location with Gaussian errors, Location with Gauss/Independent errors, Regression with Gaussian errors and Regression with Gauss/Independent errors.

2. Location with Gaussian Errors

In this case we are concerned with computing

$$\mathrm{Var}_0(\hat{\beta}) = E_0(\hat{\beta}^2) \tag{2-1}$$

for a general location and scale invariant estimator under the assumption that f is the unit Gaussian density. However, when f is the Gaussian distribution, we know what the best location (regression) and scale invariant estimator is -- our friend $\bar{y}$, the sample average. Hence, instead of computing the variance of $\hat{\beta}$, we shall try to compute the excess of the variance of $\hat{\beta}$ over that of $\bar{y}$. We now derive the important formula that allows us to do this.

Theorem 1: If $\hat{\beta}$ is any location and scale invariant estimator, then under unit Gaussian errors we have

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\bar{y}) + E(\hat{\beta} - \bar{y})^2 \tag{2-2}$$

$$= \mathrm{Var}(\bar{y}) + E\left(\frac{(\hat{\beta} - \bar{y})^2}{S^2}\right) \tag{2-3}$$

where S^2 is the usual unbiased estimate of σ^2.

Proof: We begin with (2-2) and then derive (2-3) from it. We have

$$(\hat{\beta}(y))^2 = (\hat{\beta}(y - \bar{y}) + \bar{y})^2 = (\hat{\beta}(y - \bar{y}))^2 + 2\,\hat{\beta}(y - \bar{y})\bar{y} + (\bar{y})^2 \tag{2-4}$$

Hence

$$\mathrm{Var}(\hat{\beta}) = E_0(\hat{\beta}^2) = E_0(\bar{y})^2 + 2\,E_0(\bar{y}\,\hat{\beta}(y - \bar{y})) + E_0(\hat{\beta}(y - \bar{y}))^2 \tag{2-5}$$

But $\bar{y}$ and $y - \bar{y}$ are independent so that $\bar{y}$ and $\hat{\beta}(y - \bar{y})$ are independent and hence

$$E_0(\bar{y}\,\hat{\beta}(y - \bar{y})) = 0$$

Also $E_0(\bar{y})^2 = \mathrm{Var}(\bar{y})$ and $\hat{\beta}(y - \bar{y}) = \hat{\beta} - \bar{y}$ so that $\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\bar{y}) + E(\hat{\beta} - \bar{y})^2$ which was to be proved. To prove (2-3) we need to show that

$$E(\hat{\beta} - \bar{y})^2 = E(\frac{(\hat{\beta} - \bar{y})^2}{S^2}) \tag{2-6}$$

This is done as follows

$$E(\hat{\beta} - \bar{y})^2 = E(S^2(\hat{\beta}(\frac{y - \bar{y}}{S}))^2) \tag{2-7}$$

But S is independent of $\frac{y - \bar{y}}{S}$ so that S^2 is independent of $\hat{\beta}(\frac{y - \bar{y}}{S})$ and hence (2-7) equals

$$E(S^2)\ E(\hat{\beta}(\frac{y - \bar{y}}{S})^2) = \sigma^2\ E(\frac{(\hat{\beta} - \bar{y})^2}{S^2}) \tag{2-8}$$

Since we have assumed $\sigma^2 = 1$, (2-6) follows immediately. QED

It should be noted that (2-2) only requires $\hat{\beta}$ to be location invariant while (2-3) requires both location and scale invariance.

It is quite simple to use Theorem 1 to get a Monte Carlo swindle for the variance of a location and scale invariant estimator, $\hat{\beta}$. If we were going to use naive random sampling to estimate $\mathrm{Var}(\hat{\beta})$ we would draw repeated independent samples of size N, $y = (y_1, \ldots, y_N)^T$, compute $\hat{\beta}(y)$ for each sample and average the value of $\hat{\beta}^2$ over the replications. But from (2-2) we see that we may also estimate $\mathrm{Var}(\hat{\beta})$ by computing $\hat{\beta}(y)$ and $\bar{y}$ from each sample, averaging the value of $(\hat{\beta} - \bar{y})^2$ over the replications and then adding $N^{-1} = \mathrm{Var}(\bar{y})$ to the result, i.e.,

$$\text{estimated } \mathrm{Var}(\hat{\beta}) = N^{-1} + \text{Monte Carlo Average } (\hat{\beta} - \bar{y})^2. \tag{2-9}$$

