Integrating a Robust Option into a Multiple Regression Computing Environment

Article · January 1992

DOI: 10.1007/978-1-4613-9154-8_3

William DuMouchel
Oracle Corporation

INTEGRATING A ROBUST OPTION INTO A MULTIPLE REGRESSION COMPUTING ENVIRONMENT

WILLIAM DUMOUCHEL* AND FANNY O'BRIEN*

Key words. Robust regression, Tukey's bisquare, interactive data analysis, software design.

1. The Mulreg Program.

Mulreg is an interactive computing environment for data exploration, regression modeling, and the visualization and use of regression results. The program is part of the RS/Explore® data analysis software which runs on minicomputers, mainframes, and workstations from several manufacturers and has been distributed to over 500 industrial research, development and manufacturing sites. RS/Explore provides a menu system which allows it to be used effectively by nonstatisticians, especially industrial scientists and engineers. The system is structured so as to provide a "safety net" in the form of tools and recommended analyses that simplify the methodology, that make it more intuitive, and that help the user avoid common pitfalls and inappropriate applications of regression methods.

A primary design strategy was to integrate graphics into every stage of the modeling process. The Mulreg object contains some newly developed graphical techniques for visualizing and reviewing the effects and possible interactions of predictors in a multiple regression. Extensive menus of residuals graphs and regression diagnostics graphs are provided. Special graphs have also been implemented for diagnosing the need for a transformation, for checking whether the residual variance is constant, and for inspecting the weights in a bisquare regression estimate. And once a model has been accepted, there are menus for creating graphs of estimates and confidence intervals for predictions and contrasts, as well as contour plots of the fitted response surface. DuMouchel (1988, 1989) provides more details.

The software also has menus and internal data structures that provide a framework for providing assistance to the user. Mulreg provides consistent strategies of data analysis that implement modern statistical practices and that also cover many situations which "fall through the cracks" of textbook regression recommendations.

The principle of trying to maintain a consistent "look and feel", even though the program copes with very many aspects of the general linear model, was uppermost during the design of Mulreg. Consider this list of some of the features of Mulreg:

Stepwise selection option
Robust option
Residual diagnostics
Special checks of distributional assumptions
Transformation and back-transformation of variables
Automatic support for categorical predictors

*BBN Software Products, 150 Cambridge Park Drive, Cambridge, MA 02138.

Support for a random effect
Support for unbalanced designs
Support for singular models
Generation of a wide selection of special graphs
Simultaneous confidence intervals for predicted values and contrasts

Every attempt has been made to integrate all these features smoothly, so that selecting one option would not restrict the selection of other options, except where there was a substantial conflict between options, or where the use of certain combinations of options seemed likely to encourage poor statistical practice. (For example, the user is prohibited from generating predictions from a singular model.) The goal has been to simplify the total package as much as possible, so that these useful tools would be as accessible as possible to as wide an audience as possible.

This paper focuses on one aspect of this goal: to integrate the bisquare robust regression technique into a statistical modeling software environment. Section 2 describes how the program helps the user decide between the least squares and the robust approach to estimation. Section 3 provides algorithmic details of the bisquare estimation. Section 4 discusses the computation and use of standard errors for the coefficients and of other inferential techniques. Section 5 describes the implementation of certain analyses of residuals and regression diagnostics. Section 6 contains a summary and concluding discussion.

2. Choosing Between Least Squares and Robust.

2.1 Problem Formulation. Assume the usual multiple regression formulation, $Y = (y_1, \ldots, y_n)^t$,

$$y_i = x_i\beta + \varepsilon_i, \quad i = 1, \ldots, n$$

where x_i is the ith row of an $n \times p$ design matrix X, of full rank, and where the unobserved errors ε_i are assumed to be independent and identically distributed with variance σ^2.

