General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models

L. Breiman; W. S. Meisel

Journal of the American Statistical Association, Vol. 71, No. 354 (Jun., 1976), 301-307.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28197606%2971%3A354%3C301%3AGEOTIV%3E2.0.CO%3B2-S

Journal of the American Statistical Association is currently published by American Statistical Association.

General Estimates of the Intrinsic Variability of Data in Nonlinear Regression Models

L. BREIMAN and W. S. MEISEL*

A dependent variable is some unknown function of independent variables plus an error component. If the magnitude of the error could be estimated with minimal assumptions about the underlying functional dependence, then this could be used to judge goodness-of-fit and as a means of selecting a subset of the independent variables which best determine the dependent variable. We propose a procedure for this purpose which is based on a data-directed partitioning of the space into subregions and a fitting of the function in each subregion. The behavior of the procedure is heuristically discussed and illustrated by some simulation examples.

1. INTRODUCTION

An important phase in nonlinear regression problems is the exploration of the relationship between the independent and dependent variables. Much of the current literature on nonlinear regression assumes that a parametric class of regression functions has somehow been selected and focuses on the relatively straightforward problem of minimizing the sum of the squared errors over the parameter space. (This problem is not necessarily easy, but is well-understood (see, e.g., Chamber's survey [2]).)

But many typical data analytic problems are characterized by their high dimensionality (a large number of independent variables) and the lack of any *a priori* identification of a natural and appropriate family of regression functions.

This sort of situation raises three important and difficult problems.

1. How can we select those independent variables which most significantly affect the dependent variable?
2. Once a relatively small subset of independent variables has been selected, how can an appropriate family of regression functions be chosen?
3. Do we get a good fit to the data by the best least squares fit selected within the family?

In analyzing the predictive capability of a set of independent variables, only minimal assumptions regarding the form of the regression function should be made. Otherwise, variables with high predictive capability may be discarded because the appropriate form of the regression function for that variable is not included in the study. On the other hand, if too large a class of functional forms is allowed, there is the danger of overfitting the data; that is, of fitting the random fluctuations in the data rather than the (hopefully) smooth regression functions.

The purpose of this paper is to study a different approach to the exploratory phase in nonlinear regression, and to goodness-of-fit testing. It is based on what we call *general estimates* of the *intrinsic* variability of the data. These are estimates of the standard deviation (or some other measure) of the fluctuation of the dependent variable around the true regression function, such that the estimates depend only on very general assumptions regarding the functional relationship between the dependent and independent variables.

If one has a general estimate of, say, the standard deviation of the intrinsic variability, this can be used to estimate the percent of variance explained by any subset of independent variables. This gives a method of ranking the predictive capabilities of various subsets of independent variables for the variable selection problem.

A general estimate can also be the basis for a goodness-of-fit criterion. Given a parametric family of regression functions, the comparison of a general estimate of the error variance and the minimum over the family of the residual sum of squares gives a measure of how well the parametric model fits. Mallow's C_P statistic [4] for the linear case is an example of this approach.

More specifically, suppose that our data consist of n points $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in M-dimensional space X and associated values $y_1, \ldots, y_n$ of the dependent variable. Assume the model

$$y_i = \epsilon_i + \phi(\mathbf{x}_i) \ , \quad i = 1, \ldots, n$$

where the ϵ_i are independent outcomes from a $N(0, \sigma^2)$ distribution with unknown σ^2, and $\phi(\mathbf{x})$ is unknown. The problem is to construct an estimate of σ^2 with only a minimal set of assumptions about the form of $\phi(\mathbf{x})$.

If there is enough replication in the data, then the replicated points can be used to construct a simple general estimate; but in the usual regression problem, one does not have available repeated observations at the same point. One might group values of $\mathbf{x}_i$ that are close

* L. Breiman is lecturer, Department of Mathematics, UCLA, and consultant, and W.S. Meisel is manager, both at Data Sciences Division, Technology Service Corporation, Santa Monica, CA. 90403. Research was sponsored by the Air Force Office of Scientific Research/AFSC, U.S. Air Force, under Contract No. F44620-71-C-0093. The authors wish to thank Mike Teener, who programmed and ran the simulation examples, and the referees for their reviewing work, which resulted in significant improvements.

