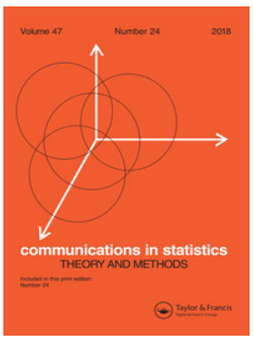

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

Robust regression using iteratively reweighted least-squares

Paul W. Holland & Roy E. Welsch

To cite this article: Paul W. Holland & Roy E. Welsch (1977) Robust regression using iteratively reweighted least-squares, Communications in Statistics - Theory and Methods, 6:9, 813-827, DOI: 10.1080/03610927708827533

To link to this article: https://doi.org/10.1080/03610927708827533

Published online: 27 Jun 2007.

Submit your article to this journal

Article views: 2032

Citing articles: 843 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=lsta20

COMMUN. STATIST.-THEOR. METH. A6(9), 813-827 (1977)

ROBUST REGRESSION USING ITERATIVELY REWEIGHTED LEAST-SQUARES

Paul W. Holland

Educational Testing Service
Princeton, NJ

Roy E. Welsch

Massachusetts Institute of Technology
National Bureau of Economic Research
Cambridge, MA

Key Words & Phrases: ROSEPACK; robust weight functions; tuning constants; small sample variances; SLASH distribution; one-step estimates.

ABSTRACT

The rapid development of the theory of robust estimation (Huber, 1973) has created a need for computational procedures to produce robust estimates. We will review a number of different computational approaches for robust linear regression but focus on one--iteratively reweighted least-squares (IRLS). The weight functions that we discuss are a part of a semi-portable subroutine library called ROSEPACK (RObust Statistical Estimation PACKage) that has been developed by the authors and Virginia Klema at the Computer Research Center of the National Bureau of Economic Research, Inc. in Cambridge, Mass. with the support of the National Science Foundation. This library (Klema, 1976) makes it relatively simple to implement an IRLS regression package.

Copyright © 1977 by Marcel Dekker, Inc. All Rights Reserved. Neither this work nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage and retrieval system, without permission in writing from the publisher.

INTRODUCTION

We will consider the standard regression model

$$y = X\beta + \varepsilon \ , \tag{1}$$

where y and ε are n by 1, X is n by p, and β is p by 1. A robust estimate for β, $\hat{\beta}$, minimizes

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - x_i\beta}{\sigma}\right) \ , \tag{2}$$

where σ is a known or previously estimated scale parameter, ρ is a robust loss function, and x_i is the ith row of X. If we let $\psi = \rho'$, then a necessary condition for a minimum is that $\hat{\beta}$ satisfy

$$\sum_{i=1}^{n} x_{ij}\, \psi\left(\frac{y_i - x_i\beta}{\sigma}\right) = 0 \text{ for all } j. \tag{3}$$

In general (3) is a set of nonlinear equations and iterative methods are required.

We define the weight function $w(r)$ as $\psi(r)/r$, and let $< >$ denote an n by n diagonal matrix. If we have a starting value $\hat{\beta}_o$ then there are three fairly obvious iteration schemes for finding the solution to (3):

$$\hat{\beta}_1 = \hat{\beta}_o + \sigma\left(X^T<\psi'\left(\frac{y-X\hat{\beta}_o}{\sigma}\right)>X\right)^{-1} X^T \psi\left(\frac{y-X\hat{\beta}_o}{\sigma}\right) \ , \tag{4}$$

$$\hat{\beta}_1 = \hat{\beta}_o + \sigma(X^TX)^{-1} X^T \psi\left(\frac{y-X\hat{\beta}_o}{\sigma}\right) \ , \tag{5}$$

$$\hat{\beta}_1 = \hat{\beta}_o + \left(X^T<w\left(\frac{y-X\hat{\beta}_o}{\sigma}\right)>X\right)^{-1} X^T<w\left(\frac{y-X\hat{\beta}_o}{\sigma}\right)>(y-X\hat{\beta}_o) \ . \tag{6}$$

The first is Newton's method, the second has been discussed extensively by Huber (1975) and Bickel (1975) and the third, reweighted least-squares, is generally attributed to Beaton and Tukey (1974).