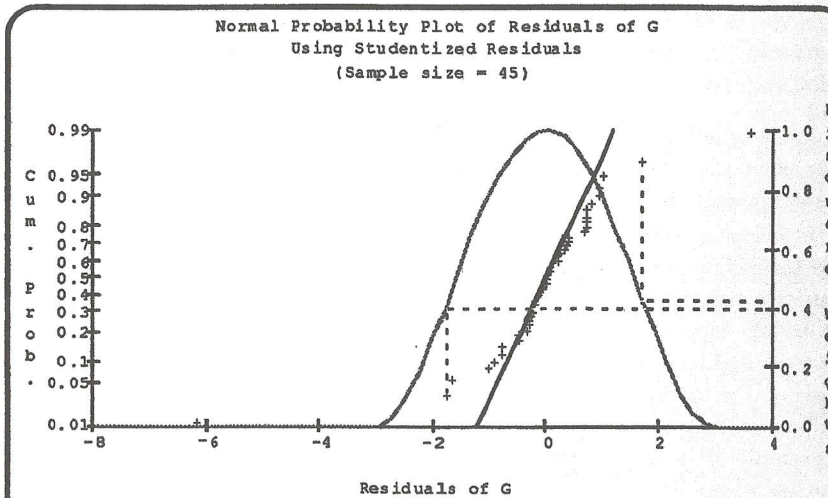
The least squares estimate of β is

$$\beta_{LS} =' (X^tX)^{-1}X^tY,$$

which has variance-covariance matrix

$$(X^tX)^{-1}\sigma^2.$$

The least-squares estimator is to be compared to a robust M-estimator. In this paper, we concern ourselves exclusively with the Tukey bisquare (biweight) estimator, but all the results are readily extended to other M-estimators. Hoaglin, Mosteller and Tukey (1983) provide further background on robust estimation.




```
MULREG.FIT.REFINE.CHECK> robust
(Checking efficiency of bisquare regression ...  Done)
Display graph of bisquare weights and residuals? [Yes]
Making graph.
The efficiency of bisquare regression (compared to least squares) is estimated
to be 164 percent for fitting this data to this model.  Therefore the system
recommends using robust bisquare regression.
Use bisquare weights? [Yes]
```

Mulreg assistance in choosing between the robust and least squares methods. The bisquare weight function is overlaid on a normal probability plot of the residuals from a robust regression. From this one can see that the most extreme positive and negative residuals each got weight 0, and the second most extreme positive and negative residuals each got weight about 0.4. (The dotted lines were added to this figure and do not appear on the computer screen.) The dialog at the bottom of the screen shows the system recommending the use of the bisquare estimation method for this data.

2.2 Estimating Relative Efficiency. The relative efficiency of the robust bisquare to the least squares method is estimated as

$$\text{EFFICIENCY} = 100\% \times \hat{\sigma}^2_{LS}/\hat{\sigma}^2_R$$

where σ^2_{LS} is the mean squared error of the least squares residuals, and where σ^2_R is a robust version of this statistic, described in detail in the next section. The assumption is that the covariance matrix of the robustly estimated coefficients is approximately

$$(X^tX)^{-1}\hat{\sigma}^2_R.$$

The Mulreg user may ask the system for advice on whether to use the least squares or the robust estimate. The system recommends the bisquare method whenever EFFICIENCY is greater than a user-specified threshold, set at 100% by default. The dialog at the bottom of the Figure shows an example of the system giving such advice.

2.3 Visualizing the Bisquare Weights. In addition to the numerical estimate of relative efficiency, the system helps the user understand the effect of the bisquare weights by displaying an enhanced normal probability plot of residuals.

In many regression modeling situations, the user expects changes in the model or in the estimation technique (for example, adding terms to the equation or making a variable transformation) to "improve" the normality of the residuals. This does not usually happen when the user chooses the robust option. Although the presence of nonnormal least-squares residuals (especially fat tails) is a cue to try a robust estimate, the residuals from the robust fit often look less normal than before. This can be confusing to someone not used to using robust methods. The Figure shows an example of the Mulreg graph designed to help the user understand the residuals from the robust fit.