© Journal of the American Statistical Association
June 1976, Volume 71, Number 354
Applications Section

together; but, for M large, not many of the $\mathbf{x}_i$ points may be close. Furthermore, measuring distance and gathering together neighboring points in a space of high dimensions is a difficult and ill-defined computational process, particularly since "distance" varies with scaling of the variables relative to one another. (See Daniel and Wood [3, Ch. 7] for an interesting way of handling the scaling problem in computing distances for the linear case.)

The essential feature of the estimation method we study is the construction of a fitting surface where the construction is data-directed. The idea is to approximate $\phi(\mathbf{x})$ by piecewise linear patches, whose sizes are determined by the data in the following way: At each stage, the space $\mathbf{X}$ is broken into regions $R_1, \ldots, R_J$ and the data in each region are fitted by linear regression. Take any one of these regions R, split it by a randomly oriented plane, fit a linear function $b_0 + \mathbf{b}\cdot\mathbf{x}$ in each of the two new subregions and use an F-ratio to determine if the split has significantly reduced the residual sum of squares. If it has, accept the split and try splitting the new subregions. If not, try another randomly selected split on R. If, after a fixed number of tries, there is no successful split of R, let it stand. The details of this algorithm are given in Section 2.

The estimate $\hat{\sigma}$ is obtained by computing the residual sum of squares about each linear segment of the final piecewise linear approximation to $\phi(\mathbf{x})$ and combining them. Thus, very little is assumed about the shape of $\phi(\mathbf{x})$ in computing $\hat{\sigma}^2$. In general, all that is required for the method to produce accurate estimates is that $\phi(\mathbf{x})$ can be "locally well-approximated" by a linear function. The radius of the "local" fitting region is determined by the sample density, and the closeness to which $\phi(\mathbf{x})$ has to be approximated in the region is determined by the requirement that the squared error in the linear approximation to ϕ over the region be small compared to σ^2.

Since we are assuming $\phi(\mathbf{x})$ unknown, it is not clear *a priori* whether $\phi(\mathbf{x})$ will be suitably smooth in the sense just stated. Therefore, we have designed some diagnostics which are discussed in Section 3.

As a test of this method, a number of simulation examples are discussed in Section 4. Polynomial regressions were also run on the examples of Section 4 to see how well σ^2 could be estimated by standard methods. The examples involved data sets of 100, 500, and 2,000 points in four dimensions, with four different values of σ^2. Our conclusions are that the algorithm is quite effective in handling large sets. Within limitations, it gives accurate estimates of σ^2. Because of its simplicity, it has sufficient computational efficiency so that problems involving thousands of data points in four dimensions can typically be run for less than ten dollars.

The appendix contains a more detailed analysis of the splitting rule which leads to an estimate of the mean square fitting error to $\phi(\mathbf{x})$.

The particular procedure just defined involves fitting a linear function $b_0 + \mathbf{b}\cdot\mathbf{x}$ to the data in each subregion. This can be generalized to the use of quadratic or higher-order polynomial fitting in each subregion.

The final functional estimate of $\phi(\mathbf{x})$ produced by the algorithm consists of discontinuous patches of hyperplanes. This is not, in most situations, a satisfactory form of approximation. However, we did not design or intend it to be used to get a good approximation to $\phi(\mathbf{x})$. Its use is to give fast and accurate estimates of σ in nonlinear situations, leading to a usable procedure for variable selection.

2. DESCRIPTION OF THE PROCEDURE

Suppose that at any stage in the process, the space is broken into the subregions $R_1, \ldots, R_K$. In each subregion, the points are fitted by a linear least-squares regression. That is, coefficients b_0, $\mathbf{b} = (b_1, \ldots, b_M)$ are found in R_j which minimize

$$D_j = \sum_{x_i \in R_j} (y_i - b_0 - \mathbf{b}\cdot\mathbf{x}_i)^2 ,$$

where $\mathbf{b}\cdot\mathbf{x}_i$ is the inner product of $\mathbf{b}$ and $\mathbf{x}_i$. Let the minimum value of D_j be D_j^*. Then if $\phi(\mathbf{x})$ is linear in $\mathbf{x}$ over R_j, an unbiased estimate of σ^2 is given by

$$\hat{\sigma}_{R_j}^2 = D_j^*/(N_j - M - 1)$$

where N_j is the number of sample points in R_j. Now pick out for checking any previously unchecked subregion R_j. Subdivide R_j by first choosing at random an M-dimensional direction vector $\mathbf{n}$. Take any plane orthogonal to $\mathbf{n}$ and move it toward R_j until it divides in half (or as closely as possible) the sample points $\{\mathbf{x}_i\}$ in R_j. Denote the two resulting subregions of R_j by R_{j1} and R_{j2}.