Newton's method is theoretically the most desirable, but it is difficult to implement because it requires ψ'; also $X^T<\psi'>X$ may be negative definite. The Huber method has desirable

computational properties since $(X^TX)^{-1}X^T$, the generalized inverse, need only be computed once. It does, however, require more iterations and is not as easy to use with existing least-squares regression packages.

The third method only requires knowing how to compute the weight function, w(r), and then it is possible to use an existing weighted least-squares algorithm or to compute the square root of w(r), form $<w^{1/2}>X$ and $<w^{1/2}>y$ and use a standard least-squares program for each step. The IRLS method generally converges somewhat faster than the Huber method, but more slowly than Newton's method.

With nonconvex ρ functions we can only hope for convergence to a local minimum and therefore a good starting value is important. We propose using the least absolute residuals estimator, $\hat{\beta}_L$ which minimizes

$$\sum_{i=1}^{n} |y_i - x_i\beta| \quad . \tag{7}$$

This can be obtained from algorithms described by Barrodale and Roberts (1973) and Bartels (1975).

Clearly a question arises about estimating σ. We have done this by using

$$\sigma = 1.48 \left[\underset{i}{\text{med}} \left|(y_i - x_i\hat{\beta}_o) - \underset{i}{\text{med}}\ (y_i - x_i\hat{\beta}_o)\right|\right] \tag{8}$$

just once, before starting the iterations. (The factor 1.48 makes this an approximately unbiased estimate of scale when the error model is Gaussian.) We do not iterate scale because there is no convergence theory when the scale is iterated except for the Huber loss function (Huber, 1975). If scale is to be iterated and/or a least-squares start is used, the best procedure is to iterate to convergence with the Huber ρ function and the Huber scaling and then use a nonconvex ρ function without further scale iteration. Hill and Holland (1977) discuss other aspects of σ defined in (8).

We will not discuss convergence criteria in this paper. Useful references are Dennis and Welsch (1976) and Huber (1975).

WEIGHT FUNCTIONS

After some years of experience with robust regression and a survey of the literature, we chose to restrict our attention to the eight weight functions given in Table I. They can be classified by the behavior of $\psi(t)$. The hard redescenders; A (Andrews et al., 1972), B (Beaton and Tukey, 1974), and T (Hinich and Talwar, 1975), all have $\psi(r) = 0$ for $|r|$ sufficiently large. The

TABLE I

Weight Functions

Name	$\rho(r)$	$\psi(r)$	$w(r)$	Range
A	$A^2[1-\cos(r/A)]$	$A \sin(r/A)$	$(r/A)^{-1}\sin(r/A)$	$\lvert r\rvert \le \pi A$
	$2A^2$	0	0	$\lvert r\rvert > \pi A$
B	$(B^2/2)[1-[1-(r/B)^2]^3]$	$r[1-(r/B)^2]^2$	$(1-(r/B)^2)^2$	$\lvert r\rvert \le B$
	$B^2/2$	0	0	$\lvert r\rvert > B$
T	$r^2/2$	r	1	$\lvert r\rvert \le T$
	$T^2/2$	0	0	$\lvert r\rvert > T$
C	$(C^2/2)\log[1+(r/C)^2]$	$r[1+(r/C)^2]^{-1}$	$[1+(r/C)^2]^{-1}$	
W	$(W^2/2)[1-\exp[-(r/W)^2]]$	$r \exp[-(r/W)^2]$	$\exp[-(r/W)^2]$	
H	$r^2/2$	r	1	$\lvert r\rvert \le H$
	$H\lvert r\rvert - H^2/2$	$\operatorname{sgn}(r)\, H$	$H\lvert r\rvert^{-1}$	$\lvert r\rvert > H$
L	$L^2 \log[\cosh(r/L)]$	$L \tanh(r/L)$	$(r/L)^{-1}\tanh(r/L)$	
F	$F^2[\lvert r\rvert/F-\log(1+\lvert r\rvert/F)]$	$r(1+\lvert r\rvert/F)^{-1}$	$(1+\lvert r\rvert/F)^{-1}$	

soft redescenders, C (Cauchy or t-likelihood) and W (Dennis and Welsch, 1976) are asymptotic to zero for large $|r|$. The last three, H (Huber, 1964), L (logistic), and F (Fair, 1974) have monotone ψ functions.

With some abuse of notation we will use the same letters to refer to the ψ-function and to the "tuning constants" for each ψ-function. Tuning constants may be used to adjust the efficiency of the resulting estimators for specific distributions.

