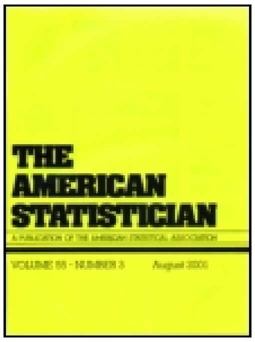

The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares

James O. Street , Raymond J. Carroll & David Ruppert

Pickens, G. T., Busby, R. C., and Mode, C. J. (1984), "MATHIST: An Interactive Program for Modeling Maternity Histories," *The American Statistician*, 38, 224–225.

Sheps, M. C., and Menken, J. A. (1973), *Mathematical Models of Conception and Birth*, Chicago: University of Chicago Press.

Soyka, M. G. (1983), "Computer-Aided Applications of Non-Markovian Macrosimulation Models to the Study of Human Populations," unpublished Ph.D. thesis, Drexel University, Library.

Weiss, G. H., and Zelen, M. (1965), "A Semi-Markov Model for Clinical Trials," *Journal of Applied Probability*, 2, 269–285.

A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares

JAMES O. STREET, RAYMOND J. CARROLL, and DAVID RUPPERT*

The 1985 *SAS User's Guide: Statistics* provides a method for computing robust regression estimates using iterative reweighted least squares and the nonlinear regression procedure NLIN. We show that the estimates are asymptotically correct, although the resulting standard errors are not. We also discuss computation of the estimates.

1. INTRODUCTION

Parameter estimates for generalized linear models (McCullagh and Nelder 1983) and robust regression (Hampel, Ronchetti, Rousseeuw, and Stahel 1986) can be computed by iteratively reweighted least squares techniques using the SAS nonlinear regression procedure NLIN. Examples of these computations are given in the 1985 *SAS User's Guide: Statistics* (SAS Institute 1985, pp. 597–605). Despite the fact that the user's guide makes no mention of standard errors, casual readers may assume standard errors from such a fitting algorithm are correct. In fact this is the case only for generalized linear regression models. That such an algorithm works for generalized linear regression models was shown by McCullagh and Nelder (1983); see also Carroll and Ruppert (1988) for similar results. In this section, we show that the standard errors in the *SAS User's Guide* are inconsistent for robust regression. In Section 2, we discuss computation of the estimates. In Section 3, we present an example to show that the use of these standard errors can give results noticeably at variance with the usual formula.

Consider an ordinary robust regression with the model

$$y_i = x_i^t \beta + \sigma \varepsilon_i,$$

where the errors $\{\varepsilon_i\}$ are independent and identically distributed. For a given estimate $\hat{\sigma}$ of the scale parameter σ, the M estimate of the regression parameter β solves

$$0 = \sum_{i=1}^{N} x_i \, \psi \, [(y_i - x_i^t \hat{\beta})/\hat{\sigma}]. \tag{1}$$

The M estimators defined by Equation (1) are not robust against the effects of leverage—that is, unusual design points. For discussion of M estimates that control for leverage, see Hampel et al. (1986, chap. 4). In Equation (1) the function ψ is usually bounded. Typical choices include $\psi(u) = \max(-k, \min(u, k))$ (Huber's function), Tukey's biweight as in the *SAS User's Guide* and the Hampel function

$$\begin{aligned}
\psi(u) &= -\psi(-u) & \\
&= u, & 0 \le u < a_H \\
&= a_H, & a_H \le u < b_H \\
&= a_H(c_H - u)/(c_H - b_H), & b_H \le u < c_H \\
&= 0, & u \ge c_H.
\end{aligned}$$

The solution $\hat{\beta}$ is usually computed by the following algorithm (see Holland and Welsch 1977):

1. Begin with initial estimates $\hat{\beta}$ of β and $\hat{\sigma}$ of σ.
2. Form the residuals $r_i = (y_i - x_i^t \hat{\beta})/\hat{\sigma}$.
3. Define weights $w_i = \psi(r_i)/r_i$.
4. Update the estimate $\hat{\beta}$ by performing a weighted least squares regression with the weights w_i.
5. Iterate until convergence.

For a nice discussion of tuning constants and a more detailed discussion of robust estimation, see Hogg (1979). Choices of $\hat{\sigma}$ are discussed in Section 2.

Let $\dot{\psi}$ be the derivative of ψ. It is well known (Bickel 1976; Schrader and Hettmansperger 1980) that the final estimate $\hat{\beta}$ is asymptotically normally distributed with mean β and covariance Λ_R, where

$$\Lambda_R = \sigma^2 \, \mathrm{E}\{\psi^2(\varepsilon)\} \left[[\mathrm{E}\, \dot{\psi}(\varepsilon)]^2 \sum_{1}^{N} x_i x_i^t \right]^{-1} . \tag{2}$$

Estimated standard errors are formed by making the following substitutions (see Bickel 1976; Schrader and Hettmansperger 1980):