In order to remove the first order effect of N when looking at different sample sizes it is customary to estimate $N\ \mathrm{Var}(\hat{\beta})$ instead. This leads to

$$\text{estimated } N\,(\mathrm{Var}(\hat{\beta})) = 1 + N(\text{Monte Carlo Average } (\hat{\beta} - \bar{y})^2) \tag{2-10}$$

What is the nature of this swindle? The main point is that it concentrates the slow Monte Carlo convergence on the excess of the $\mathrm{Var}(\hat{\beta})$ over $\mathrm{Var}(\bar{y})$ rather than allowing this to effect all of $\mathrm{Var}(\hat{\beta})$. This implies that if $\hat{\beta}$ is a very efficient estimator (has a small excess of variance) then the swindle will be more effective than if $\hat{\beta}$ has a large excess. This is because the smaller the excess, the smaller the percentage of the $\mathrm{Var}(\hat{\beta})$ that is estimated by Monte Carlo averaging. From (2-4) and (2-5) we see that the swindle utilizes the theory of the Gaussian distribution to get exact results for two of three pieces of $E(\hat{\beta}^2)$; these are

$$E(\bar{y}\,\hat{\beta}(y - \bar{y})) = 0 \tag{2-11}$$

$$\text{and} \quad E(\bar{y})^2 = N^{-1} . \tag{2-12}$$

Monte Carlo is then used to estimate the third piece.

We may use (2-3) to get more of a swindle via the estimate:

$$\text{estimated } N(\mathrm{Var}(\hat{\beta})) = 1 + N(\text{Monte Carlo Average}(\tfrac{1}{S^2}(\hat{\beta} - \bar{y})^2)) \tag{2-13}$$

However, if N is at all appreciable, $1/S^2$ will not differ much from unity so that this swindle should not significantly improve upon the earlier one unless the sample size, N, is quite small. This agrees with the folklore.

3. Location with Gauss/Independent Errors

Critical to the swindle in the Gaussian case was the ability to evaluate (2-11) and (2-12) exactly. These calculations lean heavily on properties of $\bar{y}$ and S^2 in the Gaussian distribution. It is not clear how to successfully generalize this to arbitrary symmetric unimodal densities, f. However, the class of distributions given by

$$y_i = u_i/v_i \tag{3-1}$$

where u_i is unit Gaussian and v_i is a positive random variable independent of u_i, is such that an expression analogous to (2-11) can be evaluated exactly and one analogous to (2-12) can be evaluated exactly in some cases and partially evaluated in all cases. This leads to a swindle that is not as effective as the one for the Gaussian case, but which is better than simple random sampling. The family of densities associated with (3-1) is a generalization of the t-family and contains such members as: Cauchy, t, double exponential, logistic and scale mixtures of Gaussian densities. Conditionally, given v_i, y_i is Gaussian with mean zero and variance v_i^{-2}. We may regard y_i as Gaussian with a random scale. Andrews and Mallows [1973] give conditions under which a density has the representation (3-1).

Let $v = (v_1, \ldots, v_N)$; then the key idea is that given v, we are back in much the same situation as we were in the pure Gaussian case. The only real differences are (1) now the variances are unequal and (2) we must eventually integrate over the density of v. We let

$$\bar{y} = \bar{y}(v) = \frac{\sum_i v_i^2 y_i}{\sum_i v_i^2} \tag{3-2}$$

and

$$S^2 = S^2(v) = \frac{1}{N-1} \sum_i v_i^2 (y_i - \bar{y}(v))^2. \tag{3-3}$$

Note that $\bar{y}(v)$ and $S^2(v)$ can't be computed in real data since v is not an observable but in a Monte Carlo study in which v is generated along with $u = (u_1, \ldots, u_N)^T$ to produce y, v will be available.

Now instead of knowing the best location-scale invariant estimator of β for the error distribution given by (3-1), we know an even better estimator, $\bar{y}(v)$. It is better than the best location-scale invariant estimator because it uses unobservable information. Thus we will try to compute the excess variances of $\hat{\beta}$ over that of $\bar{y}(v)$. The formula for this is given in the next theorem.