Functions A and B are essentially the same except that ψ' is continuous for B but not for A. The same comment applies to L and H respectively. The function F was created as an approximation to least absolute residuals (LAR) for nonlinear models but can be useful for linear problems as well. Performing LAR regression using iteratively reweighted least-squares (Armstrong and Frome, 1976) is not recommended. The algorithms described for finding $\hat{\beta}_L$ in (7) are better.

The C class is easily seen to give the maximum likelihood estimates for t distributions where the degrees of freedom can be related to the tuning constant, C. We should note that H and L can also be derived using maximum likelihood, but we will not discuss this aspect of robust estimation here.

The function W was developed mainly to simplify certain theoretical computations relating to the asymptotic efficiencies of robust estimators. Recent work by Schumaker and Paulson (1976) on characteristic functions may give it a more important theoretical foundation.

For monotone ψ-functions (H, L, and F), Huber (1973) has shown that under reasonable conditions fully iterated estimators defined by (2) are asymptotically Gaussian with covariance

$$V(F,\psi) = \frac{E_F(\psi^2)}{[E_F(\psi')]^2} \times (X^TX)^{-1} \quad , \tag{9}$$

when the underlying population has a cumulative distribution function, F. For nonmonotone ψ-functions, much less is known, but Collins (1976) gives some results. Monte Carlo studies of the

covariance matrix are discussed in Welsch (1975) and Gross (1977). In order to provide some guidance about the use of the eight weight functions, we have listed in Table II the values of the tuning constants such that the asymptotic efficiency in the location model, $nV(F,\psi)^{-1}$, is 95% when F is the unit Gaussian distribution.

TABLE II

Tuning Constants for 95% Asymptotic Efficiency at the Gaussian Distribution

Weight Function	A	B	C	F	H	L	T	W
Tuning Constant	1.339	4.685	2.385	1.400	1.345	1.205	2.795	2.985

Bickel (1975) has shown that the same asymptotic formulas apply in most cases to estimators which take one Newton step (4) or one Huber step (5) from a consistent starting value such as (7). These results do not appear to apply to one IRLS step (6) and we would expect these one-step estimators to be less efficient than the Newton one-step. The T estimator is a special case, since $\psi' = w$ and Table IV (in the next section) shows a high one-step efficiency.

In order to obtain asymptotic efficiencies at the unit Gaussian we needed to compute integrals of the form

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi^2(x)\, e^{-x^2/2} dx \tag{10}$$

and

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi'(x)\, e^{-x^2/2}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \psi(x)\, xe^{-x^2/2}\, dx. \tag{11}$$

Only the ψ' for T requires some care in evaluating the jump discontinuities and this can be avoided by using the right-hand side of (11).

The B, H, T, and W estimators have formulas for their asymptotic efficiencies which may be expressed in terms of incomplete chi-square (or Gamma) functions and Gaussian integrals and may, therefore, be evaluated fairly easily.

The other four estimators--A, C, F, L--required numerical integration in order to compute their asymptotic efficiencies. Step-function approximations were used with a sufficiently fine grid so that the accuracy indicated in Table II was obtained. The Monte Carlo discussed in the next section provides a crude check on these tuning constant values.

MONTE CARLO

A small Monte Carlo study was performed to check the values of the tuning constants given above. For three sample sizes ($n = 10$, 20 and 40) we obtained Gaussian efficiencies for location estimates based on these w-functions. (The $\text{eff}_E = \text{Var}(\bar{X})/\text{Var}(E)$ where E is the estimate being studied and $\bar{X}$ is the sample mean.) These efficiencies were computed for 1 through 5 iterations of reweighted least-squares using each of the eight w-functions. In these iterations σ was not changed from iteration to iteration to insure convergence.

Table III gives the results of this Monte Carlo study. We see that multiple iterations improve the efficiency of these estimators in the desired direction (i.e., increases them towards the $n = \infty$ value of 95%) but even after convergence (which is operationally defined here as five iterations) there is still wide disparity among the efficiencies of the eight estimators, and none of these exceed 94% for $n = 10$. The five iteration data from Table III is graphed in Figure 1. This graph shows the disparity in the efficiencies of the fully iterated estimates for $n = 10$, 20 and 40. While there is a clear trend for the efficiencies to increase towards the asymptotic value of 95% as $n \to \infty$, even by $n = 40$ only one of them (F) is over 94% efficient. The eight curves in Figure 1 group neatly if we examine them as we did earlier by the type of ψ-function that motivates each w-function.