Do a linear regression in each two regions, winding up with the minimum values

$$D_{j1}^* = \min_{b_0,\mathbf{b}} \Big(\sum_{x_i \in R_{j1}} (y_i - b_0 - \mathbf{b}\cdot\mathbf{x}_i)^2\Big)$$

$$D_{j2}^* = \min_{b_0,\mathbf{b}} \Big(\sum_{x_i \in R_{j2}} (y_i - b_0 - \mathbf{b}\cdot\mathbf{x}_i)^2\Big) .$$

Certainly $D_{j1}^* + D_{j2}^* \le D_j^*$, since $D_{j1}^* + D_{j2}^*$ are the result of a double linear least-squares fit to the data points in R_j. But if $\phi(\mathbf{x})$ is adequately fit in R_j by a hyperplane, then the difference $D_j^* - D_{j1}^* - D_{j2}^*$ is entirely attributable to the slightly better fit gotten to the random component by minimizing over $2(M + 1)$ parameters instead of $M + 1$ parameters. This can be quantified by using the following result.

Proposition 1: If $\phi(\mathbf{x})$ is linear in $\mathbf{x}$ on R_j, then

$$F = \left(\frac{N_j - 2(M+1)}{M+1}\right)\left(\frac{D_j^* - D_{j1}^* - D_{j2}^*}{D_{j1}^* + D_{j2}^*}\right) \qquad (2.1)$$

has an $F_{M+1,N_j-2(M+1)}$ distribution. The proof of this is a straightforward application of one of the basic theorems on distributions usually associated with analysis of variance.

If $\phi(\mathbf{x})$ is strongly nonlinear in R_j, we may get a much better fit to $\phi(\mathbf{x})$ by fitting it separately in each of regions

R_{j1} and R_{j2}. Then F would tend to be large. Thus, our decision whether to stay with the original linear fit over R_j or go to the separate fits over R_{j1} and R_{j2} is based on the size of F compared to a fixed significance level selected from the $F_{M+1,N_j-2(M+1)}$ distribution. But even if $\phi(\mathbf{x})$ is strongly nonlinear over R, the original subdivision into R_{j1} and R_{j2} may not produce a significantly better fit. For this reason we repeat the subdivision of R_j K times at most, each time choosing a new random planar direction, computing the F value and comparing it with the appropriate distribution. As soon as any split produces a significant F value, the splitting terminates, R_j is replaced by R_{j1}, R_{j2}, and the process started anew. If no significant F value is produced in K trials, then R_j is left undivided and the same process applied to any previously unchecked region.

However, if at any stage in the process a subregion contains too few points, it will not be split further. We set this lower bound at $2(M + 1)$ points (or as specified externally by the user). Therefore, a region is terminal in this procedure if either

i. It has more than $2(M + 1)$ points, and none of the attempted splittings significantly lower the variance; or
ii. It contains $2(M + 1)$ or fewer points.

Call the first type of terminal region *nonminimal*, the second type *minimal*.

When the sequential procedure terminates, the space is divided into subregions $R_1, \ldots, R_m$ containing $N_1, \ldots, N_m$ data points. We take as our estimate for σ^2 the value

$$\hat\sigma^2 = \frac{1}{n}\sum_{j=1}^{m} D_{R_j}{}^* \ . \qquad (2.2)$$

The reason for dividing by n rather than $n - m(M + 1)$ is discussed in the appendix.

The actual distributional properties of the splitting algorithm are complicated. The F values on successive attempted splits are dependent and Proposition 1 holds only for the first split. Thus, the use of the ratio (2.1) and, in fact, all of the approximations derived in the following sections should be regarded as heuristic—to be justified by example and use.