A normal probability plot of the residuals from the robust fit is displayed. The bisquare weight function, which depends only on the magnitude of the residuals, is overlaid using a second vertical axis so that the user can see how many points are being down-weighted, and by how much, in the bisquare procedure.

3. Computational Details.

The robust vector of coefficients is the solution to:

$$\sum_{i=1}^{n} w_i(y_i - x_i\widehat{\beta}_R)x_{ij} = 0, \quad j = 1, \ldots, p.$$

where

$$\begin{aligned} w_i &= (1-u_i^2)^2 && \text{if } |u_i| \le 1, \\ &= 0 && \text{if } |u_i| > 1. \end{aligned}$$

and

$$u_i = (y_i - x_i\widehat{\beta}_R)/cS\sqrt{1-h_i}$$
$$S = \text{ MAD }/0.6745.$$

The quantities h_i are the leverages, the diagonal elements of $X(X^tX)^{-1}X^t$. The MAD is the median of the absolute deviations of the residuals, which is computed after removing the smallest $(p-1)$ absolute residuals, as discussed and recommended in Hill and Holland (1977) and in Hogg's paper in Launier and Wilkinson (1979, p. 6). The tuning constant $c = 4.685$ was chosen so that the robust method would have 95% a.r.e. if normality holds.

The solution is attained by the method of iteratively reweighted least squares. Starting with the least squares solution (every $w_i = 1$), the values of $\widehat{\beta}_R, u_i, w_i$, and then the next value of $\widehat{\beta}_R$, etc. are computed in turn until convergence.

The procedure differs from the usual implementation of the bisquare method in the use of the factor $\sqrt{1-h_i}$ in the denominator of the definition of u_i. This modification, originally proposed by Huber (1981) and studied extensively by O'Brien (1984), has the effect of down-weighting high-leverage points. In simulations done in O'Brien (1984), calculating Tukey's bisquare using least squares as the initial estimate and residuals adjusted by the factor $\sqrt{1-h_i}$ at each iteration gave results at least as good as using the more robust least absolute deviations estimate as the starting estimate in the iterative procedure.

After each iteration, the value of the auxiliary scale estimate, S, is recalculated. A maximum of ten iterations is performed, and a test for convergence is made after each iteration.

A weighted bisquare fit can also be performed. That is, in addition to the weights used in the iteratively reweighted least squares computations, other weights can be used. for example, if information about the relative error variance of different observations is available, you could specify weights inversely proportional to the error variances. Each row of X and Y is premultiplied by the square root of the corresponding weight before the robust algorithm is started.

4. Estimation of Standard Errors.

4.1 Computation of the Robust MSE. Although the robust fit is computed as a weighted regression, the usual formulas for the standard errors of coefficients for a regression in which the weights are fixed in advance do not apply. Street, Carroll, and Ruppert (1988) point out that some statistical packages incorrectly use such formulas.

Let $\hat{\hat{\sigma}}_R^2$ be the sample estimate of the asymptotic variance factor as defined by Huber (1981, p. 172) or Street et al. (1988, eq. 2). In the case of the bisquare estimate the formula is

$$\hat{\hat{\sigma}}_R^2 = K^2 \frac{\sum(y_i - x_i\widehat{\beta}_R)^2(1-u_i^2)^4/(n-p)}{A^2}$$

where

$$K = 1 + \frac{p}{n}\frac{\frac{1}{n}\sum[(1-u_i^2)(1-5u_i^2) - A]^2}{A^2}$$

and

$$A = \frac{1}{n}\sum(1-u_i^2)(1-5u_i^2).$$

In the above summations, only terms for which $|u_i| < 1$ are included in the sum.