The first impression one gets from Figure 1 is that the w-functions approach 95% efficiency at different rates as $n \to \infty$. But the other possibility is that they are differentially sensitive to the estimated scale. To test this, we ran a second Monte Carlo

study with $\sigma^2 = 1$ which was the correct variance for the Gaussian variables used in the Monte Carlo. The results are contained in Table IV. This allowed us to see the effect of estimating σ compared to knowing the correct scale. Figure 2 summarizes the results of that study for the fully iterated estimators (i.e., after 5 iterations). Except for estimator T at n = 10, all of the estimators are very nearly 95% efficient at all values of n. Hence we conclude that the variation exhibited across the eight estimators in Figure 1 is primarily due to differential sensitivity to the effect of estimating the scale. In general, the effect of estimating the scale has been swept under the rug in previous studies of robust estimation and perhaps these results will bring attention to the

TABLE III

Gaussian Efficiencies with Estimated Scale

Iteration	Weight Function								
	A	B	C	F	H	L	T	W	M
	n=10 (1000 replications)								
1	85.5	85.7	88.2	87.8	90.3	88.8	83.9	86.8	70.2
2	86.8	87.0	90.8	92.0	91.9	91.7	83.7	88.6	
3	87.0	87.2	91.3	93.1	92.2	92.3	83.7	89.0	
4	87.0	87.3	91.5	93.4	92.3	92.5	83.7	89.1	
5	87.1	87.3	91.5	93.5	92.3	92.5	83.7	89.1	
	n=20 (1500 replications)								
1	88.3	88.4	89.5	88.1	91.4	89.5	87.1	89.0	67.6
2	90.2	90.3	92.4	92.6	93.1	92.7	87.1	91.3	
3	90.5	90.6	92.9	93.7	93.4	93.4	87.1	91.7	
4	90.5	90.7	93.0	94.0	93.4	93.5	87.1	91.8	
5	90.5	90.7	93.1	94.1	93.4	93.5	87.1	91.8	
	n=40 (750 replications)								
1	90.3	90.3	90.3	88.4	92.2	90.2	91.1	90.4	65.8
2	92.3	92.3	93.2	93.0	93.8	93.4	90.6	92.8	
3	92.6	92.6	93.7	94.0	94.0	93.9	90.6	93.1	
4	92.6	92.7	93.8	94.3	94.1	94.0	90.6	93.2	
5	92.6	92.7	93.8	94.4	94.1	94.0	90.6	93.2	