3. DIAGNOSTICS

As mentioned in Section 1, this method is capable of giving accurate estimates of σ^2 only if $\phi(\mathbf{x})$ satisfies some smoothness conditions. The exact conditions are complicated and arise out of a complex interaction between the local linearity of $\phi(\mathbf{x})$, the local density of the points $\{x_i\}$, and the variance σ^2. If $\phi(\mathbf{x})$ does not meet the smoothness requirements, then our estimate $\hat\sigma^2$ is biased upward. In fact, it is actually an estimator of σ^2 plus a term which measures the square of the error in approximating $\phi(\mathbf{x})$ locally by linear functions. (We discuss this further later.) Since $\phi(\mathbf{x})$ is unknown, we usually have no *a priori* way of knowing whether $\phi(\mathbf{x})$ satisfies the relevant conditions. Diagnostics for checking the validity of the results become important. We propose some diagnostics in this section; their effectiveness is illustrated in the examples of Section 4.

3.1 Expected Approximation Error

For any region R containing N data points $\mathbf{x}_1, \ldots, \mathbf{x}_N$, let $e_R{}^2$ be the mean squared bias in the linear fit to $\phi(\mathbf{x})$. That is,

$$e_R{}^2 = \sum_i (\phi(\mathbf{x}_i) - E(\hat b_0) - E\hat{\mathbf{b}}\cdot\mathbf{x}_i)^2/N \ . \qquad (3.1)$$

where $\hat b_0$, $\hat{\mathbf{b}}$ are the estimated regression coefficients. It follows that

$$e_R{}^2 = \min \sum_i (\phi(\mathbf{x}_i) - a_0 - \mathbf{a}\cdot\mathbf{x}_i)^2)/N \ . \qquad (3.2)$$

That is, e_R is the error of the least-squares hyperplane fit to $\phi(\mathbf{x})$ at the points of $\{\mathbf{x}_i\}$. By using the first form of $e_R{}^2$, we conclude that

$$ED_R{}^* = (N - M - 1)\sigma^2 + Ne_R{}^2 \ . \qquad (3.3)$$

Using (2.2), the expectation of the estimate is

$$E\hat\sigma^2 = \left(\frac{n - m(M + 1)}{n}\right)\sigma^2 + \frac{1}{n}\sum_{j=1}^{m} N_j e_{R_j}{}^2$$
$$= \left(\frac{n - m(M + 1)}{n}\right)\sigma^2 + e^2 \ , \qquad (3.4)$$

where

$$e^2 = \frac{1}{n}\sum_{j=1}^{m} N_j e_{R_j}{}^2 \ .$$

In the appendix, we give an argument leading to the approximation

$$E\hat\sigma^2 \simeq \sigma^2 \ . \qquad (3.5)$$

Denote by $\hat\phi(\mathbf{x})$ the terminal piecewise linear regression surface produced by the procedure. The mean square approximation error is defined as

$$W^2 = \frac{1}{n}\sum_{.1}^{n} (\phi(\mathbf{x}_i) - \hat\phi(\mathbf{x}_i))^2 \ . \qquad (3.6)$$

In the appendix, we show that if σ^2 is not small compared to e^2, an estimate for W^2 is given by

$$\hat W^2 = [(2(M + 1)m)/n]\hat\sigma^2 \ , \qquad (3.7)$$

where m is the number of terminal regions. This approximation holds up fairly well when compared to the actual values of W^2 computed in the simulation examples.

3.2 Minimal Terminal Regions

In (3.7) we begin to see the effects of dimensionality, sample size, σ^2, and the curvature of the function $\phi(\mathbf{x})$ on W^2. The more highly curved ϕ, the more splitting will occur, and the larger m will be.

The size of the smallest possible terminal regions generated by the program is governed by the requirement that it contain more than, say, L data points. The lowest attainable values of e^2 depends on how much $\phi(\mathbf{x})$ departs from linearity in $\mathbf{x}$ across such minimal regions. If σ^2 is not substantially larger than the minimal value of e^2, then the algorithm may keep splitting regions in an attempt to get a better approximation to $\phi(\mathbf{x})$. In this situation, a proportion of the terminal regions will be minimal. *This is a good diagnostic of the effectiveness of the algorithm.* The existence of points in minimal terminal regions indicates that the function is not "approximately linear" over regions of size determined by the density of the data points $\{\mathbf{x}_i\}$. Here "approximately linear" means that e_R^2 is small compared to σ^2.