Theorem 2: If $\hat{\beta}$ is any location and scale invariant estimator, then if the errors are given by (3-1) we have

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\bar{y}(v)) + E(\hat{\beta} - \bar{y}(v))^2 \tag{3-4}$$

$$= \mathrm{Var}(\bar{y}(v)) + E\left(\frac{(\hat{\beta} - \bar{y}(v))^2}{S^2(v)}\right) \tag{3-5}$$

Proof:

We first show (3-4) and then derive (3-5) from it. Given v we may compute $\bar{y} = \bar{y}(v)$ so that we have

$$(\hat{\beta}(y))^2 = (\hat{\beta}(y - \bar{y}))^2 + 2\,\hat{\beta}(y - \bar{y})y + (\bar{y})^2$$

and hence taking conditional expectations we get

$$E[(\hat{\beta}(y))^2|v] = E[(\hat{\beta}(y - \bar{y}))^2|v] + 2\,E[\bar{y}\,\hat{\beta}(y - \bar{y})|v] + E[(\bar{y})^2|v]$$

However given v, $\bar{y}$ and $y - \bar{y}$ are independent so the middle term vanishes. Then taking expectations over v we get (3-4). To prove (3-5) we need to show that

$$E(\hat{\beta} - y(v))^2 = E\left(\frac{(\hat{\beta} - y(v))^2}{S^2(v)}\right)$$

We have

$$E[(\hat{\beta} - y(v))^2|v] = E[S^2(v)\hat{\beta}(\frac{y - \bar{y}(v)}{S(v)})^2|v] \tag{3-6}$$

but given v, S(v) is independent of $(y - \bar{y}(v))/S(v)$ so that $S^2(v)$ is independent of $\hat{\beta}(\frac{y - \bar{y}(v)}{S(v)})$. Hence (3-6) equals

$$E[S^2(v)|v]\, E[\hat{\beta}(\frac{y - \bar{y}(v)}{S(v)})^2|v]$$

$$= E[\frac{(\hat{\beta} - \bar{y}(v))^2}{S^2(v)}|v]$$

Taking expectations over v proves the result. QED

To use Theorem 2 to get a swindle we need a little more work, namely we need to be able to compute $\mathrm{Var}(\bar{y}(v))$. In general this is difficult, but the following result helps a little.

Theorem 3: $\mathrm{Var}(\bar{y}(v)) = E(\frac{1}{\sum_i v_i^2})$. (3-7)

Proof: Condition on v. QED

Depending on what the distribution of V_i is, the simplification implied by Theorem 3 may or may not lead to an exact solution. When qv_i^2 has a chi-square distribution with q degrees of freedom, then we may obtain an exact result for $\mathrm{Var}(\bar{y}(v))$. It is:

$$E(\frac{1}{\sum_i v_i^2}) = \frac{q}{Nq - 2} \tag{3-8}$$

The swindle now comes in two forms depending on whether or not we have an exact formula for $\mathrm{Var}(\bar{y}(v))$. As before we give it for estimating $N\,\mathrm{Var}(\hat{\beta})$. If $\mathrm{Var}(\bar{y}(v))$ is known then use:

$$\text{Estimated } N\,(\mathrm{Var}(\hat{\beta})) = N\,\mathrm{Var}\,\bar{y}(v) + N(\text{Monte Carlo Average}(\hat{\beta} - \bar{y}(v))^2) \tag{3-9}$$

If $\mathrm{Var}(\bar{y}(v))$ cannot be found exactly then use

$$\text{Estimated N } (\mathrm{Var}(\hat{\beta})) = N(\text{Monte Carlo Average}(\frac{1}{\Sigma\ v_i^2})) + N(\text{Monte Carlo Average}(\hat{\beta} - \bar{y}(v))^2) \qquad (3\text{-}10)$$

As before, some extra swindle may be gained from using (3-5) rather than (3-4), but unless N is small the gains are not likely to be appreciable.

