

## An Experiment with Experimental Proc RobustReg

Shuang Lu, Merck & Co., Inc., Rahway, NJ

### ABSTRACT

Proc RobustReg is an experimental procedure in SAS/STAT® version 9. It implements the most commonly used robust regression techniques, including M (Maximum likelihood-like) estimation, LTS estimation, S estimation and MM estimation. A macro developed by the Merck Research Laboratories, Merck & Co., Inc. also carries out M estimation of General Linear Models. This paper provides a comparison of the output from an experimental data test and suggests possible improvements in Proc RobustReg, based on the author's experience with the in-house macro.

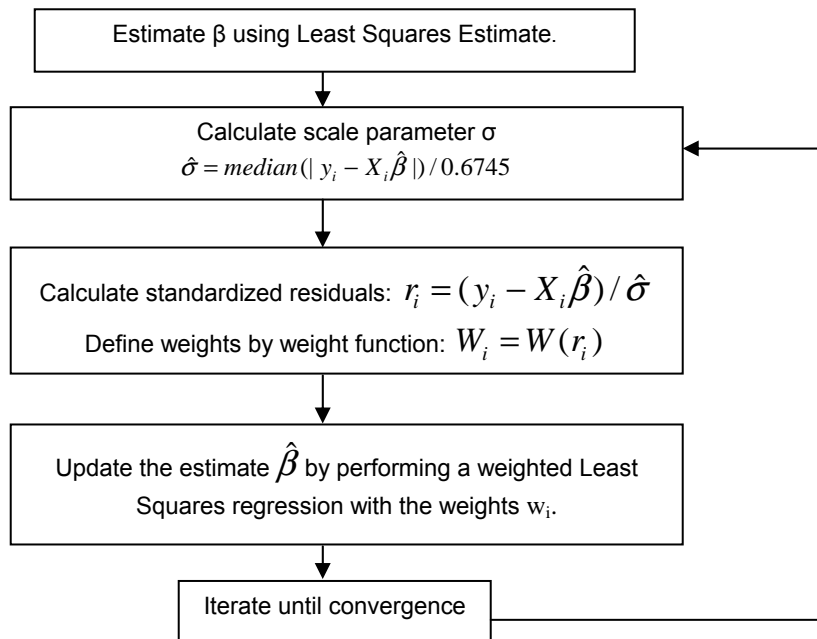
### INTRODUCTION

In 1995, in order to meet the needs of fitting robust regression models with clinical data possibly contaminated with outliers, Merck Research Laboratories, Merck & Co., Inc. developed a SAS macro to perform M estimation of General Linear Models (the macro). The macro performs M estimates and hypothesis testing by an iteratively re-weighted least squares algorithm using Proc GLM (Street, Carroll and Ruppert, 1988).

Proc RobustReg, which became available recently as an experimental procedure in SAS/STAT version 9, implements the most commonly used robust regression techniques, including M estimation, LTS estimation, S estimation and MM estimation. It is useful to see how differently the M estimation method of the macro and Proc RobustReg perform and what future improvements might be made to SAS.

### DESCRIPTION OF ALGORITHM

Proc RobustReg is based on the Iteratively Re-weighted Least Squares (IRLS) algorithm.



SAS provides 3 options for calculating the scale parameter  $\sigma$ , including the one shown above, the Huber method and the Tukey method.

### CHOOSING THE TUNING CONSTANT

The tuning constant  $c$  in robust regression determines the robustness of the estimator to outliers and the efficiency of the estimator in the absence of outliers. Robustness refers to the performance over a range of error distributions. Efficiency refers to how well the method performs when the errors do have a normal distribution.

For the Huber and bisquare weight functions, the smaller the tuning constant, the more robust the model is, and the less efficiency the model has. The bigger the tuning constant, the less robust the model is, and the more efficiency the model has. The tuning constant is generally picked to give reasonably high efficiency in the normal case. In