As one might expect, the more curved $\phi(\mathbf{x})$ and the smaller σ, the more points fall into minimal terminal regions. If the number of points in minimal terminal regions indicates that the algorithm is in difficulty, what can be done? We believe that a logical step is to use the same algorithm again, but replacing the linear fitting functions by quadratic functions.

As the sample size increases, the density of the points $\{\mathbf{x}_i\}$ increases, and the size of the regions containing L adjacent points decreases. Thus, with increased sample size, the biasing effect of the local nonlinearity of $\phi(\mathbf{x})$ is decreased.

However, even if $\phi(\mathbf{x})$ is very smooth, there is no guarantee that the procedure will work well. For instance, in one dimension, let $\phi(x) = \sin x$, $-4\pi \leq x \leq 4\pi$. For large sample size, the original linear fit will be nearly on the x-axis and splitting the points in half will produce two linear fits, both closely following the x-axis. Our opinion is that this type of occurrence is rare in real data. In our simulation examples, inaccurate estimates of σ^2 were always due to the lack of appropriate local linearity of $\phi(\mathbf{x})$ relative to the sample size and magnitude of σ^2. But to provide an additional safeguard, one might modify the selection of splitting planes so that the fraction of points split off as well as the direction of the plane is random.

The splitting algorithm used raises a number of questions, most of which are difficult to answer in an analytically precise way. For instance, if the underlying function $\phi(\mathbf{x})$ is nonlinear in the region R_j, then (2.1) is a ratio of two noncentral χ^2 variables, and the decision to accept the split may be based more on the relative sizes of the noncentrality parameters than on the reduction of the residual sums of squares. (See the appendix for a more detailed discussion.) In practice, our simulation examples and calculations both indicate that the effect of the nonlinearity of $\phi(\mathbf{x})$ is to induce more "acceptable" splits. But this is not a particularly harmful drawback, as oversplitting will not generally produce much bias in the estimates.

3.3 Repetition

Another important question is how sensitive the value of $\hat{\sigma}^2$ is to the random selection of splitting hyperplanes. This depends on how well $\phi(\mathbf{x})$ can be locally approximated by linear functions. In numerical examples, repetitions of the process were carried out using the same data.

We conclude that if $\phi(\mathbf{x})$ can be adequately approximated with the given sample size and distribution, then the procedure provides accurate estimates of σ^2 which vary only a few percent when independent repetitions of the process are carried out. On the other hand, if the procedure cannot provide accurate estimates of σ, then the major part of the error is the functional approximation error. Generally, this error is sensitive to the selection of splitting planes and varies considerably when the process is repeated on the same data. Therefore, another good (but more expensive) diagnostic is to make a few repeated runs and check the variability of the estimate $\hat{\sigma}^2$. This variability did not occur in the particular simulation example presented in the text (see Table 3), but has occurred in a number of other examples we have explored.

4. TWO SIMULATION EXAMPLES

Two examples were designed, both four-dimensional.

Example I:

$$\phi_1(\mathbf{x}) = \exp\left[-\tfrac{1}{6}(x_1^2 + x_2^2 + x_3^2 + x_4^2)\right].$$

The sample points $\{\mathbf{x}_1\}$ were chosen by selecting a random direction in four-space and then proceeding in that direction from (0, 0, 0, 0) a distance R, where R was uniformly distributed on [0, 6].

Example II:

$$\phi_2(\mathbf{x}) = 1.5\phi_1(\mathbf{x}) + 1.0\phi_1(\mathbf{x} - (6, 6, 6, 6)) .$$

Here half of the sample points were chosen as in Example I. The other half were chosen in a similar way except that the origin was taken to be (6, 6, 6, 6).

In Example I, $\phi(\mathbf{x})$ has a single peak at the origin. In Example II, it has two unequal peaks, the larger at the origin and the smaller at (6, 6, 6, 6).

