Behavior Research Methods, Instruments, & Computers
1988, 20 (6), 579-582

ROBREG: An APL procedure for fitting robust regression lines using Huber's M-estimate

JON ANSON
Ben Gurion University of the Negev, Beer Sheba, Israel

The sensitivity of ordinary least squares (OLS) regression to bias by points that lie outside the homoscedastic, multivariate-normal point cloud has been well documented (see, e.g., Huber, 1973, 1977, 1981; Mosteller & Tukey, 1977). This sensitivity has led to several proposals of robust solutions, each of which seeks to restrict the influence of outliers on the estimated coefficients. The present procedure uses Huber's monotonic winsorizing function to control the effect of errant points. The basic OLS approach is maintained, but outlying values are treated as contamination, and are brought closer to the point cloud in order to reduce their pull on the estimated regression line (for a discussion of alternative residual-adjusting functions, see Goodall, 1983; for an APL procedure based on Tukey's biweight function, see McNeil, 1977, p. 177). The procedure iteratively winsorizes outlying residuals and then reestimates the coefficients, in order to minimize the sum of the squares of the adjusted residuals. The winsorizing function is the following:

$$\Psi(r) = \begin{array}{lll} -c & \text{if} & r < -c \\ r & \text{if} & |r| \leq c \\ c & \text{if} & r > c, \end{array}$$

where r is the standardized residual and c is the tuning value beyond which r is winsorized. This tuning constant is usually set between 1 for heavily contaminated samples (>10% of the sample) and 2 for light contamination (<1% of the sample). If contamination is suspected but its extent unknown, set $c=1.5$.

The Algorithm and the Procedure

The procedure uses Huber's M-estimate (see also Dutter, 1975) and iterates through two basic steps, first estimating the scale (root mean square error) of the winsorized residuals, then adjusting the regression coefficients, based on these modified residuals. The algorithm and formulas are reproduced in Appendix A, and the procedure in Appendix B; see also Huber, 1981, pp. 179-183, 172-175, and 195-198; for the proof of convergence, see pp. 184-188.) After convergence has been achieved, standard errors and t values for the regression coefficients are estimated, as well as overall R^2 and F values for the regression. The major procedure is ROBREG, which calls in the subprocedure WINSOR. The procedures were written and tested in IBM APL on an IBM PC with 256K of memory. Local variables are represented by lowercase letters (underlines in other versions of APL), and global variables by uppercase letters.

Input. ROBREG is a dyadic procedure of the form *Y* ROBREG *X*. The left argument is a vector *Y*, the dependent or criterion variable, and the right argument is the matrix of the *X* independent or regressor variables. If a constant term is desired in the equation, a column of 1s should be included in the *X* matrix. A global vector, PAR, sets the parameters of the iteration, and needs to be preset before running the procedure:

PAR[1]: The maximum number of iterations required. A warning is printed if iterations do not converge, and summary statistics are not computed.

PAR[2]: The convergence criterion (usually 0.001 or less).

PAR[3]: c, the tuning constant or standardized distance beyond which the residuals are winsorized.

PAR[4]: Bias correction factor, β, for computing the scale estimate (Huber, 1981, p. 180).

PAR[5]: A relaxation factor, q, to speed up convergence (Huber, 1979, pp. 182-183).

The tuning constant (PAR[3]) will generally be set to 1, 1.5, or 2, and PAR[4] and PAR[5] may then be read in from Table 1.

Output. ROBREG returns a three-column matrix, TH. If $\varrho(X)=(n,p)$ then the first p rows of TH contain the coefficients (θ), their standard errors, and t values. Row $p+1$ contains R^2, the mean square error, and F; and row $p+2$ contains R^2 adjusted for degrees of freedom (Theil, 1971, pp. 178-179), and the numerator $(p-1)$

Table 1
Parameter Values for Turning Points

Tuning Constant	PAR[4] (β)	PAR[5] (q)
1	0.516059	1.46479
1.5	0.778465	1.15422
2	0.920537	1.04767

Table 2
Output Matrix TH

Column 1	Column 2	Column 3
θ_1	$SE(\theta_1)$	$t(\theta_1)$
θ_2	$SE(\theta_2)$	$t(\theta_2)$
.	.	.
.	.	.
θ_p	$SE(\theta_p)$	$t(\theta_p)$
R^2	$\sqrt{MSE}$	F
R^2_{adj}	df_{num}	df_{denom}

Address correspondence to Jon Anson, Department of Social Work, Ben Gurion University of the Negev, 84105 Beer Sheba, Israel.

