



The Reduced Monotonic Regression Method

Michael J. Schell; Bahadur Singh

Journal of the American Statistical Association, Vol. 92, No. 437. (Mar., 1997), pp. 128-135.

Stable URL:

<http://links.jstor.org/sici?sici=0162-1459%28199703%2992%3A437%3C128%3ATRMRM%3E2.0.CO%3B2-1>

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

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/astata.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

The JSTOR Archive is a trusted digital repository providing for long-term preservation and access to leading academic journals and scholarly literature from around the world. The Archive is supported by libraries, scholarly societies, publishers, and foundations. It is an initiative of JSTOR, a not-for-profit organization with a mission to help the scholarly community take advantage of advances in technology. For more information regarding JSTOR, please contact support@jstor.org.

The Reduced Monotonic Regression Method

Michael J. SCHELL and Bahadur SINGH

Medical researchers often desire to categorize patients into monotonic response groups based on the relationship between continuous variables. Isotonic regression fits consist of level sets of increasing value, for which the estimated response is constant. However, the number of level sets obtained is often large, preventing simple description. This article introduces two new nonparametric methods called reduced isotonic regression and reduced monotonic regression, the latter being a two-sided extension of the former for use when the direction of the trend is unknown. Using a backward elimination algorithm, the new procedures reduce the number of level sets by combining those whose values do not differ greatly. For the statistical relations examined here, the reduced monotonic method averaged at most 30% of the number of level sets obtained for isotonic regression. The method is illustrated with an example that examines the relationship between risk factors for survival among children with leukemia. In simulation studies, the reduced monotonic method fits the data as closely as alternative methods that combine isotonicity and smoothing, while improving greatly on isotonic regression. The method is also related to changepoint models of normally distributed sequences.

KEY WORDS: Changepoint analysis; Isotonic regression; Monotonicity; Nonparametric regression; Order-restricted inference.

1. INTRODUCTION

In medical research two variables are often monotonically related to each other; for example, amount of exercise and serum cholesterol level. However, many of these relationships are decidedly nonlinear. Isotonic regression theory (see Robertson, Wright, and Dykstra [RWD] 1988) provides a nonparametric solution to this problem. The solution, which is composed of level sets for which the estimated response is constant, minimizes the sum of squares of the model to the data under the isotonicity restriction; that is, where $E(Y|X = x)$ is nondecreasing in X for an independent variable X and dependent variable Y .

The isotonic regression fit reduces the description of n points to $l \leq n$ level sets that could be used to model the population by yielding a model consisting of l more-or-less homogeneous subpopulations—the fewer the level sets, the simpler the model. Some level sets—particularly those with few elements or with small differences in step sizes from their neighbors—could be amalgamated with adjacent neighbors to improve the parsimony of the model. We describe a procedure for implementing this with a backward elimination procedure. We refer to this method and a two-sided extension of it as the reduced isotonic regression and the reduced monotonic regression methods. A similar idea was proposed by Bacchetti (1989), who developed an additive isotonic model with a dichotomous response variable.

We illustrate the new method by an example by Behm et al. (1992), which is analyzed in Section 3. Behm et al. evaluated the prognostic importance of the expression of a leukemia cell surface antigen, CD45, in 187 children diagnosed with B-lineage acute lymphocytic leukemia (ALL). Cell surface antigens are important features of the biology of a cell, because they define how a cell will inter-

act with other cells or extracellular compounds. Childhood acute leukemia is not a homogeneous disease, and investigators frequently examine differences in antigen expression to explain responsiveness to treatment (often measured by survival). The normal function of a given antigen is usually unknown. A cancer researcher wants to know whether this antigen contributes to the heterogeneity of response that is seen. Understanding the function of an antigen is a lengthy process and will probably not be undertaken by cancer researchers unless prognostic significance is found. One aspect of the analysis of CD45 was to examine its relationship to accepted prognostic factors, which may provide additional biological insight into patient heterogeneity. For example, if CD45 expression is associated with elevated white blood count, this might suggest that patients with higher CD45 counts have more aggressive disease. Patients were dichotomized at 20% CD45 expression, because the prevailing theory was that low, but nonzero expression was probably due to contaminating cells and that true expression levels should be either 0% or 100%. DNA index and white blood cell count are two prognostic factors found to be associated with the presence of CD45 ($CD45 \geq 20\%$). It is of interest to know whether the dichotomization of CD45 at 20% represents a natural cutpoint for the data.

While epidemiologists and other medical researchers often wish to divide research subjects into risk groups by categorizing some continuous variable, methods that form the categories on the basis of the specific values obtained from the data remain controversial (Altman, Lausen, Sauerbrei, and Schumacher 1994; Hilsenbeck, Clark, and McGuire 1992). One popular approach is to form the groups by maximizing the differences of some statistic between them. The predominant criticism is leveled at methods for which no adjustment of significance level is made when multiple splits are considered. Miller and Siegmund (1982) derived the asymptotic distribution of the maximal chi-squared statistic for the association of a dichotomized continuous variable and a second dichotomous variable. Hawkins

Michael J. Schell is Associate Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599. Bahadur Singh is a Visiting Fellow at the University of Missouri, Columbia, MO 65211. The first draft of the article was written while the first author was in the Department of Medicine, University of California, Irvine, CA. The authors thank Fred Behm and St. Jude Children's Research Hospital for the data used in the example, Ellen Banta for assistance with the simulation studies, Sean O'Brien for the SAS program, and the referees for helpful comments.

(1977) and Worsley (1979) presented solutions based on likelihood ratio statistics to a changepoint problem, where a continuous independent variable is dichotomized to model a single shift in mean level in a normally distributed dependent variable. The method introduced here could also be formulated as a changepoint model, where an unknown number of monotonic shifts may occur. This method, like the maximal chi-squared and likelihood ratio changepoint statistics, adjusts for the multiple comparisons involved.

In recent years methods have been described that combine the isotonic regression estimator with a smoothing function in a sequential fashion. The smoothing step confers the double advantage of reducing model complexity (by decreasing the degrees of freedom of the fit, as defined in Sec. 4) and yielding a continuous model. Friedman and Tibshirani (1984) recommended first smoothing and then isotonicizing the data; Mukerjee (1988) recommended the reverse sequence. Mammen (1991) provided theoretical results under certain conditions to determine when each approach would be better. Although a reduced monotonic regression fit is discontinuous, some comparisons of it with these alternative approaches are valuable. The reduced isotonic regression and reduced monotonic regression methods are developed in Section 2. The latter method is illustrated with a medical example in Section 3 and compared to other monotonic regression methods via simulation in Section 4.