Gaussian noise ϵ_i with variance σ^2 was added to each functional value $\phi(\mathbf{x}_i)$. To trace the effect of σ on the procedures, we used the values

$$\sigma = .4, .2, .1, .05$$

in different runs. To check the effect of sample size, we used $n = 100$, 500, and 2,000. The sample size of 100 does not sample the four-variable function sufficiently to retrieve its form accurately, and is included as an example of a situation where the "true" model may not be retrievable by any method.

The same data sets and additive noise were run on a standard polynomial regression, starting with only linear terms, then adding the quadratic and finally the cubic terms. In the cubic regression, 35 coefficients were estimated.

The estimates of σ by the intrinsic variability algorithm (INVAR) and polynomial regressions are presented in Table 1. Since the actual sample standard deviation of the noise differs from the underlying σ, its value B is given. The same set of 2,000 normal deviates was used throughout, which accounts for the systematically low values of the sample σ. They were generated using the approximation given in [1, Ch. 26.2.22, p. 933] and the CDC uniform random number generator.

1. Examples I and II

Sample size	σ	Sample σ	INVAR $\hat{\sigma}$	Estimated σ using cubic regression	% variance due to $\phi(\mathbf{x})$	% variance explained by INVAR	% variance explained by cubic regression
				a. Example I			
100	.4	.378	.341	.431	46.9	56.8	31.0
100	.2	.189	.284	.270	76.7	47.4	52.4
100	.1	.095	.121	.207	92.6	88.0	64.9
100	.05	.047	.128	.186	97.8	85.5	69.4
500	.4	.394	.398	.429	42.3	41.1	31.6
500	.2	.197	.189	.266	75.3	77.2	54.9
500	.1	.099	.113	.206	91.5	90.2	67.3
500	.05	.049	.078	.188	98.1	95.1	71.4
2000	.4	.394	.385	.430	40.5	43.2	29.2
2000	.2	.197	.198	.264	74.5	74.3	54.2
2000	.1	.098	.102	.203	92.4	91.8	67.6
2000	.05	.049	.061	.185	98.0	96.9	71.9
				b. Example II			
100	.4	.378	.402	.434	61.8	56.8	49.7
100	.2	.189	.284	.332	85.7	67.6	55.8
100	.1	.095	.247	.306	96.8	78.7	56.1
100	.05	.047	.244	.301	98.9	70.5	55.2
500	.4	.394	.378	.485	57.3	60.7	35.24
500	.2	.197	.207	.349	84.4	82.8	51.1
500	.1	.099	.129	.305	95.6	92.5	58.0
500	.05	.049	.095	.293	98.9	95.8	60.0
2000	.4	.394	.398	.480	57.4	54.2	33.4
2000	.2	.197	.206	.342	83.7	82.2	50.9
2000	.1	.098	.112	.299	95.5	94.1	58.2
2000	.05	.049	.088	.288	98.8	96.3	60.3

The last three columns of Table 1 restate the results in terms of the percent of variance explained:

$$V = 100[1 - (\sigma_I^2/\sigma_y^2)] ,$$

where

σ_I^2 = the actual or estimated intrinsic variance

σ_y^2 = the variance of the dependent variable across the sample.

The 100-sample case produces an underdetermined problem, and except for the largest σ, both the algorithm and cubic regression produce rather poor results for both examples. One hundred observations in four dimensions cannot lead to an accurate representation of a complex surface, unless the surface is of a parametric family with a low-dimensional parameter space.

Clearly, INVAR is producing more accurate estimates than cubic regression. As sample size increases, there is very little actual (or expected) improvement in the cubic regression. The INVAR algorithm takes advantage of increased sample size to do more splitting and track the surface more accurately. This is illustrated by the increasing number of terminal regions (Table 2).

2. Number of Terminal Regions

σ	n		
	100	500	2000
		Example I	
.4	5	9	26
.2	4	20	43
.1	8	24	70
.05	7	33	100[a]
		Example II	
.4	4	15	29
.2	5	20	54
.1	6	32	100[a]
.05	6	42	100[a]

[a] 100 terminal regions was the maximum possible as the algorithm was employed.

Although the surface in Example II is bumpier than that in Example I, INVAR's accuracy degenerates only slightly. We would expect that the cubic regressions have more difficulty in fitting the second surface, and this turns out to be so.