Table 3
Input Data for Trial Run and Reproduced *Y* Values

ID	Stack Loss (*Y*)	Air Flow	Water Temperature	Acid Concentration	Reproduced *Y* Values	
					OLS	Robust
1	42	80	27	89	38.765	39.095
2	37	80	27	88	39.917	39.229
3	37	75	25	90	32.444	32.873
4	28	62	24	87	22.302	21.822*
5	18	62	22	87	19.712	19.740
6	18	62	23	87	21.007	20.781
7	19	62	24	93	21.389	21.014
8	20	62	24	93	21.389	21.014
9	15	58	23	87	18.144	17.577
10	14	58	18	80	12.733	13.315
11	14	58	18	89	11.364	12.102
12	13	58	17	88	10.221	11.196
13	11	58	18	82	12.429	13.045
14	12	58	19	93	12.050	12.604
15	8	50	18	89	5.639	5.694
16	7	50	18	86	6.095	6.098
17	8	50	19	72	9.520	9.025
18	8	50	19	79	8.455	8.082
19	9	50	20	80	9.598	8.989
20	15	56	20	82	13.588	13.525
21	15	70	20	91	23.238	23.527*

*Outlier case with winsorized residual.

and denominator $(n-p)$ degrees of freedom for F (see Table 2). The global variable, YRESIDS, is a five-column matrix containing the serial number of the observations, the observed Y values, the reproduced Y, the winsorized residual, and an indicator of winsorization (1=yes, 0=no). From this, the raw residuals and the robustified Y values may easily be calculated. The global variable I indicates the number of iterations required.

Application

Dutter (1976) reported a FORTRAN package, LINDWR, to compute robust regression estimates using a variety of residual-adjusting functions. I compare ROBREG with Dutter's results using a set of data reported by Brownlee (1965). The data are in Table 3, together with the Y values reproduced by OLS and robust regression, and the results from the two procedures are compared in Table 4 (tuning constant $c=1.5$, convergence=0.001).

For the OLS fit, the ROBREG results match Dutter's (1976) to six places, his reported level of precision. For the robust fit, results match at least to the third decimal place. The reproduced Y values, reported in Table 3, were identical with Dutter's analyses for both OLS and ROBREG. Because my procedure required only 10 iterations, compared with Dutter's 13, I suspect that part, or even all, of this difference in the robust coefficients stems from the different computers used. Substantively, two of the cases have been identified as outliers, although in this particular analysis the effect on the coefficients and on the predicted Y values was slight. As a test of the procedure's insensitivity to outliers (which would express themselves in the OLS starting values; see Appendix A), the procedure was rerun, setting the starting values of the coefficients as random numbers between −100 and 100. In 10 such runs, the coefficients converged to the reported values in 16 or fewer iterations.

Conclusion

The ROBREG procedure is a short, simple to implement procedure that estimates robust regression coefficients, as well as summary statistics. It may be used to identify outlier dependent data points in a multivariate

Table 4
Comparison of OLS and Robust Regression Results

Regression Coefficients	OLS			ROBUST		
	θ	$SE(\theta)$	$t(\theta)$	θ	$SE(\theta)$	$t(\theta)$
Constant	−39.920	11.9	−3.9	−41.107	10.6	−3.9
Air Flow	0.716	0.135	5.3	0.801	0.121	6.6
Water Temperature	1.295	0.368	3.5	1.041	0.329	3.2
Acid Concentration	−0.152	.152	−1.0	−0.135	0.140	−1.0
R^2		0.914			0.931	
$\sqrt{MSE}$		3.243			2.915	
$F(3,17)$		59.9			76.1	
R^2(adjusted)		0.898			0.918	

point cloud, or in place of standard OLS results in fitting bivariate or multivariate regression lines.

REFERENCES

BROWNLEE, K. A. (1965). *Statistical theory and methodology in science and engineering*. New York: Wiley.

DUTTER, R. (1975). *Numerical solution of robust regression problems: Computational aspects, a comparison* (Research Rep. No. 7). Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, 8006 Zurich.

DUTTER, R. (1976). *Computer linear robust curve fitting program* (Research Rep. No. 10). Fachgruppe fuer Statistik, Eidgenoessische Technische Hochschule, 8006 Zurich.

GOODALL, C. (1983). M-estimators of location: An outline of the theory. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 339-403). New York: Wiley.

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 799-821.

HUBER, P. J. (1977). *Robust statistical procedures* (Regional Conference Series in Applied Mathematics, No. 27). Philadelphia: Society for Industrial and Applied Mathematics.

HUBER, P. J. (1981). *Robust statistics*. New York: Wiley.

MOSTELLER, F., & TUKEY, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.

MCNEIL, D. R. (1977). *Interactive data analysis*. New York: Wiley.