In this case the swindle has two things going against it. Most obviously, if $E(\frac{1}{\Sigma\ u_i^2})$ can't be computed exactly, and must be estimated by simple random sampling then not only are we using Monte Carlo to estimate the excess variance, we are also using it to estimate a portion of $\mathrm{Var}(\bar{y}(v))$. Secondly, even if $\mathrm{Var}(\bar{y}(v))$ can be computed exactly, it may not be a very large portion of the total variance of $\tilde{\beta}$. This is because $\bar{y}(v)$ is a better estimator than any location and scale invariant estimator, i.e.,

$$\mathrm{Var}(\bar{y}(v)) \leq \mathrm{Var}(\hat{\beta}) \qquad (3\text{-}11)$$

for any such estimator $\hat{\beta}$. Thus relative to any given location-scale invariant estimator $\mathrm{Var}(\bar{y}(\hat{\beta}))$ may be very small, even relative to the best such estimator. Because of these problems, the swindle should not be as effective here as it is for the Gaussian case.

4. The Regression Case

With the preparation given in the previous two sections we may move easily to the regression case. We first treat the case of Gaussian errors and then Gauss/Independent errors. The theorems are stated without proof since they are completely analogous to the corresponding ones for the location case.

Gaussian Errors

If the errors are Gaussian, then we have the following basic result.

Theorem 4: If $\hat{\beta}$ is any regression and scale invariant estimator and $\hat{\beta}_{LS}$ the usual least squares estimator, then

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}(\hat{\beta}_{LS}) + E((\hat{\beta} - \hat{\beta}_{LS})(\hat{\beta} - \hat{\beta}_{LS})^T) \tag{4-1}$$

$$= \mathrm{Cov}(\hat{\beta}_{LS}) + E((\frac{1}{S^2})(\hat{\beta} - \hat{\beta}_{LS})(\hat{\beta} - \hat{\beta}_{LS})^T) \tag{4-2}$$

where S^2 is the usual unbiased estimate of σ^2 based on the least squares residual mean-square.

As was true for Theorem 1, (4-1) only requires regression invariance, while (4-2) requires both regression and scale invariance.

Since $\mathrm{Cov}(\hat{\beta}_{LS})$ is given exactly by

$$\mathrm{Cov}(\hat{\beta}_{LS}) = (X^T X)^{-1} \tag{4-3}$$

we are led to the following Monte Carlo swindle formula.

$$\text{Estimated } \mathrm{Cov}(\hat{\beta}) = (X^T X)^{-1} + \text{Monte Carlo Average}((\hat{\beta} - \hat{\beta}_{LS})(\hat{\beta} - \hat{\beta}_{LS})^T) \tag{4-4}$$

As before when N is small (actually when N-p is small) there may be some additional advantage to basing the swindle on (4-2) rather than (4-1), but otherwise the improvement over (4-4) is not likely to be noticeable.

Gauss/Independent Errors

When the errors have the structure given by (3-1), we may define, for each v, these quantities:

$$\hat{\beta}(v) = (X^T\langle v^2\rangle X)^{-1}X^T\langle v^2\rangle y \tag{4-5}$$

$$\hat{y}(v) = X\hat{\beta}(v) \tag{4-6}$$

and

$$S^2(v) = (N-p)^{-1}\sum_i (y_i - \hat{y}_i(v))^2 \quad . \tag{4-7}$$

Where $\langle v^2\rangle$ is the diagonal matrix based on $v^2 = (v_1^2, \ldots, v_N^2)^T$. Then we have the following theorem.

Theorem 5: If $\hat{\beta}$ is any regression and scale invariant estimator and the errors are given by (3-1) then

$$\mathrm{Cov}(\hat{\beta}) = \mathrm{Cov}(\hat{\beta}(v)) + E((\hat{\beta} - \hat{\beta}(v))(\hat{\beta} - \hat{\beta}(v))^T) \tag{4-8}$$

$$= \mathrm{Cov}(\hat{\beta}(v)) + E((\frac{1}{S^2(v)})(\hat{\beta} - \hat{\beta}(v))(\hat{\beta} - \hat{\beta}(v))^T) \tag{4-9}$$

In order to use Theorem 5 to get a swindle we need to be able to compute $\mathrm{Cov}(\hat{\beta}(v))$. As before, this is generally difficult, but can be partially accomplished from the following result.

Theorem 6: $\mathrm{Cov}(\hat{\beta}(v)) = E(X^T\langle v^2\rangle X)^{-1}$ (4-10)

Proof: Condition on v. QED.