2. REDUCED ISOTONIC REGRESSION AND REDUCED MONOTONIC REGRESSION METHODS

Let $(X_i, Y_i; w_i), i = 1, \dots, n$ denote n ordered pairs where $X_1 \leq X_2 \leq \dots \leq X_n, Y_i = \mu_i + \varepsilon_i, \mu_1 \leq \dots \leq \mu_n, w_i, i = 1, \dots, n$ are the weights, and $\varepsilon_i \sim \text{iid } N(0, \sigma^2)$. Let $n(x) \equiv E(Y|X = x)$. Thus $n(x)$ is a nondecreasing function of x . Let μ_i^* denote the isotonic regression estimator of μ_i . One tests for a statistically significant trend in the isotonic regression estimator by testing $H_0: \mu_1 = \dots = \mu_n$ versus $H_1: \mu_1 \leq \dots \leq \mu_n$, with at least one strict inequality, using the statistic

$$r_I^2 = \frac{\sum_{i=1}^n w_i (\mu_i^* - \hat{\mu})^2}{\sum_{i=1}^n w_i (Y_i - \hat{\mu})^2},$$

where $\hat{\mu} = \sum_{i=1}^n w_i Y_i / \sum_{i=1}^n w_i$. Note that r_I^2 is a special case of the likelihood ratio statistic \bar{E}_{01}^2 defined by RWD (1988, p. 63). This choice of notation emphasizes the similarity between this statistic and the usual Pearson r^2 . For the other models presented, r^2 is defined analogously to be the sum of squares for regression divided by the total adjusted sum of squares. The value of r_I^2 lies between 0 and 1 and represents the proportion of variability explained by the model. The fit is composed of a random number of level sets. The estimator is the least squares estimator over the class of nondecreasing functions. Consequently, $r_I^2 > \text{Pearson } r^2$ (assuming that the regression slope is nonnegative) and any other nondecreasing regression fit where the Y values are not transformed. Next, reduced isotonic regression, an analog of isotonic regression, is motivated and developed using a backward elimination procedure.

2.1 Backward Elimination Procedure

Consider combining level sets m and $m + 1$ for some $1 \leq m \leq l - 1$. Let

$$\mu_i^{*m} = \begin{cases} \mu_i^*, & \text{if } i \notin B_m \cup B_{m+1} \\ \frac{W_m U_m + W_{m+1} U_{m+1}}{W_m + W_{m+1}}, & \text{if } i \in B_m \cup B_{m+1}, \end{cases} \quad (1)$$

where B_1, \dots, B_l denote the ordered level sets of $\mu^* = (\mu_1^*, \dots, \mu_n^*), W_j = \sum_{i \in B_j} w_i$, and $U_j = \mu_j^*, i \in B_j$. Then the new fit will be preferred if

$$\sum_{i=1}^n w_i (\mu_i^* - \hat{\mu})^2 - \sum_{i=1}^n w_i (\mu_i^{*m} - \hat{\mu})^2 \quad (2)$$

is statistically nonsignificant. To assess significance, we first determine the distribution of Equation (2). Given (1), (2) can be simplified to

$$\begin{aligned} & \sum_{i \in B_m} w_i (\mu_i^* - \hat{\mu})^2 + \sum_{i \in B_{m+1}} w_i (\mu_i^* - \hat{\mu})^2 \\ & - \sum_{i \in C} w_i (\mu_i^{*m} - \hat{\mu})^2 \\ & = W_m (U_m - \hat{\mu})^2 + W_{m+1} (U_{m+1} - \hat{\mu})^2 \\ & - (W_m + W_{m+1}) (\hat{\mu}_m - \hat{\mu})^2 \\ & = W_m U_m^2 + W_{m+1} U_{m+1}^2 - (W_m + W_{m+1}) (\hat{\mu}_m)^2 \\ & = W_m (U_m - \hat{\mu}_m)^2 + W_{m+1} (U_{m+1} - \hat{\mu}_m)^2 \\ & = \sum_{i \in C} w_i (\mu_i^* - \hat{\mu}_m)^2, \end{aligned} \quad (3)$$

where $C = B_m \cup B_{m+1}$ and $\hat{\mu}_m = \mu_i^{*m}, i \in C$, the common mean of all Y values in C . Straightforward expansion shows that (3) = $W_m W_{m+1} \Delta^2 / (W_m + W_{m+1}) = \Delta^2 / [(1/W_m) + (1/W_{m+1})]$, where $\Delta = U_{m+1} - U_m$. Thus the statistic is larger if the difference in the level set means is larger or if the weight sizes of the level sets are larger. We can derive the distribution of (3) only under H_0 for the case $l = 2$. In this case (3) divided by σ^2 , conditional on l , is the T_{01} statistic for the data Y_i , which has a χ_1^2 distribution, as shown in lemma C of RWD (1988, p. 72). Let

$$\begin{aligned} A_m &= \sum_{i \in C} w_i (\mu_i^* - \hat{\mu}_m)^2 \\ & \div \left[\sum_{i=1}^n w_i (Y_i - \hat{\mu})^2 - \sum_{i=1}^n w_i (\mu_i^* - \hat{\mu})^2 \right] / (n - l). \end{aligned}$$

Note that under H_0

$$\sum_{i=1}^n w_i (Y_i - \hat{\mu})^2 - \sum_{i=1}^n w_i (\mu_i^* - \hat{\mu})^2 = \sum_{i=1}^n w_i (Y_i - \mu_i^*)^2 \quad (4)$$

$\sim \sigma^2 \cdot \chi_{n-l}^2$, conditional on l (RWD 1988, p. 73). Expressions (3) and (4) are independent, conditional on l , because they are functions of $\{U_j; j = 1, \dots, l\}$ and $\{Y_i - U_j; i \in B_j, j = 1, \dots, l\}$, which are independent (RWD 1988, p. 73). Thus under H_0 if $l = 2$, then A_m has an F distribution with 1 and $n - 2$ degrees of freedom conditional on l .