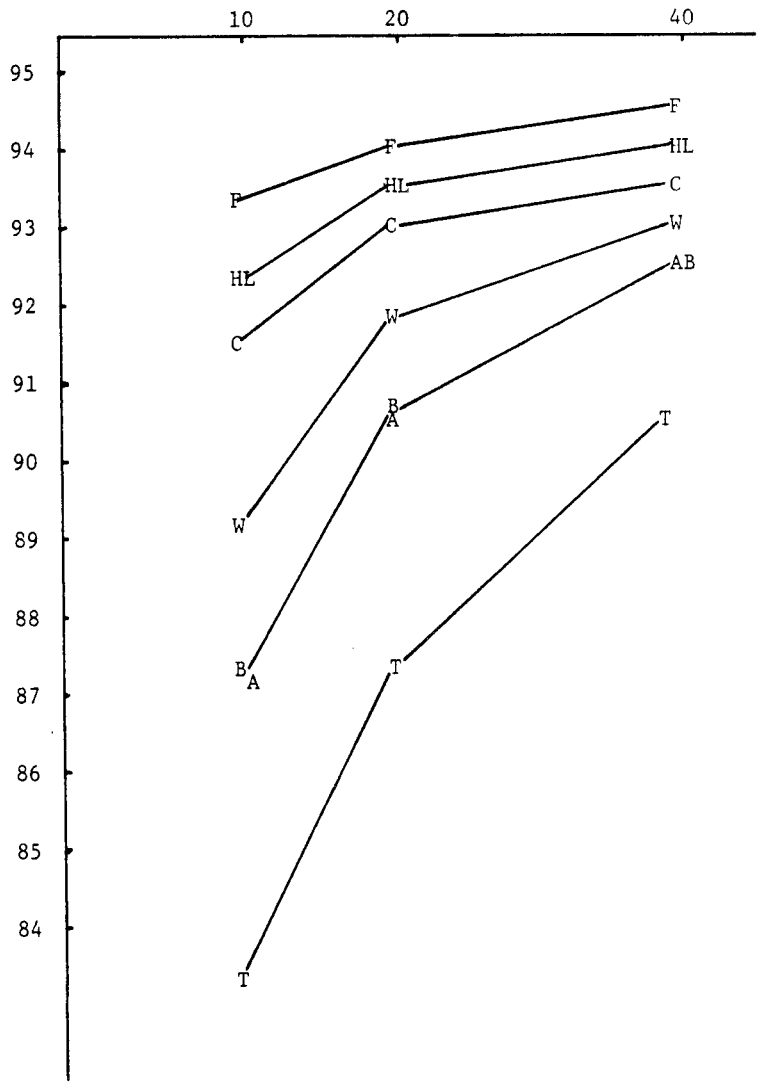

FIG. 1

Gaussian Efficiencies with Estimated Scale (5 Iterations)

TABLE IV

Gaussian Efficiencies with Known Scale

Iteration	Weight Function								
	A	B	C	F	H	L	T	W	M
	n=10 (1000 replications)								
1	92.2	92.2	91.6	89.4	93.4	91.3	92.8	92.1	70.2
2	94.4	94.4	94.4	93.5	95.0	94.3	93.2	94.5	
3	94.7	94.7	94.9	94.5	95.2	94.9	93.2	94.9	
4	94.7	94.8	95.0	94.8	95.3	95.0	93.2	95.0	
5	94.8	94.8	95.1	94.9	95.3	95.0	93.2	95.0	
	n=20 (1500 replications)								
1	92.3	92.3	91.3	88.9	93.1	90.9	94.7	92.0	67.6
2	94.6	94.6	94.3	93.4	94.8	94.1	94.9	94.5	
3	94.9	94.9	94.8	94.4	94.9	94.7	94.9	94.9	
4	95.0	95.0	94.9	94.7	95.0	94.8	94.9	94.9	
5	95.0	95.0	94.9	94.8	95.0	94.9	94.9	95.0	
	n=40 (750 replications)								
1	92.8	92.8	91.6	89.0	93.3	91.2	94.6	92.4	65.8
2	94.8	94.8	94.4	93.6	94.7	94.3	94.7	94.7	
3	95.0	95.0	94.8	94.5	94.8	94.7	94.7	94.9	
4	95.0	95.0	94.9	94.8	94.8	94.8	94.7	95.0	
5	95.0	95.0	94.9	94.8	94.8	94.9	94.7	95.0	

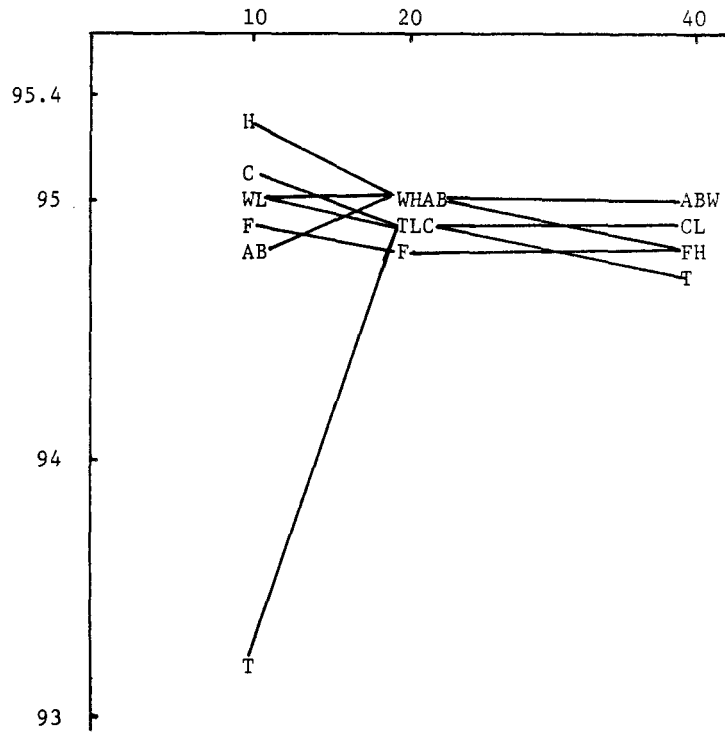

FIG. 2

Gaussian Efficiencies with Known Scale (5 Iterations)

fact that it should be more carefully considered. One implication might be that each w-function needs it own version of the median absolute deviations scaling (8) with a multiplier that depends on the w-function and n.

