

# A Comparison of OLS and Robust Regression using S-PLUS

---

**Randall E. Schumacker**  
University of North Texas

**Michael P. Monahan**  
University of North Texas

**Robert E. Mount**  
Dallas Independent School District

---

Researchers need to consider robust estimation methods when analyzing data in multiple regression. The ordinary least squares estimation of regression weights in multiple regression is affected by outliers, non-normality, multicollinearity, and missing data. It is therefore important to check the accuracy and stability of estimates using robust estimation methods. The ordinary least squares, least-trimmed, and MM parameter estimation methods were compared in the present study, which reported widely different results between ordinary least squares and the robust estimation methods. An example was given using The High School and Beyond data set with the S-PLUS statistical package. Textbooks articulate many different robust regression methods, however, only a few software packages permit robust regression estimation.

---

**P**arametric statistics use a sample statistic as an estimate of the population parameter; hence a researcher is making an inference about the value of the population parameter given knowledge of a sample statistic. An estimate (estimator) is therefore a value of a sample statistic that provides information about the population parameter. When conducting parametric statistical tests the researcher should be concerned with the properties of estimators (Glass & Hopkins, 1984). The properties of an estimate are important because one wants the sample statistic to be an accurate and stable estimate of the population parameter. The properties of estimators are unbiased, consistent, efficient, and sufficient. An estimate is *unbiased* when the mean of the sampling distribution of the statistic equals the population parameter being estimated. An estimate is *consistent* if the sample statistic gets closer to the population parameter as sample size increases. An estimate is *efficient* if it doesn't vary much from sample to sample (variance error or sampling error); hence it refers to the precision of estimation. The variance error of the statistic is the variance of the sampling distribution of the statistic. An estimate is *sufficient* when no other sample statistic is a better estimate of the population parameter.

Another concern is how aspects of the data can affect the sample statistic as an estimate of the population parameter. For example, the Pearson correlation coefficient is attenuated when computed with (1) missing data, (2) a restricted range of scores, (3) non-interval level data, or (4) non-normal data. This concern is magnified in the field of parametric statistics when a researcher discovers that the Pearson correlation coefficient is used in numerous multi-variable methods (i.e. multiple regression, path analysis, factor analysis, canonical correlation, and discriminant analysis) to name a few.

## Robust Regression

In multiple regression, ordinary least squares (OLS) estimation is used if assumptions are met to obtain regression weights when analyzing data. The assumptions of OLS are that residual errors should be normally distributed, have equal variance at all levels of the independent variables (homoscedasticity), and be uncorrelated with both the independent variables and with each other. However, if the data contains missing data, outliers, non-normality, or multicollinearity among variables then sample estimates and results can be misleading (Ho & Naugher, 2000). A researcher has the option of ignoring problems in the data set; deleting subjects; or accommodating. Accommodation involves the use of robust estimation methods in computing sample estimates in multiple regression.

A raw score regression equation with an intercept, two predictor variables, and the residual error term can be written as:  $Y = a + b_1X_1 + b_2X_2 + e$ . Each dependent variable value,  $Y$ , is comprised of an intercept value,  $a$ , two  $b$  regression coefficients *times* corresponding  $X$  independent variable values, and a unique error value,  $e$ . The error term or residual value is the difference between the  $Y$  dependent variable value and a corresponding predicted  $Y$  value ( $\hat{Y}$ ).

Computing an intercept term and estimating a set of  $b$  coefficients that minimize the difference between the  $Y$  and predicted  $Y$  values solves the regression equation for a sample of data. Although there are many  $b$  coefficient estimation methods, the method used most often is that of OLS (OLS). The OLS estimation method optimizes the fit of the model by minimizing the sum of the squared deviations between the actual

and predicted  $Y$  values,  $\Sigma(Y - \hat{Y})^2$ , or denoted as  $e_i^2$ . The OLS estimation method in residual form can be represented as:

$$\text{Min} \sum_{i=1}^n e_i^2$$

In contrast, Rousseeuw (1984) developed the least trimmed squares (LTS) robust estimation method, given by:

$$\text{Min} \sum_{i=1}^h (e_i^2),$$

where  $e_{(1)}^2, e_{(2)}^2, \dots, e_{(n)}^2$ , are the ordered squared residuals, from smallest to largest, and the value of  $h$  must be determined based on trimming the data values. Depending on the value of  $h$  and the outlier data configuration, LTS can be very efficient. In fact, if the exact same numbers of outlying data points are trimmed, this method is computationally equivalent to OLS. However, if there are more outlying data points than are trimmed, LTS method is not as efficient. Conversely, if there is more trimming than there are outlying data points, then some good data will be excluded from the computation. Huber (1973, 1981) developed a group of estimators called M-estimators, which are based on the idea of replacing the squared residuals,  $e_i^2$ , with another function of the residuals, given by:

$$\text{Min} \sum_{i=1}^h \rho(e_i^2),$$

where  $\rho$  is a symmetric function with a unique minimum at zero. M-estimates are calculated using iteratively reweighted least squares (IRLS). In IRLS, the initial fit is calculated, and then a new set of weights is calculated based on the results of the initial fit. The iterations are continued until a specified number of iterations are finished or a convergence criterion is met. MM-estimation is a special type of M-estimation, developed by Yohai (1987). In MM-estimation, a robust M-estimate is given by:

$$\sum_{i=1}^n \rho \frac{y_i - x_i^T \beta}{\hat{s}}; c,$$

where  $s$  is a robust scale estimate for the residuals and  $\rho(\bullet; c)$  is a convex weight function of the residuals with a tuning constant  $c$  that is an optimal symmetric *bounded* loss function. An alternate choice for the estimation function is:

$$\sum_{i=1}^n x_i \psi \frac{y_i - x_i^T \beta}{\hat{s}}; c,$$

where  $\psi$  is a nonmonotonic function. MM estimation involves a three-stage procedure. In the first stage, a robust, high breakdown estimator is computed. S-estimation is used for the first stage. In the second state, a robust M-estimate is computed using the initial step S-estimate residuals. In the third stage, the final M-estimation regression parameters are computed. The MM estimation method in S-PLUS 2000 (1999) has a variety of statistics for diagnostics and inference including  $p$ -values,  $R^2$  values, tests for bias, robust  $F$ -test, and residual scale estimates.

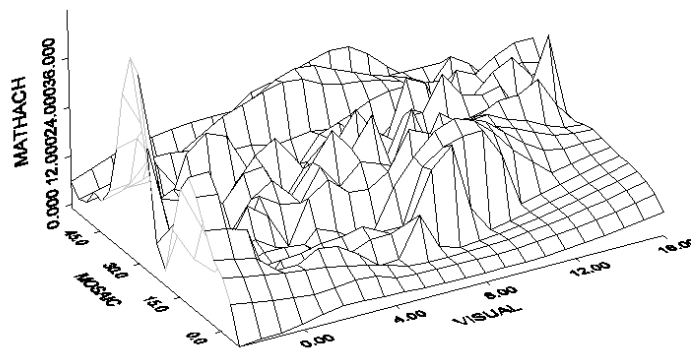
There are various robust estimation methods presented in textbooks (Berk, 1990; Birkes & Dodge, 1993; Fox, 2000; Staudte & Sheather, 1990; and Wilcox, 1997, to name only a few), but many of the robust methods are not available in statistical software packages, so only LTS and MM robust estimation were compared to OLS regression estimation in the present study.

### Method

The High School and Beyond Data Set available in Hinkle, Wiersma, and Jurs (1998) was used to calculate traditional and robust multiple regression parameter estimates in the S-PLUS 2000 (1999) statistical package. The original data set contained information from a survey of 28,240 high school seniors, however, a random sample of 500 seniors was selected for the analysis in this study. The dependent variable was score on a 25-item math achievement test (mathach) and the two-predictor variables were scores on a 16-item visualization test (visual) and a 56-item mosaic pattern test (mosaic). Figure 1 depicts a three dimensional plot of the data values. The OLS (traditional), LTS, and MM (robust) regression results were compared using the S-PLUS statistical software.

**Results**

The results were different between the OLS, LTS, and MM regression methods. Table 1 compared the regression weights (parameter estimates) and  $R^2$  values achieved using each estimation method. The  $Y$ -intercept term ( $a$ ) was much larger using OLS than either of the other two robust estimation methods. The intercept term is computed as:



**Figure 1.** 3D Plot (Math Achievement predicted by Visual and Mosaic)

$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2$ , so estimation of the regression weights was critical to the computation of the  $Y$ -intercept term. The regression weights for the two-predictor variables are computed as:

$$b_1 = r_{yx_1.x_2} \frac{s_y}{s_{x_1}} \text{ and } b_2 = r_{yx_2.x_1} \frac{s_y}{s_{x_2}} \quad (\text{Pedhazur, 1997}).$$

Some aspects of the data are obviously affecting the partial correlation coefficients used in the estimation of the regression weights because they are different between the ordinary least squares and robust estimation methods, especially for the visual data. The  $R^2$  model fit statistic that indicates how much variance in  $Y$  is explained by knowledge of the two independent variables is also very different. If our intention were to maximize the prediction of the dependent variable (i.e., maximize  $R^2$  and minimize  $\Sigma e_i^2$ ), then the MM-robust regression method would be our choice for the analysis for this data. The LTS robust regression method reduced the data set by fifty subjects due to the trimming of outlier data points. Significant outliers that affect the correlation coefficient could have been revealed if the researcher had edited the data and conducted box-whisker plots or tests of non-normality.