The backward elimination procedure to obtain the reduced isotonic regression estimator of size α is as follows:

1. Obtain the usual isotonic regression estimator with l level sets. This can be done by the pool-adjacent-violators algorithm (see RWD 1988, p. 8).
2. Compute $A = \min_m A_m, m = 1, \dots, k - 1$, where k is the current number of level sets.
3. If $A \leq F_{1,n-k;\alpha^*}$ and $A = A_r$ for some r , then combine level sets r and $r + 1$, using (1), and let $k = k - 1$.
4. Repeat Steps 2 and 3 until $A > F_{1,n-k;\alpha^*}$.

The choice of α^* is discussed later in this section. The resulting estimator is termed the reduced isotonic regression estimator. Note that the reduced isotonic regression fit is identical to the isotonic regression fit when the data perfectly observe the trend, because the denominator of A_m is 0 in that case. To obtain the reduced antitonic regression fit (where $\mu_1 \geq \dots \geq \mu_n$), multiply the Y_i by -1 before applying the backward elimination procedure. To obtain the reduced monotonic regression fit (where $\mu_1 \leq \dots \leq \mu_n$ or $\mu_1 \geq \dots \geq \mu_n$), first choose the standard isotonic or the antitonic fit, whichever has the largest r_J^2 , and then apply Steps 2–4.

Several observations can be made regarding reduced isotonic regression and reduced monotonic regression. In many applications, the investigator has strong a priori beliefs that the monotonic trend will be in a particular direction. This is usually true of dose–response relations. In other instances, however, an investigator is unsure of the directionality of the trend but still believes that a trend should exist. Indeed, the usual practice in the linear regression model is to obtain the least squares estimate for the slope and not restrict it to be nonnegative or nonpositive prior to data analysis. Reduced monotonic regression retains the monotonicity assumption of linear regression but relaxes the linearity requirement, because, as Fairley, Pearl, and Verducci (1987) noted in their title, there can be a “penalty for assuming that a monotone regression is linear.” The decision to use reduced isotonic regression or reduced monotonic regression corresponds to whether one wishes to perform a one- or two-sided test for monotonic trend. The reduced monotonic regression fit is discontinuous, even though the underlying regression will often be presumed to be continuous. Though not ideal, this property is shared by many other estimators, including the isotonic regression estimator itself, the Kaplan–Meier estimator, and the empirical distribution function. The backward elimination procedure described here is superficially identical to the variable selection procedure. However, it is quite different in one respect; for an appropriate choice of $\hat{\alpha}^*$, the existence of multiple level sets can be used as evidence of a monotonic trend at a defined α level, as described explicitly in the next paragraph.

The reduced isotonic (monotonic) regression estimator obtained depends on the choice of α^* . Using $\alpha^* = 1$ gives the usual isotonic regression estimator, whereas using $\alpha^* = 0$ gives a single level set at the overall mean of the de-

pendent variable. With the appropriate choice of α^* , one can have an α -level test of H_0 versus H_1 as defined in Section 2 (for the reduced monotonic regression, $H_1: \mu_1 \leq \dots \leq \mu_k$ with at least one strict inequality or $\mu_1 \geq \dots \geq \mu_k$ with at least one strict inequality) by rejecting H_0 if and only if the reduced isotonic (monotonic) regression fit has more than one level set. Using α^* in the backward elimination procedure does not yield an α^* -level test, because A has an exact F distribution only when $l = 2$. For instance, when $\alpha^* = .05$, the backward elimination procedure did not yield singleton level sets in 502 and 639 of 1,000 simulated samples of random noise data for $n = 50$ and 200. This finding is comparable to that of Miller and Siegmund (1982), who dichotomized a continuous variable to yield a maximal 2×2 chi-squared test. With the smaller group restricted to be at least $\frac{1}{10}$ of the entire sample, they showed that under the null hypothesis, the maximally selected chi-squared statistic will exceed the nominal 5% level 49% of the time. They determined the appropriate asymptotic critical value using Brownian bridge theory. For our application, we do not fix the number of level sets at 2 and do not know of a theoretical solution. Let us call the α^* that yields an α -level test of H_0 versus H_1 the stepwise α level.

Using simulation methods, we obtained estimates of α^* that produce α -level tests for a given α under H_0 . We applied the backward elimination algorithm to 100,000 samples of random noise data ($X_i \sim \text{iid } U[-1, 1], Y_i \sim \text{iid } N(0, 1)$, with X_i independent of Y_i) for various n to estimate α^* (denoted $\hat{\alpha}^*$) for both reduced isotonic regression and reduced monotonic regression analysis. The simulations were done using RANUNI and RANNOR in SAS. Table 1 provides $\hat{\alpha}^*$ and 95% confidence limits from the simulation study. For $\alpha = .05$, $\hat{\alpha}^*$ is the 5,000th smallest “ p value” when only two level sets remain, and the 4,865th and 5,135th lowest p values give the 95% confidence interval limits (obtained from binomial theory). Note that $\hat{\alpha}^*$ decreases rapidly for $n < 50$, but flattens out considerably for $n > 50$. For $\alpha = .05$ and $20 \leq n \leq 800$, $\hat{\alpha}^* \approx .012 * n^{-.361}$. In no case was a simulation dataset found to have both an

Table 1. Simulated $\hat{\alpha}^*$ for the Reduced Isotonic Regression and Reduced Monotonic Regression Estimators for Selected α

n	$\alpha = .10$		$\alpha = .05$		$\alpha = .01$	
	$\hat{\alpha}^*$	diff	$\hat{\alpha}^*$	diff	$\hat{\alpha}^*$	diff
<i>Reduced Isotonic Regression</i>						
10	.0316	.0005	.0136	.0003	.00234	.00013
20	.0219	.0003	.0095	.0002	.00158	.00007
50	.0147	.0002	.0062	.0001	.00099	.00005
200	.0092	.0002	.0038	.0001	.00058	.00003
800	.0063	.0001	.0026	.0001	.00041	.00002
<i>Reduced Monotonic Regression</i>						
10	.0136	.0003	.0062	.0002	.00113	.00007
20	.0095	.0002	.0042	.0001	.00074	.00005
50	.0062	.0001	.0028	.0001	.00047	.00003
200	.0038	.0001	.0017	.0001	.00028	.00002
800	.0026	.0001	.0011	.0000	.00020	.00001

NOTE: Adding and subtracting diff from $\hat{\alpha}^*$ comprises at least a 95% confidence interval for α^* .