So far we have only given results at the Gaussian distribution. We also did a small study of the SLASH distribution which has the form

$$Y = \frac{G}{U} \tag{12}$$

where G is Gaussian (0,1) and U is uniform on (0,1) and independent of G. Table V gives the variance of each estimator for 1 to 5 iterations. For the SLASH, the median (M) does better than all of

TABLE V

Slash Variances with Estimated Scale

Iteration	Weight Function								
	A	B	C	F	H	L	T	W	M
	n=10 (1000 replications)								
1	7.70	7.70	8.27	9.98	9.51	9.23	9.25	7.76	7.39
2	8.12	8.12	8.78	11.38	10.07	10.01	9.33	8.17	
3	8.24	8.24	8.92	11.90	10.17	10.23	9.34	8.28	
4	8.27	8.26	9.96	12.09	10.19	10.29	9.34	8.31	
5	8.28	8.27	8.97	12.15	10.20	10.31	9.34	8.31	
	n=20 (1000 replications)								
1	6.16	6.16	6.62	8.41	7.82	7.65	7.28	6.19	6.26
2	6.48	6.48	6.96	9.55	8.28	8.25	7.56	6.49	
3	6.57	6.57	7.05	9.94	8.36	8.41	7.56	6.57	
4	6.60	6.59	7.07	10.06	8.37	8.44	7.56	6.59	
5	6.60	6.60	7.07	10.10	8.38	8.45	7.56	6.60	
	n=40 (500 replications)								
1	5.98	5.99	6.74	8.50	7.80	7.72	7.13	6.11	6.63
2	6.20	6.21	7.04	9.55	8.25	8.29	7.38	6.32	
3	6.27	6.28	7.12	9.91	8.33	8.43	7.40	6.38	
4	6.28	6.29	7.14	10.02	8.34	8.47	7.40	6.39	
5	6.29	6.30	7.14	10.05	8.34	8.48	7.40	6.40	

the fully iterated estimators except for A, B and W when n=40. These tables also illustrate the phenomenon that multiple iterations tend to increase the variance under highly non-Gaussian error (like the SLASH). Figure 3, graphs the data from the 5 iteration columns of Table V for n=10, 20, 40. It shows that under the SLASH, the ordering of non-Gaussian variances is in agreement with the Gaussian efficiencies. Thus except for T, the ordering of the variances under the SLASH is exactly the opposite of the ordering of the Gaussian efficiencies in Figure 1. This suggests that if the differential sensitivity to scale estimation were removed for these estimators, their behavior under the SLASH would be the same. The estimator T is a possible exception and its

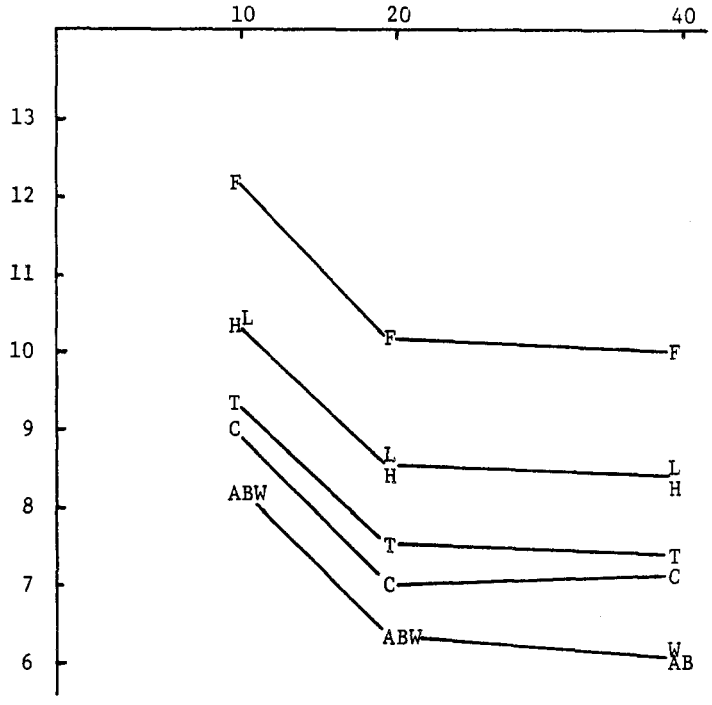

FIG. 3

SLASH Variances with Estimated Scale (5 Iterations)

poorer performance under both the Gaussian and the SLASH distribution suggests that a more detailed examination is warranted.