The Mulreg program modifies this asymptotic variance factor to be more conservative if there are not many degrees of freedom for error. It assumes that

$$V(\widehat{\beta}_R) = (X^tX)^{-1}\hat{\sigma}_R^2$$

where

$$\begin{aligned} \hat{\sigma}_R^2 &= \hat{\hat{\sigma}}_R^2 && \text{if } \hat{\hat{\sigma}}_R^2 > \hat{\sigma}_{LS}^2 \\ &= \left(n\hat{\hat{\sigma}}_R^2 + p^2\hat{\sigma}_{LS}^2\right)/(n+p^2) && \text{otherwise .} \end{aligned}$$

The rationale behind this adjustment is as follows: Since the asymptotic theory (see Huber (1981)) upon which the use of $\hat{\hat{\sigma}}^2_R$ is based assumes that $n \gg p^2$, a conservative approach to estimation of standard errors is to not allow the standard error of the robust estimate to fall much below the standard error of the least squares estimate unless n is large compared to p^2. The definition of $\hat{\sigma}^2_R$ given above does this.

4.2 Uses of the Robust MSE. The most straightforward use of $\hat{\sigma}^2_R$ is to compute standard errors of coefficients for use in confidence intervals and t-ratios for each coefficient, as in Gross (1977). In addition, Mulreg has special menus for computation of confidence intervals for predictions and contrasts, which also assume that the covariance matrix of the robust coefficients is $(X^tX)^{-1}\hat{\sigma}^2_R$.

The anova table for a robust regression is built up as follows: The robust version of the residual sum of squares is defined as

$$SS_{\text{resid}} = (n-p)\hat{\sigma}^2_R.$$

Sums of squares due to regression effects are then defined in terms of a contrast matrix C which identifies an hypothesis $C\beta = 0$ to be tested. The robust sum of squares "due to C" is defined as

$$SS_{\text{due-to-}C} = (C\widehat{\beta}_R)^t[C(X^tX)^{-1}C^t]^{-1}C\widehat{\beta}_R$$

which allows construction of an anova table with the use of the F-statistic

$$F = \frac{SS_{\text{due-to-}C}/q}{SS_{\text{resid}}/(n-p)}$$

if C has rank q. Under the null hypothesis qF has an asymptotic (for n large) Chisquared distribution with q degrees of freedom. The system uses the more conservative $F(q, n-p)$ reference distribution for F.

In Mulreg, the stepwise regression option also has a bisquare version, which uses the bisquare fitting method at each step. Using the anova theory discussed above, the system computes values of p-to-remove for each term in the model, but cannot compute p-to-enter for terms not currently in the equation, so backward elimina-
compute p-to-enter for terms not currently in the equation, so backward elimina-
tion of terms is recommended. During the robust stepwise procedure, the iterative bisquare uses the estimate and weights from the previous fit as a starting point for determining a new bisquare estimate. Thus, the first step in the robust stepwise procedure starts with the least squares estimate to get the bisquare estimate, but for subsequent steps, it starts from the previous bisquare estimate.

If duplicate rows of X are available in the sample, an anova test for lack of fit uses the pseudo-residuals, defined as

$$\begin{aligned} \tilde{e}_{jk} &= e_{jk}(1-u^2_{jk})^2 \quad && \text{if } u^2_{jk} < 1, \\ &= 0 && \text{otherwise}, \end{aligned}$$

where e_{jk} is the jth residual in the kth group of duplicates, having n_k members.

Then, if $\bar{\tilde{e}}_k$ is the mean of the $\tilde{e}_{jk}$, the sum of squares of the pseudo-residuals is partitioned in the usual way:

$$\sum_k \sum_j \tilde{e}^2_{jk} = \sum_k n_k \bar{\tilde{e}}^2_k + \sum_k \sum_j (\tilde{e}_{jk} - \bar{\tilde{e}}_k)^2$$

to form an approximate F-statistic for lack of fit.