Table 2. Comparison of Methods for the CD45 Example

Method	LOGWBC		DI	
	r^2	df	r^2	df
I	.173	11	.327	8
R	.159	3	.286	2
S	.194	11.4	.288	3.6
SI	.137	6.2	.282	3.6
IS	.108	3.3	.273	3.3
L	.094	2	.284	2
CUT20	.066	2	.209	2

isotonic and antitonic fit with two or more level sets when the $\hat{\alpha}^*$ corresponding to $\alpha \leq .05$ was used. Thus for $10 \leq n \leq 800$, reduced monotonic regression can be regarded as a two-sided version of reduced isotonic regression, and the $\hat{\alpha}^*$ giving an α -level test for monotonic regression will give an $\alpha/2$ -level test for reduced isotonic regression. Throughout the remainder of this article, we use the estimated stepwise α level where $\alpha = .05$. We now illustrate the use of reduced monotonic regression with a medical example.

3. AN EXAMPLE

The data for this example have been described in Section 1. It was reasonable to assume that the relationship between the percent of expression of CD45 and both the logarithm

of the white blood count (LOGWBC) and DNA index (DI) would be monotonic, but it was unknown whether low or high values of CD45 would be associated with high LOGWBC. Thus we performed reduced monotonic regression (R) for LOGWBC and DI on CD45 expression to assess the choice of 20% CD45 expression as a cutpoint. The percentage of leukemia cells from an individual patient that express CD45 ranged from 0% to 99%, with a median of 85%. We dropped three outliers from DI before analyzing the data. Increasing CD45 percentage was associated with increasing LOGWBC and decreasing DI. Because expression was obtained in percent units, numerous ties exist for CD45. Thus we applied reverse secondary sorts to the two regressions to prevent the isotonicization step from having multiple level sets in a single CD45 value. Because there are 68 distinct CD45 values, one might expect the stepwise α level $\hat{\alpha}^* \approx .0026$ for $\alpha = .05$, using the formula in Section 2. A simulation study of 1,000 random noise datasets, using the exact groupings of CD45 values, yielded the estimate $\hat{\alpha}^* = .0034$. The strata identified for LOGWBC (Fig. 1a) were .88 for $CD45 \leq 46$ ($n = 46$), 1.18 for $48 \leq CD45 \leq 96$ ($n = 117$), and 1.66 for $CD45 \geq 97$ ($n = 24$), yielding $r^2 = .159$, whereas $r^2 = .066$ when the data were dichotomized at 20%. The strata for DI (Fig. 1b) were 1.14 for $CD45 \leq 56$ ($n = 54$) and 1.03 for $CD45 \geq 57$ ($n = 130$), giving $r^2 = .286$, compared to $r^2 = .209$ for the split at

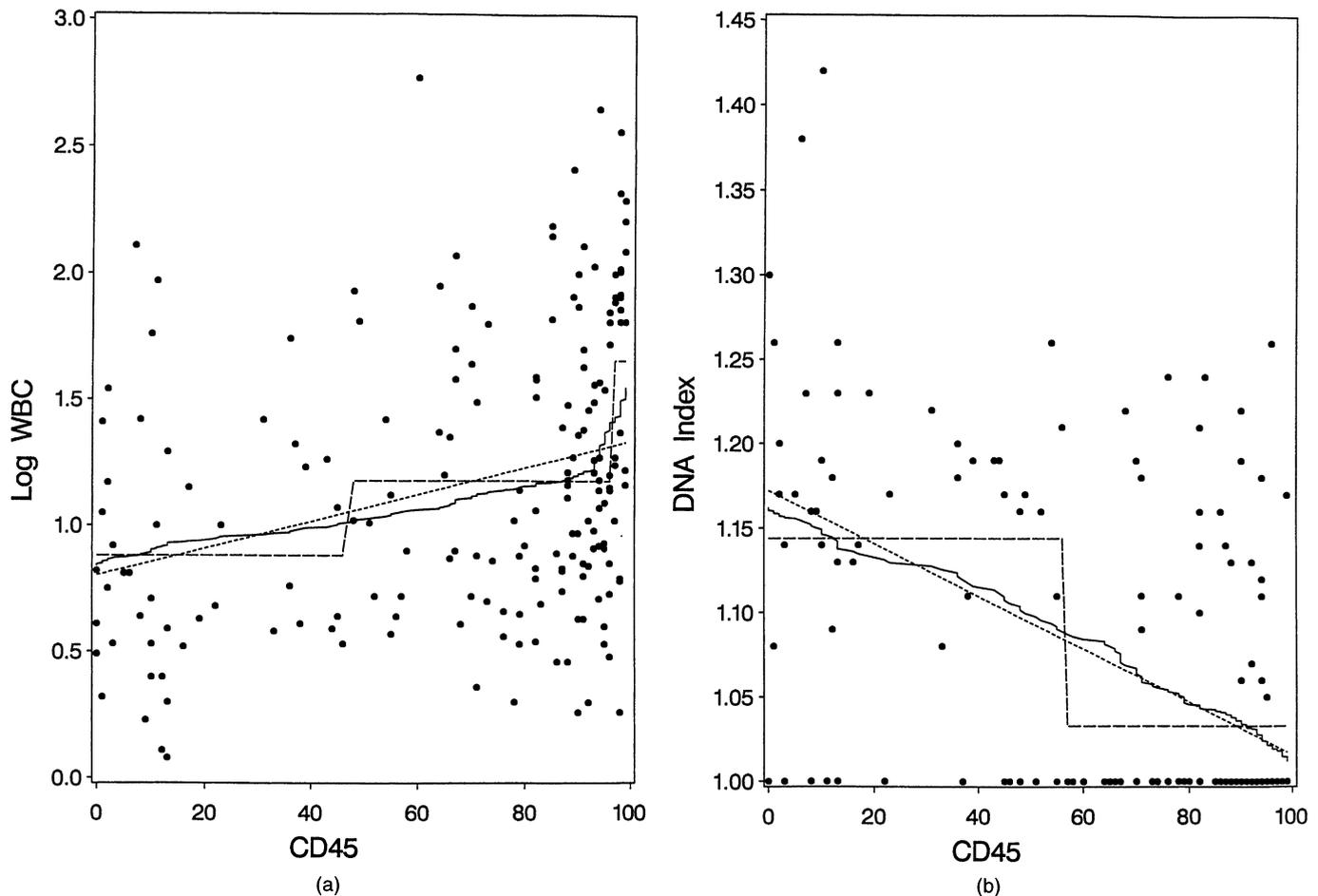


Figure 1. The Functions for Linear Regression (Short-Dashed Line), Isotonized-Then-Smoothed Regression (Solid Line), and Reduced Monotonic Regression (Long-Dashed Line), Along With the Data for the Regression of (a) LOGWBC and (b) DI on CD45.