There do not appear to be too many cases where $E(X^T\langle v^2\rangle X)^{-1}$ can be computed exactly so that either approximations or Monte Carlo estimates must be used. Again we get two swindle formulas depending on what we use for $\mathrm{Cov}(\hat{\beta}(v))$. If $\mathrm{Cov}(\hat{\beta}(v))$ is known or can be well approximated, then use

$$\text{Estimated Cov}(\hat{\beta}) = \text{Cov}(\hat{\beta}(v)) + \text{Monte Carlo Average}((\hat{\beta} - \hat{\beta}(v))(\hat{\beta} - \hat{\beta}(v))^T) \tag{4-11}$$

If $E(X^T\langle v^2\rangle X)^{-1}$ must be estimated by Monte Carlo use

$$\text{Estimated Cov}(\beta) = \text{Monte Carlo Average}((X^T\langle v^2\rangle X)^{-1}) + \text{Monte Carlo Average}((\hat{\beta} - \hat{\beta}(v))(\hat{\beta} - \hat{\beta}(v))^T) \tag{4-12}$$

In the regression case, because we will usually have to use Monte Carlo is estimate $E(X^T\langle v^2\rangle X)^{-1}$ it is likely that the swindle will not produce much of an improvement over simple random sampling.

5. Some Final Remarks

The basic result that underlies all of this discussion is (2-2). This is a special case of a general result that holds for the best location invariant estimator for any given distribution (some conditions may be necessary). This result is given but not proved in the next theorem.

Theorem 7: If $\hat{\beta}$ is any location invariant estimator and $\hat{\beta}_0$ is the best location invariant estimator, then

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\hat{\beta}_0) + E(\hat{\beta} - \hat{\beta}_0)^2 \qquad (5\text{-}1)$$

It is evident that if $\hat{\beta}_0$ can be computed easily, and if $\mathrm{Var}(\hat{\beta}_0)$ can be computed exactly then (5-1) provides a basis for a Monte Carlo swindle. Unfortunately, neither $\hat{\beta}_0$ nor its variance can be easily computed for many cases outside of the Gaussian. The use of the $\hat{\beta}(v)$ is a compromise for this state of affairs.

Role of Configurations

In the discussion of this Monte Carlo swindle by Andrews, *et al.*, [1972] the concept of a "configuration" plays a prominent role. In the development here no such concept arises. We comment on this briefly, now.

A configuration is a sample (the y's) adjusted in a particular way. One important example of a configuration is

$$y - X\hat{\beta}_{LS} \qquad (5\text{-}2)$$

We note that a configuration is unchanged by the addition of $X\gamma$ to y for any choices of γ. A scale invariant configuration is given by

$$(y - X\hat{\beta}_{LS})/S \ . \qquad (5\text{-}3)$$

The reader is referred to Andrews, *et al.*, [1972] for the use of configurations in the derivation of the swindles discussed in the previous sections.

In our derivation here there is a place for regression or regression and scale invariant configurations, but they are not central to the swindle as such, rather they may be used to make some of the computing more efficient. In (4-4) we may compute the Monte Carlo Average in one of two ways:

$$\text{Monte Carlo Average of } ((\hat{\beta} - \hat{\beta}_{LS})(\hat{\beta} - \hat{\beta}_{LS})^T) \tag{5-4}$$

$$\text{or " " " " } (\hat{\beta}(y - X\hat{\beta}_{LS})\hat{\beta}(y - X\hat{\beta}_{LS})^T) \tag{5-5}$$

If we are going to compare several estimators then from a computational stand point it makes sense to compute $\hat{\beta}_{LS}$, then form the configurations, $y - X\hat{\beta}_{LS}$, and then compute $\hat{\beta}$ on these rather than on y. This saves a large number of subtractions. If a regression and scale swindle is going to be used then rather than

$$\text{Monte Carlo Average of } (\frac{1}{S^2}(\hat{\beta} - \hat{\beta}_{LS})(\hat{\beta} - \hat{\beta}_{LS})^T) \tag{5-6}$$

it is more efficient to form the configurations given by (5-3) and compute (5-6) via

$$\text{Monte Carlo Average of } (\hat{\beta}(\frac{y - X\hat{\beta}_{LS}}{S})\hat{\beta}(\frac{y - X\hat{\beta}_{LS}}{S})^T) \tag{5-7}$$

Similar remarks hold for the configurations that arise from the swindle in the non-Gaussian case.