With low values of σ^2 the process changes in character from a piecewise linear regression method into a procedure that attempts to approximate a function $\phi(\mathbf{x})$ by hyperplane segments. For any sizable region, R, suppose that $\sigma^2 \ll e_R^2$. Then, defining

$$\Delta^2 = e_R^2 - \tfrac{1}{2}(e_{R_1}^2 + e_{R_2}^2) ,$$

we have

$$EF \simeq \frac{N - 2(M+1)}{M+1} \cdot \frac{2\Delta^2}{e_{R_1}^2 + e_{R_2}^2} ,$$

and the splitting continues as long as this ratio is significant. If the function is peculiar, this splitting condition may not yield a good approximation method. Whether it does or not, what we are left with when the process terminates is not an estimate of σ^2, but the value of e^2.

At a higher sample density, the difficulties that the procedure has in adapting to the data is somewhat reflected in the percentage of points turning up in Minimal Terminal Regions.

Finally, to check the effects of the random selection of splitting hyperplanes, five iterations each of Example I ($\sigma = .2$) and Example II ($\sigma = .05$) were carried out, using the same data for each set of five runs with $n = 500$ (Table 3).

At the suggestion of one of the referees, we also ran INVAR on pure noise ($\phi(\mathbf{x}) \equiv 0$) at a variety of sample

3. Iterations of INVAR on Same Data

Iteration	$\hat{\sigma}$ Example I	$\hat{\sigma}$ Example II
1	.199	.112
2	.184	.110
3	.197	.110
4	.214	.110
5	.206	.124
(Sample σ)	(.197)	(.049)

sizes and variances. In each case, the initial splits were rejected and the data fitted by a single hyperplane.

4.1 Informal Observations on Algorithm Parameters

The $2(M+1)$ bound used in defining a terminal region was selected because if a region has fewer than this number of points, and it is split, then at least one of the two subregions does not have enough points to define a hyperplane. In our simulation work, we used the .99 significance level and set $K = 5$, where K is the maximum allowed number of test splittings of a subregion. We found that lower significance levels led to too much splitting induced by random improvement. The value $K = 5$ was a compromise between computational efficiency and statistical optimality. However, we conjecture that, if the data are splittable, then the split will occur fairly early in the trials, and a sequence of five failures carries a strong implication that further trials will also end in failure. If K is increased, the significance level for F should also be increased.

5. CONCLUSION

We have presented an algorithm for estimating the intrinsic variability of a dependent variable about an unknown underlying nonlinear model, given a set of independent variables and a limited sample of data exemplifying the relationship. Such estimates can be used for (1) efficiently comparing the efficacy of alternative groups of independent variables, and (2) estimating the limiting accuracy achievable by a good choice of nonlinear form for the model without having to discover that form.

The results of theoretical analyses and several examples, as well as use in several applications by the authors, indicates that the method provides good estimates most of the time, and, when it doesn't, provides diagnostics which allow detection of the failure.

APPENDIX

In this appendix we derive (3.4), using crude approximations, showing that $\hat{\sigma}^2$ is approximately unbiased, and equation (3.7) for the estimated mean square approximation error to $\phi(\mathbf{x})$.

Suppose that R is a nonminimal terminal region. From (3.3),

$$E(D^* - D_1^* - D_2^*) = (M+1)\sigma^2 + N\Delta^2,$$

where

$$\Delta^2 = e_R^2 - \tfrac{1}{2}(e_{R_1}^2 + e_{R_2}^2).$$

This latter term is the decrease in the mean square error when $\phi(\mathbf{x})$ is approximated by the best linear fits individually over R_1 and R_2 as compared with the best linear fit over the original set R.