Table 3. Estimated Minimum Significant r^2 Values at $\alpha = .05$

Method	r^2				
	$n = 10$	$n = 20$	$n = 50$	$n = 200$	$n = 800$
S	.692	.601	.356	.099	.028
I	.729	.457	.225	.070	.021
R	.633	.376	.173	.049	.013
SI	.545	.326	.140	.036	.009
IS	.459	.244	.091	.024	.006
L	.399	.197	.078	.019	.005

NOTE: Adding and subtracting .014, .009, .005, .004, and .001 from r^2 for $n = 10, \dots, 800$ comprises at least a 95% confidence interval (except that .018 is needed for S when $n = 50$). The results for I and L are exact.

20%. The p values for combining the remaining adjacent level sets, estimated from the simulations, are $p = .021$ and $.002$ for LOGWBC and $p < .001$ for DI. These regression functions challenge the theory that patients should be dichotomized at roughly 20% CD45 expression. In fact, because we believe that a simple parametric model should be preferred when it yields an adequate fit, the linear fit is preferred for the regression of DI on CD45.

For comparative purposes, we examined five other regression methods: isotonic regression (I), smoothed regression (S), smoothed-then-isotonized regression (SI), isotonized-

then-smoothed regression (IS), and linear regression (L). Although we are examining monotonic trends here (except possibly for S), we retain the word "isotonic" for these other methods because the term is well established for them. (Note: The IS regression function is not necessarily monotone for an arbitrary smoother.) The smoother that we used was the running mean smoother with the span size determined by the cross-validation method described by Friedman and Tibshirani (1984). The running mean smoother is a less than ideal choice of smoother when many ties exist in the data, as is the case here. It does permit easy computation of degrees of freedom (see Sec. 4.1), which is difficult for many smoothers. Using an approximate F -test method to compare regression models, as recommended by Hastie and Tibshirani (1990, sec. 3.9), R outperformed the other regression models in fitting both LOGWBC and DI, although DI is well modeled using linear regression and should be favored in practice (see Table 2). Figure 1 shows the fits for R, L, and IS. Table 2 also shows the inadequacy of CUT20, where the data are dichotomized at 20%.

4. COMPARISON OF REGRESSION METHODS

In this section we compare the reduced monotonic regression estimator developed in Section 2 to the other five

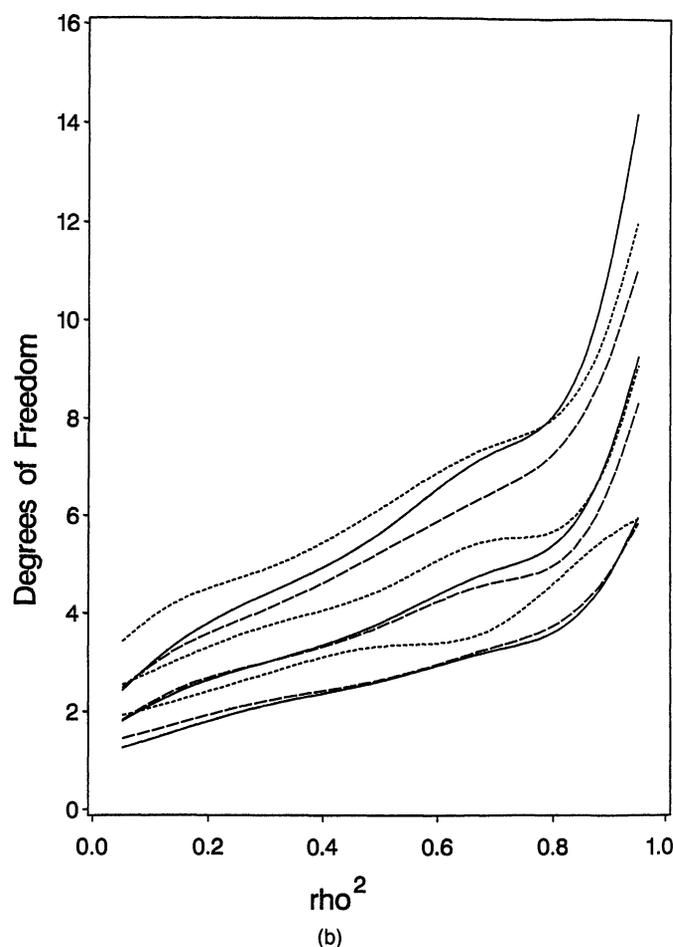
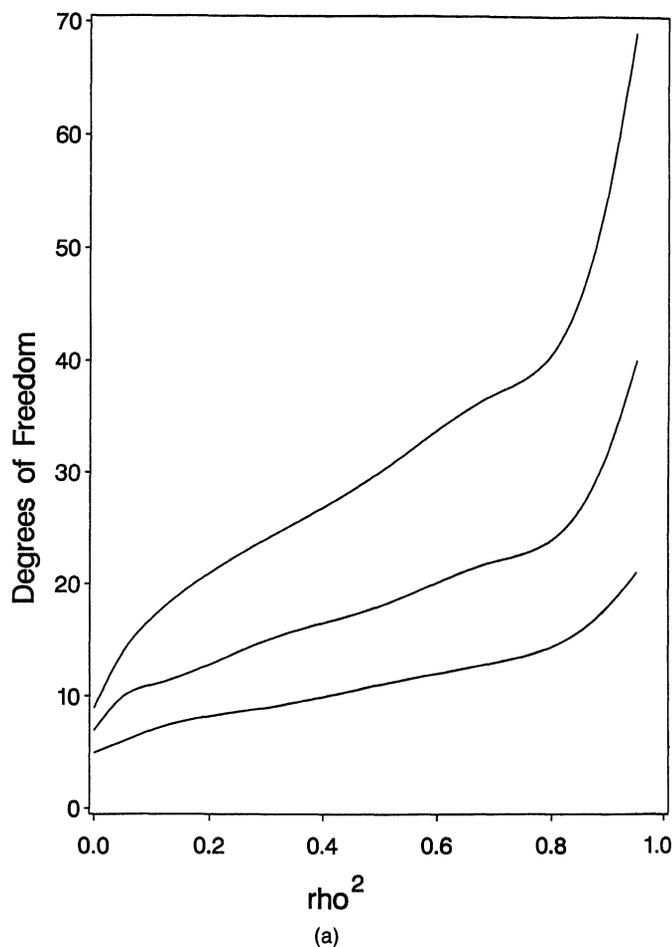