THEIL, H. (1971). *Principles of econometrics*. New York: Wiley.

APPENDIX A
The Algorithm and Computing Formulas

1. Setting Up

   Y is an n-element vector of the dependent variable.

   X is an $n \times p$ matrix of independent variables.

   i. Compute $cii = \sqrt{(X'X)^{-1}}$

   ii. Set iteration counter $i=0$

   iii. Compute coefficients $\theta_{(0)}$ and residuals $z_{(0)}$ by OLS.

   iv. Compute the mean square error (standard error of the estimate):

$$\sigma^2_{(0)} = \frac{1}{(n-p)\beta} \Sigma z^2_{(0)}.$$

2. The Iterations

   i. Compute the winsorized mean square error:

$$\sigma^2_{(i+1)} = \frac{1}{(n-p)\beta} \Sigma \psi \frac{(z_{(i)})^2}{\sigma_{(i)}} \cdot \sigma^2_{(i)}.$$

   ii. Compute the winsorized residuals:

$$\Delta_{(i)} = \sigma_{(i+1)} \cdot \psi \frac{z_{(i)}}{\sigma_{(i+1)}}.$$

   iii. Adjust the computed coefficients:

$$q\tau = q\Delta X(X'X)^{-1}$$

$$\theta_{(i+1)} = \theta_{(i)} + q\tau,$$

   where q is the relaxation factor defined by PAR [5].

   iv. If

$$\min \frac{|q\tau|}{cii \cdot \sigma_{(i+1)}} < \epsilon$$

   and

$$|\sigma_{(i+1)} - \sigma_{(1)}| < \epsilon,$$

   with ϵ the convergence criterion PAR[2], set $\theta = \theta_{(i+1)}$, and proceed to Summary Statistics. Otherwise, return for further iteration.

3. Summary Statistics

   i. Set k, the correction factor (Huber, 1981, pp. 173-174) for winsorizing:

$$k = 1 + \frac{p}{n} \frac{1-m}{m},$$

   where m is the relative frequency of nonwinsorized residuals, and n and p are, as above, the number of cases and of regressor variables, respectively.

   ii. Estimate the winsorized Y values:

$$y' = \hat{y} + k\Delta/m,$$

   where $\hat{y}$ = the predicted values of y, and compute R^2, F, and the standard errors of the coefficients, using these values of y' and $k\Delta/m$ as if they were original y values and OLS residuals.

APPENDIX B
APL Procedure Listings

```
     ▼ROBREG[□]▼
    ▼ TH←Y ROBREG C;cii;del;qt;nsig;q;z;sig;dsig;wz;rsq;y;f;df;m;k
[1]   q←PAR,PAR[4]*-/⍴C
[2]   cii←(1 1⍉⌹(⍉C)+.×C)*0.5
[3]   z←Y-C+.×TH←Y⌹C
[4]   sig←÷q[6]÷z+.×z
[5]   sig←sig,sig*0.5
[6]   →IT+I←0
[7]  ITER:z←Y-C+.×TH
[8]  IT:nsig←sig[1]÷q[6]÷wz+.×wz←q[3]WINSOR z÷sig[2]
[9]   dsig←|sig[2]-¯1↑nsig←nsig,nsig*0.5
```

APPENDIX B (Continued)

```
[10]  sig←nsig
[11]  del←sig[2]×q[3]WINSOR z÷sig[2]
[12]  TH←TH+q[5]×qt←del≤C
[13]  →((∨/q[2]≤|(qt÷sig[2]×cii),dsig÷sig[2])^q[1]>I←I+1)/ITER
[14]  ⍎(q[1]≤I)/('→0×⍴□←''ITERATIONS DID NOT CONVERGE ''')
[15]  YRESIDS←(↑⍴Y),Y,(y←C+.×TH),del,[1.5]del⌿z
[16]  k←1+(÷/⍴C)×¯1+÷m←(del+.=z)÷⍴z
[17]  del←÷m÷del×k
[18]  TH←TH,÷/TH←TH,[1.5]cii×((del←del+.×del)÷(-/⍴C))*0.5
[19]  rsq←÷1+del÷y+.×y←y-(+/y)÷⍴y
[20]  f←rsq×(-/⍴C)÷(df←¯1+1↓⍴C)×1-rsq
[21]  TH←TH,[1]2 3⍴rsq,sig[2],f,(rsq-df×(1-rsq)÷-/⍴C),df,-/⍴C
      ▼
```

```
      ▼WINSOR[□]▼
    ▼ Z←C WINSOR Y
[1]   Z←⌊/C,[1.5]⌐/Y,[1.5]-C
      ▼
```

(Revision accepted for publication March 16, 1988.)