Furthermore, since

$$E(D_1^* + D_2^*) = (N - 2(M+1))\sigma^2 + N(e_{R_1}^2 + e_{R_2}^2)/2,$$

the F-ratio used for splitting has, in general, a double noncentral F-distribution with noncentrality parameters

$$\delta_1^2 = N\Delta^2/\sigma^2, \qquad \delta_2^2 = N(e_{R_1}^2 + e_{R_2}^2)/2\sigma^2.$$

The expectation of this ratio can be approximated as

$$E(F) \simeq \left(1 + \frac{N\Delta^2}{(M+1)\sigma^2}\right)\left[\frac{L}{L+\delta_2^2}\right],$$

where $L = N - 2(M+1)$. We assume that the term in brackets is near unity for the moment, and write

$$E(F) \simeq 1 + [N\Delta^2/((M+1)\sigma^2)].$$

At the significance levels we used, for $N - 2(M+1) \geq 15$, $M \geq 5$, the critical F-values range from about two to four. Therefore, the splitting will usually keep going in a region if splits can be found such that $E(F) \geq 2$, or

$$[N\Delta^2/((M+1)\sigma^2)] \geq 1.$$

Writing this splitting condition as

$$\Delta^2 \geq ((M+1)/N)\sigma^2$$

makes it clear that for the same $\phi(\mathbf{x})$, the smaller σ^2 is, the more splitting will go on. To define what effect this will have on the final error of approximation to $\phi(\mathbf{x})$, we make the assumption that if e_R^2 cannot be reduced by $(M+1)\sigma^2/N$ by splitting, then

$$e_R^2 \simeq ((M+1)/N)\sigma^2 \tag{A.1}$$

leading to

$$e^2 \simeq \frac{m(M+1)}{n}\,\sigma^2.$$

Substituting this result into (3.4) gives (3.5).

Let W_R^2 be the mean square approximation error to $\phi(\mathbf{x})$ over the region R (see (3.6)). Then, by an argument virtually identical to that in [3, Ch. 6, pp. 86–7], we get that

$$EW_R^2 = e_R^2 + ((M+1)/N)\sigma^2.$$

From (A.1)

$$EW_R^2 = ((2(M+1))/N)\sigma^2.$$

Putting the regions together gives

$$EW^2 = \frac{1}{n}\sum_k N_k EW_{R_k}^2 = \frac{2(M+1)m}{n}\,\sigma^2\,. \tag{A.2}$$

As we might expect, our ability to approximate the regression function is limited by the variability of the data. The larger σ^2 is, the poorer our ability to approximate $\phi(\mathbf{x})$, for n fixed.

If $\phi(\mathbf{x})$ were exactly linear in each of the regions $R_1, \ldots, R_m$, then

$$EW_{R_j}^2 = ((M+1)/N_j)\sigma^2$$

so that

$$EW^2 = (((M+1)m)/n)\sigma^2 \tag{A.3}$$

This is one-half as large as the general approximation given in (A.2).

The usefulness of (A.2) is that it can be combined with $\hat{\sigma}^2$ to give

4. Estimates of Fitting Error

	n					
σ	100		500		2000	
	W	$\hat{W}$	W	$\hat{W}$	W	$\hat{W}$
			Example I			
.4	.24	.24	.23	.18	.16	.14
.2	.22	.19	.14	.12	.10	.09
.1	.11	.11	.09	.08	.07	.06
.05	.12	.11	.07	.07	.06	.04
			Example II			
.4	.34	.26	.25	.20	.19	.15
.2	.27	.20	.18	.13	.14	.11
.1	.24	.19	.14	.10	.10	.08
.05	.25	.19	.10	.08	.08	.06

the estimate (3.7)

$$\hat{W}^2 = ((2(M+1)m)/n)\hat{\sigma}^2$$

for W^2, thus providing a measure of how closely the function $\phi(\mathbf{x})$ is approximated by the piecewise linear fit. The estimate (3.7) was checked in the numerical examples of Section 4 against the actual value of

$$\frac{1}{n}\sum_{1}^{n}(\phi(x_i) - \hat{\phi}(x_i))^2$$

and held up surprisingly well, as shown in Table 4.

[*Received January 1974. Revised October 1975.*]

REFERENCES

[1] Abramowitz, M. and Stegun, L.A., *Handbook of Mathematical Functions*, Washington, D.C.; National Bureau of Standards, 1964.

[2] Chambers, John M., "Fitting Nonlinear Models: Numerical Techniques," *Biometrika*, 60, No. 1 (1973), 1–13.

[3] Daniel, C. and Wood, F.S. (assisted by J.W. Gorman), *Fitting Equations to Data*, New York; Wiley-Interscience, 1971.

[4] Mallows, C.B., "Choosing Variables in a Linear Regression: A Graphical Aid," presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas, May 7–9, 1964.