*James O. Street is Senior Statistician, Medical Data Services, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, Connecticut 06877. Raymond J. Carroll is Professor, Department of Statistics, Texas A&M University, College Station, Texas 77843. David Ruppert is Professor, School of Operations Research, Cornell University, Ithaca, New York 14853. Carroll's research was supported by Air Force Office of Scientific Research Contract F-49620-85-C-0144. Ruppert's research was supported by National Science Foundation Grant MCS-8100748.

$$a = N^{-1} \sum_{i=1}^{N} \dot{\psi}(r_i) \rightarrow \mathrm{E}\{\dot{\psi}(\varepsilon)\}, \qquad \hat{\sigma} \rightarrow \sigma; \tag{3}$$

and

$$\lambda\, b \rightarrow \mathrm{E}\psi^2(\varepsilon), \tag{4}$$

where

$$b = (N-p)^{-1} \sum_{i=1}^{N} \psi^2(r_i), \quad \lambda = 1 + (p/N)(1-a)/a. \tag{5}$$

The correction term λ was discussed by Huber (1981, pp. 172–175).

The procedure NLIN treats the weights $\{w_i\}$ as if they were fixed and known a priori. This is the crux of the problem, because robust regression is one instance where the randomness of the weights is crucial. As shown later, NLIN pretends that $\hat{\beta}$ is asymptotically normally distributed with mean β and covariance Λ_{NLIN}, where

$$\Lambda_{\mathrm{NLIN}} = d\, \Lambda_R$$

$$d = \mathrm{E}\{\epsilon\psi(\epsilon)\}\, [\mathrm{E}\{\dot{\psi}(\epsilon)\}]^2 [\mathrm{E}\{\psi(\epsilon)/\epsilon\}\mathrm{E}\{\psi^2(\epsilon)\}]^{-1}. \tag{6}$$

To see (6), consider the following argument. If we pretend that the weights are fixed, then the estimated covariance matrix is Δ, where

$$\Delta = (N-p)^{-1} \sum_{i=1}^{N} w_i\, [y_i - x_i^t\hat{\beta}]^2 \left[\sum_{i=1}^{N} x_i\, x_i^t\, w_i\right]^{-1}.$$

By standard asymptotic theory,

$$(N-p)^{-1} \sum_{i=1}^{N} w_i\, [y_i - x_i^t\hat{\beta}]^2$$

$$= \hat{\sigma}^2 (N-p)^{-1} \sum_{i=1}^{N} \psi(r_i) r_i \xrightarrow{p} \sigma^2\, \mathrm{E}\{\varepsilon\psi(\varepsilon)\}$$

and

$$N^{-1} \sum_{i=1}^{N} x_i x_i^t\, w_i - N^{-1} \sum_{i=1}^{N} x_i\, x_i^t \mathrm{E}\{\varepsilon\psi(\varepsilon)\} \xrightarrow{p} 0.$$

This verifies (6). Since $d \neq 1$, it follows that the NLIN standard errors are inconsistent. In Section 2 we present an example where the difference between estimates of Λ_{NLIN} and Λ_R is not trivial.

That Λ_{NLIN} and Λ_R can be different was also discussed by Gross (1977). For a particular Tukey biweight ψ function he showed that Λ_R is approximately 40% larger than Λ_{NLIN} at the normal distribution. Because of this, to make sure that using Λ_{NLIN} does not lead to smaller coverage probabilities, he used t values that are approximately 25% larger than they would be using Λ_R. Thus the alternatives seem to be to estimate Λ_R directly or to adjust the NLIN standard errors for the inconsistency.

2. COMPUTATION

Suppose first that one has computed robust estimates $\hat{\beta}$ and $\hat{\sigma}$ of β and σ. The easiest device we know of for computing standard errors is through the use of pseudovalues (Bickel 1976). Defining (a, b, λ) as in Equations (3)–(5), the pseudovalues are

$$\tilde{y}_i = x_i^t\, \hat{\beta} + (\lambda\hat{\sigma}/a)\, \psi[r_i].$$

If one runs a linear regression replacing the responses $\{y_i\}$ by the pseudovalues $\{\tilde{y}_i\}$, the estimated covariance is asymptotically correct, being

$$\hat{\Lambda}_R = b\, \lambda(\hat{\sigma}/a)^2 \left[\sum_{i=1}^{N} x_i\, x_i^t\right]^{-1}. \tag{7}$$

Tests and confidence intervals using the pseudovalues are also asymptotically correct. Auxiliary quantities such as R^2 would not be meaningful when computed using pseudovalues. An alternative approach to hypothesis testing was discussed by Schrader and Hettmansperger (1980).

It thus remains to consider numerical calculation of $\hat{\sigma}$ and $\hat{\beta}$. For a given value of $\hat{\sigma}$, the algorithm discussed in Section 1 can be employed. For a given $\hat{\beta}$, there are two common estimates of σ. The first is based on the median absolute deviation (MAD). The resulting estimate of σ is defined by

$$\hat{\sigma} = \mathrm{MAD}/.6745 = \mathrm{Med}\{|y_i - x_i^t\, \hat{\beta}|\}/.6745. \tag{8}$$

The division by .6745 is made so that for normally distributed data, $\hat{\sigma}$ is an estimate of the standard deviation. Hill and Holland (1977) suggested that for smaller sample sizes, the MAD in (8) be replaced by the modified estimator

(normalized) MAD

$$= \mathrm{Med}\{\text{largest } N-p+1 \text{ of the } |y_i - x_i^t\, \hat{\beta}|\}.$$

The MAD and normalized MAD are easily calculated.