The Gaussian part of this study used the swindles discussed in Holland (1973). The random number generators were obtained from Marsaglia et al. (1975).

ACKNOWLEDGMENTS

This research was supported by National Science Foundation Grants DCR75-08802 and 76-14311DSS to the National Bureau of Economic Research, Inc. We wish to thank the referee for several helpful suggestions.

BIBLIOGRAPHY

Andrews, D., Bickel, P., Hampel, F., Huber, P., Rogers, W., & Tukey, J. (1972). Robust Estimates of Location: Survey and Advances. Princeton: Princeton University Press.

Andrews, D.F. (1974). A robust method for multiple linear regression. Technometrics 16, 523-31.

Armstrong, R.D. & Frome, E.L. (1976). A comparison of two algorithms for absolute deviation curve fitting. J. Amer. Statist. Assoc. 71, 328-30.

Barrodale, I. & Roberts, F.D.K. (1973). An improved algorithm for discrete L_1 approximation. SIAM J. Numer. Anal. 10, 839-48.

Bartels, R.H., Conn, A.R., Sinclair, J.W. (1975). Minimization techniques for piecewise differentiable functions: L_1 solution to an overdetermined linear system. Tech. Report 230, Dept. of Math. Sciences, Johns Hopkins University, Baltimore, Maryland.

Beaton, A.E. & Tukey, J.W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics 16, 147-85.

Bickel, P.J. (1975). One-step Huber estimates in the linear model. J. Amer. Statist. Assoc. 70, 428-34.

Collins, J.R. (1976). Robust estimation of a location parameter in the presence of asymmetry. Ann. Statis. 4, 68-85.

Dennis, J.E. & Welsch, R.E. (1976). Techniques for nonlinear least squares and robust regression. 1976 Proc. Amer. Statist. Assoc. Statist. Comp. Section. Washington, D.C.: American Statistical Association. 83-87.

Fair, R.C. (1974). On the robust estimation of econometric models. Ann. Econ. Social Measurement 3, 667-78.

Gross, A.M. (1977). Confidence intervals for bisquare regression estimates. To appear in J. Amer. Statist. Assoc.

Hill, R.W. & Holland, P.W. (1977). Two robust alternatives to least squares regression. To appear in J. Amer. Statist. Assoc.

Hinich, M.J. & Talwar, P.P. (1975). A simple method for robust regression. J. Amer. Statist. Assoc. 70, 113-19.

Holland, P.W. (1973). Monte Carlo for robust regression: the swindle unmasked. WP 10, Computer Research Center for Economics and Management Science, National Bureau of Economic Research, Cambridge, Massachusetts.

Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Statist. 35, 73-101.

Huber, P.J. (1973). Robust regression: asymptotics, conjectures, and Monte Carlo. Ann. Statist. 1, 799-821.

Huber, P.J. (1975). Robust methods of estimation of regression coefficients. Presented to 2nd Int. Summer School on Problems of Model Choice and Regres. Anal. at Rheinhardsbrum, G.D.R., November 8-18.

Klema, V. (1976). An iteratively reweighted least squares system. Proc. Bicentennial Conf. Math. Program., to appear.

Marsaglia, G., Ananthanarayanan, K. & Paul, N. (1972). The McGill random number package "super-duper". Unpublished notes and program distributed at the 1972 Amer. Statist. Assoc. meeting in Montreal.

Schumaker, A.D. & Paulson, A.S. (1976). Robust data analysis based on characteristic functions I: Univariate data and linear models. Research Report No. 37-76-P5. School of Management, Rensselaer Polytechnic Institute, Troy, New York.

Welsch, R.E. (1975). Confidence regions for robust regression. 1975 Proc. Amer. Statist. Assoc. Statist. Comp. Section. Washington, D.C.: American Statistical Association. 36-42.