Figure 2. Plot of the Coefficient of Determination, ρ^2 , Versus the Median Degrees of Freedom for the Linear Statistical Relation of the (a) I (Solid Line) and (b) R (Solid Line), SI (Short-Dashed Line), and IS (Long-Dashed Line) Regression Methods for $n = 50$ (Bottom Set), 200 (Middle Set), and 800 (Top Set), Based on 1,000 Simulations. Because the median degrees of freedom for R are a small integer, the mean was used instead to provide smoother curves.

Table 4. Median df/df_I Ratios (%)

Method	Overall			$\rho^2 = .30$			$\rho^2 = .95$		
	<i>n</i>			<i>n</i>			<i>n</i>		
	50	200	800	50	200	800	50	200	800
S	35	27	22	34	28	24	36	27	20
SI	33	25	20	32	26	22	33	24	18
IS	28	22	17	26	22	18	30	21	16
R	26	22	20	24	22	18	29	23	20

NOTE: The overall ratio includes $.3 \leq \rho^2 \leq .95$.

regression methods described in Section 3 for the example. Using 10,000 samples of random noise data as described in Section 2, Table 3 shows the estimated minimum r^2 values that are significant at $\alpha = .05$ for the different regression methods and sample sizes, with the exact values for isotonic regression obtained using methods described by Kudo and Yao (1982). This table shows one disadvantage of increased model flexibility. I and S have much higher minimum r^2 values that are statistically significant, particularly for small sample sizes. On the other hand, L has significantly lower limits than other methods and would be preferable when the true correlation is very weak and consequently could not be demonstrated by the other methods.

We carried out a Monte Carlo experiment to compare the six regression methods for $n = 50, 200$, and 800 . (Recall that for reduced monotonic regression, we are using $\hat{\alpha}^*$ to give a .05-level test.) We examined five statistical relations: linear ($Y = aX + \varepsilon$), mean shift ($Y = \varepsilon$ if $X < 0$; $Y = a + \varepsilon$ if $X > 0$), elbow ($Y = \varepsilon$ if $X < 0$; $Y = aX + \varepsilon$ if $X > 0$), sigmoidal ($Y = a\Phi(2.5X) + \varepsilon$), and exponential ($Y = ae^X + \varepsilon$), where $X \sim U[-1, 1]$, $\varepsilon \sim N(0, 1)$, Φ is the Gaussian cumulative distribution function and a is determined for each statistical relation and n so that the coefficient of determination, ρ^2 , would be .05, .10(.20).90, and .95. We simulated each situation 1,000 times, and used two measures to compare the regression methods: $r^2 - \rho^2$ to assess overfitting or underfitting by the model, and degrees of freedom of the fit (as described below in the next section), which is a measure of model complexity.

4.1 Degrees of Freedom

Let $\hat{f}(x_i), i = 1, \dots, n$ be called a regression estimator or a smoother. If $\hat{f} = \mathbf{S}\mathbf{y}$ for some $n \times n$ matrix $\mathbf{S} = \{S_{ij}\}$ and $\mathbf{y} = (y_1, \dots, y_n)^T$, then we call \mathbf{S} a smoother matrix and define the degrees of freedom (df) of the estimator to be $df = \text{tr}(\mathbf{S})$ (Hastie and Tibshirani 1990, p. 52). For linear regression, $df = 2$. For the isotonic regression estimator, df depends on the data. Suppose that the fit is composed of l level sets with p_k points in level set k . Then $S_{ij} = 1/p_k$ if i and j are in level set k , and $S_{ij} = 0$ otherwise. Thus $df = \text{tr}(\mathbf{S}) = \sum_{i=1}^n S_{ii} = \sum_{k=1}^l p_k \cdot (1/p_k) = l$, the number of level sets. Similarly, df for \hat{f}_R , the reduced monotonic fit, is the resultant number of level sets after the backward elimination.

4.2 Method Comparisons

Let df_I denote the estimated median number of degrees of freedom for the isotonic regression method, with similar terms for the other regression methods. The df_I increases as ρ^2 and n increase, as shown in Figure 2a for the linear statistical relation. The curves are similar for the other statistical relations, with the degrees of freedom being nearly identical for the exponential relation and 80%–92% as high for the sigmoidal and the elbow relations. (The mean shift relation is described separately, as the methods behave differently with it.) The S, SI, IS, and R estimators follow a similar pattern (Fig. 2b) but require only one-fifth to one-third of the degrees of freedom as I requires. R and IS have the lowest df/df_I ratios, given in Table 4. Trends in the ratios can also be seen as a function of ρ^2 .

Figure 3 shows the median difference of $r^2 - \rho^2$ for three of the statistical relations (linear, sigmoidal, and exponential). Because all $r^2 - \rho^2 > 0$, the methods “overfit” the data, with the overfitting decreasing by roughly one-half when the sample size is quadrupled. However, the other nonparametric methods have at most one-half of the overfitting found for I. The median r^2/ρ^2 values for linear regression are 1, .98, .94, and .80 for the linear, sigmoidal, exponential, and elbow statistical relations, regardless of n and ρ^2 . Thus, except for the first two statistical relations, L is not competitive with the other estimators considered here.