An alternative estimate of σ is Huber's Proposal 2, the usual form of which is the solution to

$$(N-p)^{-1} \sum_{i=1}^{N} \psi^2\, [(y_i - x_i^t\, \hat{\beta})/\hat{\sigma}] = E_Z \psi^2(\varepsilon), \tag{9}$$

where $E_Z\psi^2(\varepsilon)$ is the expected value of $\psi^2(\varepsilon)$ when ε has a standard normal distribution. The right side of (9) is again chosen so that for normally distributed data, $\hat{\sigma}$ estimates the standard deviation. Solving (9) requires iteration. If $\hat{\sigma}_o$ is the present estimae of σ, then the next step in the iteration is defined by

$$\hat{\sigma}^2 = (N-p)^{-1} \sum_{i=1}^{N} w_i^2\, (y_i - x_i^t\, \hat{\beta})^2 / E_Z\psi^2(\varepsilon), \tag{10}$$

where, as before, $w_i = \psi(r_i^{(o)})/r_i^{(o)}$, with $r_i^{(o)} = (y_i - x_i^t\, \hat{\beta})/\hat{\sigma}_o$.

In practice, one would calculate $\hat{\sigma}$ and $\hat{\beta}$ for a fixed number of iterations or until convergence. The calculation is easily programmed in any matrix language such as SAS/IML, GAUSS, or APL.

3. AN EXAMPLE

To illustrate the foregoing, we computed estimates and standard errors for the data used in the SAS/NLIN example—namely, a regression of the U.S. population against

Table 1. Parameter Estimates and Standard Errors for the Example

	β_0	β_1	β_2
Least squares			
Estimates	50.73	97.09	51.40
Standard errors	.96	1.05	1.93
Huber method			
Estimates	50.98	98.37	52.44
Standard errors as in (7)	.45	.49	.90
Standard errors via SAS/NLIN	.56	.64	1.12
Hampel method			
Estimates	51.08	98.85	52.83
Standard errors as in (7)	.36	.39	.73
Standard errors via SAS/NLIN	.30	.35	.60
Tukey biweight with $\sigma = 2$			
Estimates	51.14	98.82	52.68
Standard errors as in (7)	.39	.43	.79
Standard errors via SAS/NLIN	.35	.41	.71

time. The model is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \sigma \varepsilon_i,$$

where $\{y_i\}$ is the U.S. population in millions at year $t_i = 1780 + 10i$ ($i = 1, \ldots, 19$). Rather than use the actual year, we centered and standardized so that

$$x_i = (t_i - 1880)/90.$$

As with any polynomial model with equally spaced time points, there is a bit of a problem with leverage here, since the highest leverage value is .38; however, we will proceed with the usual analyses. We computed parameter estimates and standard errors using least squares, the Huber method with $\psi(x) = \max(-1.25, \min(x, 1.25))$ and the Hampel method with $a = 1.25$, $b = 3.5$, and $c = 8.0$. The robust methods produced estimates of σ by Proposal 2 [see Eqs. (9) and (10)]. The results of the calculations are given in Table 1. For comparison, we also reproduce the results using the Tukey biweight as in the SAS/NLIN manual, where $\hat{\sigma} = 2$. The SAS/NLIN standard errors are about 30% larger than our estimates when using the Huber method and about 20% smaller than our estimates when using the Hampel method.

[*Received March 1987. Revised November 1987.*]

REFERENCES

Bickel, P. J. (1976), "Another Look at Robustness: A Review of Reviews and Some New Developments," *Scandinavian Journal of Statistics*, 3, 145–168.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting In Regression*, New York: Chapman & Hall.

Gross, A. M. (1977), "Confidence Intervals for Bisquare Regression Estimates," *Journal of the American Statistical Association*, 72, 341–351.

Hampel, F. R., Ronchetti, E., Rousseeuw, P. J., and Stahel, W. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley.

Hill, R. W., and Holland, P. W. (1977), "Two Robust Alternatives to Robust Regression," *Journal of the American Statistical Association*, 72, 828–833.

Hogg, R. V. (1979), "Statistical Robustness: One View of Its Use in Applications Today," *The American Statistician*, 33, 108–115.

Holland, P. W., and Welsch, R. E. (1977), "Robustness Regression Using Iteratively Reweighted Least-Squares," *Communications in Statistics—Theory and Methods*, 6, 813–827.

Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley.

McCullagh, P., and Nelder, J. (1983), *Generalized Linear Models*, New York: Chapman & Hall.

SAS Institute, Inc. (1985), *SAS User's Guide: Statistics* (Version 5), Cary, NC: Author.

Schrader, R. M., and Hettmansperger, T. P. (1980), "Robust Analysis of Variance Based Upon a Likelihood Ratio Criterion," *Biometrika*, 67, 93–101.