5. Residuals and Regression Diagnostics.

When exploring residuals from the bisquare fit, the Studentized residuals are defined to be

$$e^*_i = \frac{y_i - x_i\widehat{\beta}_R}{\sqrt{1-h_i}\hat{\sigma}_R}.$$

It is not necessary to compute a version of $\hat{\sigma}^2_R$ which omits use of observation i, since the nature of the robust estimation ensures that $\hat{\sigma}^2_R$ is not inflated by extreme outliers.

When the user asks for a normal probability plot of the residuals, the bisquare weight function, which depends only on the studentized residuals, is also plotted as discussed in Section 2 and shown in the Figure.

The leverage values h_i are the same for the least squares and the robust fit, since leverages are intended to be description of the design space and not of the responses.

The measure of influence, Cook's distance, can be represented as a function of the leverage and the studentized residual; the robust version differs from the least squares version only in the use of e^*_i as defined above.

The Mulreg system also uses the residuals from the robust regression to check the data for heterogeneity of variance. As described in BBNSPC (1988, ch. 5) interquartile ranges of the robust residuals are computed separately for subsets of the cases corresponding to various partitions of the design space. If a large variation in the IQRs is found, the system can recommend the use of a weighted regression to produce a more efficient estimate.

6. Summary.

This paper has shown that robust regression can be combined with versions of most of the usual least squares modeling tools.

Studentized residuals and other regression diagnostic statistics have robust analogs. The normal probability plot of residuals has a different interpretation in the robust context, and it is recommended that this plot be combined with information about the bisquare weights, as shown in the Figure.

We also recommend modifying the usual bisquare algorithm to provide somewhat greater resistance to high leverage points, as discussed in Section 3. This adjustment also seems to lead to quicker and more reliable numerical convergence, as shown by O'Brien (1984).

The computation and use of the asymptotic variance formulas are very important to making the most of robust regression. Since users may not have sample sizes large enough to justify use of this approximation, we recommend an ad hoc tempering of the variance estimate which depends on the ratio p^2/n.

Once the approximation of the covariance matrix of the coefficients by $(X^tX)^{-1}\hat{\sigma}_R^2$ is accepted, all of the usual inference procedures are available. However, backward stepwise selection is computationally more attractive than forward selection, since asymptotic formulas for p-to-remove, but not p-to-enter, are readily available at each step. Finally a comparison of $\hat{\sigma}_R^2$ to $\hat{\sigma}_{LS}^2$ leads to an estimate of the relative efficiency of robust to least-square estimation in any given situation, and provides a natural basis for recommending one procedure over the other.

REFERENCES

BBN Software Products Corporation, RS/Explore Statistical Appendices, Cambridge, Massachusetts (1988).

DuMouchel, W., *Graphical representations of main effects in a polynomial regression on several predictors*, Proceedings of the 20th Symposium on the Interface, George Mason University (1988).

DuMouchel, W., *The structures, design principles, and strategies of Mulreg*, Proceedings of the Second International Workshop on Artificial Intelligence and Statistics (1989).

Gross, Alan M., *Confidence Intervals for Bisquare Regression Estimates*, Journal of the American Statistical Association, 72 (1977), pp. 341–354.

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

Hoaglin, D.C., Mosteller F., and Tukey, J., *Understanding Robust and Exploratory Data Analysis*, New York: John Wiley and Sons (1983).

Huber, Peter J., *Robust Statsitics*, New York: John Wiley and Sons.

Launier, R.L., and Wilkinson, G.N. (eds.), *Robustness in Statistics*, New York: Academic Press (1979).

O'Brien, F.L., *Polyefficient and polyeffective linear regression estimators and the absolute polyefficiency of the biweight regression estimator*, Ph. D. thesis, Princeton University (1984).

Street, J.O., Carroll, R.J., and Ruppert, D., *A Note on Computing Robust Regression Estimates via Iteratively Reweighted Least Squares*, The American Statistician, 42 (1988), pp. 152–154.