A good procedure for modeling a statistical relation should yield a fit where $r^2 \approx \rho^2$ under correct model specification and that is parsimonious; that is, requiring few df for the fit. Because the nonparametric methods all overfit the data, the desired method will have the lowest r^2 and lowest df. In this regard IS, SI, and R are all preferable to S and greatly outperform I. Thus we now restrict our comparisons to IS, SI, and R. IS has the least amount of overfitting, whereas SI is generally comparable to or slightly better than R. R and IS have remarkably similar df for $n = 50$ and 200 for $\rho^2 \leq .5$. For larger ρ^2 and n , $df_{IS} < df_R$. SI has the highest median df, except that $df_{SI} < df_R$ for $\rho^2 \geq .90$ when $n = 200$ and for $\rho^2 \geq .50$ when $n = 800$ for sigmoidal data and less frequently for linear and elbow data. Even in these exceptional cases, the 90th percentile df values exceed the median values by 1 for R but by at least 6.4 for SI. It is likely that some other choice of smoother would improve SI by decreasing df and the overfitting for the statistical relations considered here. Final preference between R, IS, and SI will also depend on such issues as continuity, which SI (where one joins the response values obtained for successive points after the isotonic step) and IS enjoy, compared to the simplicity of the fit, which favors R.

4.3 Mean Shift Statistical Relation

The mean shift statistical relation is ideally suited to the reduced monotonic regression estimator; $df_R = 2$ for all n and ρ^2 . In fact, for $\rho^2 \geq .30$, at least 96% of the samples result in two level sets. In contrast, df_I is roughly 8, 10, and 13 for $n = 50, 200$, and 800 . The df_{IS}/df_I and df_{SI}/df_I ratios increase from 20% to 60% and from 30% to 70% as

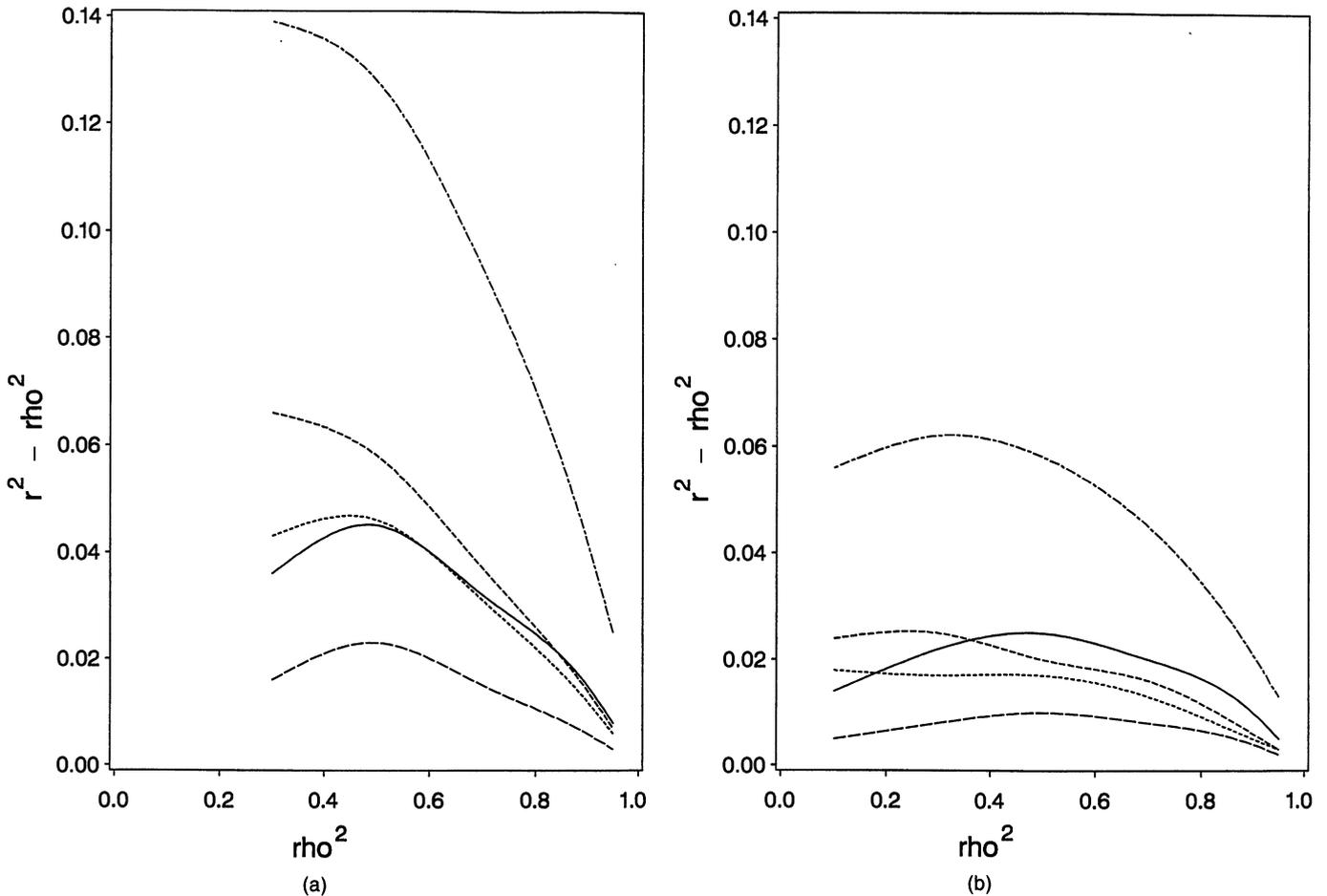


Figure 3. Overfitting (Defined as Median $r^2 - \rho^2$) of the Nonparametric Methods for the Linear, Sigmoidal, and Exponential Statistical Relations as a Function of ρ^2 , the True Coefficient of Determination, for (a) $n = 50$ and (b) $n = 200$, Based on 1,000 Simulations. (-----) L; (——) IS; (---) S; (-.-) SI; (——) R.

ρ^2 increases from .05 to .95, regardless of n . The df_5 values are even higher, with values of 13, 34, and 67 for $n = 50, 200,$ and 800 when $\rho^2 = .95$. IS underfits the mean shift statistical relation by roughly .04, .02, and .01 for $n = 50, 200,$ and 800 , whereas SI underfits by one-half that amount for larger ρ^2 .