**Table 1.** A Comparison of OLS, LTS, and MM Regression Methods

Method	$N$	Intercept	Slopes		$R^2$
			Visual	Mosaic	
OLS	500	5.87	0.66	.13	.22
LTS	450	2.99	0.83	.19	.34
MM	500	2.48	1.10	.15	.42

In order for robust estimation procedures to be used by researchers, they must be accessible and easily performed. A preliminary review of several popular statistical packages revealed that robust methods were not currently available in many popular statistical packages (Table 2). Consequently, researchers are unable to easily check whether violations of assumptions are affecting sample estimates with the exception of the SAS and S-PLUS statistical packages.

**Table 2.** Robust Regression Methods Included in Selected Statistical Software Packages.

Statistical Software Package (version)	Robust Regression Methods						
	GM	MultipleStage GM	M	MM	LAV	LMS	LTS
Mathematica 4.1	no	no	no	no	no	no	no
Minitab 13.21	no	no	no	no	no	no	no
SAS 8.1	no	no	no	no	yes	yes	yes
S-PLUS 2000	no	no	yes	yes	yes	yes	yes
SPSS 10.0	no	no	no	no	no	no	no
Statistica 6.0	no	no	no	no	no	no	no

Source: Anderson (2000).

### Summary

Multiple regression is a popular statistical technique used by researchers to predict or explain relationships between a dependent variable and multiple independent variables. Researchers are concerned with either interpreting the regression weights to determine the relative importance of the predictor variables (explanation) or predicting the dependent variable (interpreting the  $R^2$  value). Real data sets, such as the one illustrated here, often contain missing data, outliers, non-normality, and multicollinearity that violate the assumptions of OLS estimation in multiple regression. An alternative is to use robust regression methods. Consequently, researchers in the social sciences should consider a comparison of OLS parameter estimates with robust regression parameter estimates. Ryan (1997) argued that a robust regression estimation technique should have these characteristics: (1) perform almost as well as OLS when the latter is the appropriate choice (i.e., when the errors are normally distributed with the data being free of mistakes and influential data points), (2) perform much better than OLS when the conditions in (1) are not satisfied, and (3) not be overly difficult to compute or understand.

---

### References

- Anderson, C. (2000). *A Comparison of Five Robust Regression Methods with Ordinary Least Squares: Relative Efficiency, Bias, and Test of the Null Hypothesis*. Unpublished dissertation, University of North Texas, Denton, Texas.
- Berk, R. A. (1990). A Primer on Robust Regression. In J. Fox & J. S. Long (Editors), *Modern Methods of Data Analysis* (pp. 292-323). Newbury Park, CA: Sage Publications, Inc.
- Birkes, D. & Dodge, Y. (1993). *Alternative Methods of Regression*. New York, NY: John Wiley & Sons, Inc.
- Glass, G.V. & Hopkins, K.D. (1984). *Statistical Methods in Education and Psychology* (2<sup>nd</sup> Ed.). Englewood Cliffs, NJ: Prentice Hall.
- Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1998). *Applied Statistics for the Behavioral Sciences* (4<sup>th</sup> Ed.). New York: Houghton Mifflin Company.
- Ho, K. & Naugher, J. (2000). Outliers Lie: An Illustrative Example of Identifying Outliers and Applying Robust Models. *Multiple Linear Regression Viewpoints*, 26(2), 2-6.
- Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo. *The Annals of Statistics*, 1, 799-821.
- Huber, P. J. (1981). *Robust Statistics*. New York, NY: John Wiley & Sons, Inc.
- Pedhazur, E.J. (1997). *Multiple Regression in Behavioral Research* (3<sup>rd</sup> Ed.). Orlando, FL: Harcourt Brace & Company.
- Rousseeuw, P. J., (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880.
- Ryan, T. P. (1997). *Modern Regression Methods*. New York, NY: John Wiley & Sons, Inc.
- S-PLUS 2000 (1999). *Guide to Statistics, Volume I*. Seattle, Washington, Mathsoft, Inc.
- Staudte, R.G. & Sheather, S.J. (1990). *Robust Estimation and Testing*. New York, NY: John Wiley & Sons, Inc.
- Wilcox, R. R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*. San Diego, CA: Academic Press.
- Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(2), 642-656.

---

Send correspondence to: Randall E. Schumacker, Department of Educational Technology and Research, College of Education, University of North Texas, P.O. Box 311 337, Denton, TX 762039.  
 Email: [rschumacker@unt.edu](mailto:rschumacker@unt.edu) website: <http://www.coe.unt.edu/schumacker>

---