5. DISCUSSION

We have introduced two new regression methods—reduced isotonic regression and reduced monotonic regression—to model data presumed to be isotonic (monotonic). The methods apply a backward elimination procedure to isotonic regression, greatly reducing the number of level sets obtained. The resulting number of levels depends on the choice of α^* in the backward elimination procedure. When α^* is the stepwise α level, a single level set is obtained with probability $1 - \alpha$ under the null hypothesis of constant mean. Thus it discretizes a continuous dependent variable with a simple model with minimal overfitting, a desirable feature in many applications. Another attractive feature of the reduced monotonic regression method is that it is invariant to a monotonic transformation of the independent random variable, a property not shared by most nonparametric regression methods.

Bacchetti (1989, p. 292) introduced the additive isotonic model, which is comprised of multiple isotonic independent variables. Bacchetti saw the need “to pool blocks even when there are no more violators, as long as each such pool does not increase the criterion by more than a specified amount” to yield a simpler model not as prone to overfitting, and he proposed an ad hoc solution. Extension of the backward elimination procedure to his method could provide a more solid basis for combining blocks for this model.

A strong connection exists between the reduced monotonic regression method and the likelihood ratio change-point statistic. Worsley (1979) derived the distribution of the likelihood ratio statistic for the single changepoint problem for a sequence of n independent normally distributed random variables with mean μ_i and common but unknown variance σ^2 . The changepoint problem that he tested was $H_0: \mu_i = \mu, i = 1, \dots, n$ versus $H_1: \mu_i = \mu, i = 1, \dots, k$ and $\mu_i = \mu' \neq \mu, i = k + 1, \dots, n$, where $\mu, \mu',$ and k are unknown. Worsley provided exact and approximate critical values of a t distribution with $n - 2$ df for $n \leq 50$. The tail probabilities associated with these critical values are nearly identical to those in Table 1. For $n = 10, 20,$ and 50 , Worsley obtained the values .0138, .0096, and .0060 for $\alpha = .10$; .0064, .0042, and .0028 for $\alpha = .05$; and .00114, .00064,

and .00042 for $\alpha = .01$. The similarity between his results and ours makes sense when one recalls that under H_0 , backward elimination will usually ultimately result in a test of a single shift in mean level. The reduced monotonic regression method could be used to fit a changepoint model, with the advantage that multiple monotonic shifts in mean are allowed.

We have performed simulations comparing the fits of five monotonic regression methods for several monotonic statistical relations, with the goal of achieving a parsimonious fit where the $r^2 \approx \rho^2$, the true coefficient of determination. Overall, the reduced monotonic regression (R) and isotonized-then-smoothed regression (IS) methods performed best, when a cross-validated running mean smoother was used. Linear regression underfits nonlinear statistical relations, resulting in poor comparative performance. The other four methods overfit the data somewhat, but the degree of overfitting declines with increasing sample size. Isotonic regression (I) substantially overfits the data and requires many degrees of freedom. IS performed consistently better than smoothed-then-isotonized regression (SI). Overall, R performed very well, except that the degrees of freedom are high for large n and ρ^2 . Only R fit the mean shift data well. The regression of LOGWBC on CD45 represents a practical situation involving a rapid shift in response level. In that instance IS underfit the data, and SI required many degrees of freedom; R avoided both problems and provided the best fit.

Different choices of smoother (e.g., loess, kernel, smoothing spline) should be examined for IS and SI. We would expect a judicious choice to reduce the degrees of freedom and overfitting seen here. We used the running mean smoother due to the ease of calculating the degrees of freedom of the estimator when combining it with the isotonization step and because the fit is invariant with respect to transformations of the independent variable.

The reduced monotonic regression method is a powerful and flexible method for modeling monotonic data. Given its many attractive features and simplicity, the reduced mono-

tonic regression method could prove to be a valuable addition to those currently used by applied statisticians. SAS or FORTRAN programs for performing reduced isotonic and reduced monotonic regression are available from the first author on request.

[Received April 1994. Revised May 1996.]

REFERENCES

- Altman, D. G., Lausen, B., Sauerbrei, W., and Schumacher, M. (1994), "Dangers of Using 'Optimal' Cutpoints in the Evaluation of Prognostic Factors," *Journal of the National Cancer Institute*, 86, 829-835.
- Bacchetti, P. (1989), "Additive Isotonic Models," *Journal of the American Statistical Association*, 84, 289-294.
- Behm, F. G., Raimondi, S. C., Schell, M. J., Look, A. T., Rivera, G. K., and Pui, C.-H. (1992), "Lack of CD45 Antigen on Blast Cells in Childhood Acute Lymphoblastic Leukemia is Associated With Chromosomal Hyperdiploidy and Other Favorable Prognostic Features," *Blood*, 77, 1011-1016.
- Fairley, D., Pearl, D. K., and Verducci, J. S. (1987), "The Penalty for Assuming That a Monotone Regression is Linear," *The Annals of Statistics*, 15, 443-448.
- Friedman, J., and Tibshirani, R. (1984), "The Monotone Smoothing of Scatterplots," *Technometrics*, 26, 243-250.
- Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.
- Hawkins, D. M. (1977), "Testing a Sequence of Observations for a Shift in Location," *Journal of the American Statistical Association*, 72, 180-186.
- Hilsenbeck, S. G., Clark, G. M., and McGuire, W. L. (1992), "Why Do So Many Prognostic Factors Fail to Pan Out?" *Breast Cancer Research and Treatment*, 22, 197-206.
- Kudo, A., and Yao, J. S. (1982), "Tables for Testing Ordered Alternatives in an Analysis of Variance Without Replications," *Biometrika*, 69, 237-238.
- Mammen, E. (1991), "Estimating a Smooth Monotone Regression Function," *The Annals of Statistics*, 19, 724-740.
- Miller, R., and Siegmund, D. (1982), "Maximally Selected Chi Square Statistics," *Biometrics*, 38, 1011-1016.
- Mukerjee, H. (1988), "Monotone Nonparametric Regression," *The Annals of Statistics*, 16, 741-750.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988), *Order-Restricted Statistical Inference*, New York: Wiley.
- Worsley, K. J. (1979), "On the Likelihood Ratio Test for a Shift in Location of Normal Populations," *Journal of the American Statistical Association*, 74, 365-367.