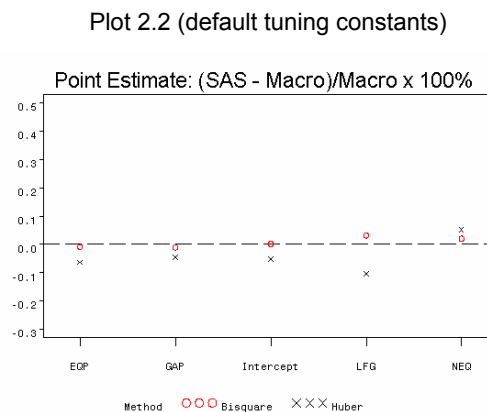
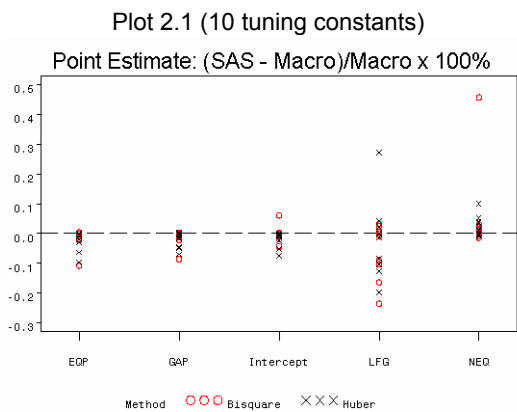
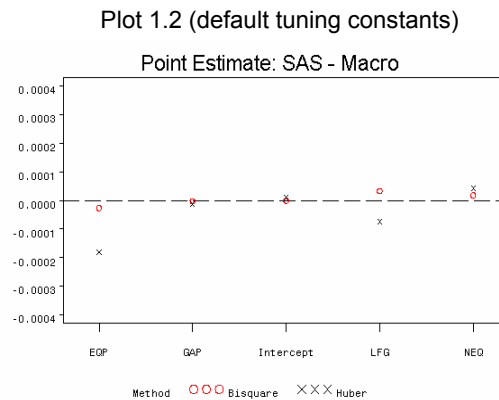
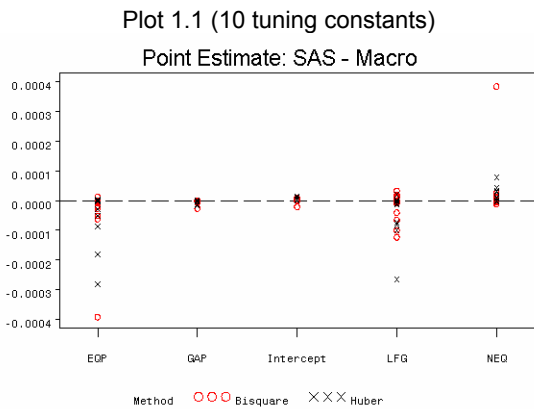
particular,  $c = 1.345$  for the Huber or  $c = 4.685$  for the bisquare weight function produce 95% efficiency when the errors are normal, and still offer protection against outliers.

### RESULT OF COMPARING OUTPUT OF AN EXPERIMENTAL DATA SET

Zaman, Rousseeuw, and Orhan (2001) used the "Growth Study" of De Long and Summers to show how robust techniques improve the OLS results. We also used this data set to show the differences of the output statistics ( parameter estimates, standard errors and p-values of parameter estimates ) produced by the two procedures, by fitting the following robust regression model:  $GDP = \beta_0 + \beta_1 LFG + \beta_2 GAP + \beta_3 EQP + \beta_4 NEQ + \varepsilon$ , where response variable is GDP ( GDP growth per worker ), independent variables are LFG ( labor force growth ), GAP ( relative GDP gap ), EQP ( equipment investment ) and NEQ ( non-equipment investment ).

We ran the two procedures with the same weight function, tuning constant, convergence criteria and statistical model. For the Huber weight function, we used a set of 10 tuning constants ( 0.1, 0.5, 1, 1.2, 1.345, 1.5, 2, 3, 4 and 1000); we used another set of 10 ( 0.1, 0.5, 1, 2.4, 4.685, 5, 6, 7, 8 and 1000) for the bisquare weight function.

In the following plots, every circle represents the differences ( either absolute difference: SAS – Macro or relative difference: ( SAS-Macro ) / Macro x 100% ) of the output statistics, corresponding to a specified tuning constant, when both procedures use the bisquare weight function; every 'x' represents the cases of the Huber weight function. Using these specific experimental data and those 10 tuning constants, the two procedures gave very little difference in parameter estimates, and some slight differences in standard errors and p-values. This might be due to the way that the macro calculates the scale parameter :  $\hat{\sigma} = MAD(y_i - X_i\hat{\beta}) / 0.6745$ , which is different from that of SAS.





The tables below show the details of the output statistics when the default tuning constant is used. The relative differences  $(SAS - Macro) / Macro \times 100\%$  of p-values is just a supplemental tool for comparison. A big relative difference (-72%, -81%) doesn't necessarily mean that the statistical significance will be very different.

Weight Function	Tuning Constant	Variable	SAS PE	Macro PE	SAS - Macro	(SAS-Macro) / Macro*100%
huber	1.345	Intercept	-0.0217	-0.0217	0.0000	-0.0526
huber	1.345	LFG	0.0735	0.0735	-0.0001	-0.1018
huber	1.345	GAP	0.0236	0.0236	0.0000	-0.0457
huber	1.345	EQP	0.2863	0.2865	-0.0002	-0.0626
huber	1.345	NEQ	0.0809	0.0809	0.0000	0.0534
bisquare	4.685	Intercept	-0.0247	-0.0247	0.0000	0.0019
bisquare	4.685	LFG	0.1041	0.1040	0.0000	0.0322
bisquare	4.685	GAP	0.0250	0.0250	0.0000	-0.0097
bisquare	4.685	EQP	0.2968	0.2968	0.0000	-0.0088
bisquare	4.685	NEQ	0.0885	0.0885	0.0000	0.0207

Weight Function	Tuning Constant	Variable	SAS SE	Macro SE	SAS - Macro	(SAS-Macro) / Macro*100%
huber	1.345	Intercept	0.0098	0.0096	0.0003	2.8737
huber	1.345	LFG	0.1897	0.1844	0.0053	2.8737
huber	1.345	GAP	0.0088	0.0085	0.0002	2.8737
huber	1.345	EQP	0.0624	0.0607	0.0017	2.8737
huber	1.345	NEQ	0.0333	0.0324	0.0009	2.8737
bisquare	4.685	Intercept	0.0097	0.0094	0.0002	2.5028
bisquare	4.685	LFG	0.1867	0.1821	0.0046	2.5028
bisquare	4.685	GAP	0.0086	0.0084	0.0002	2.5028
bisquare	4.685	EQP	0.0614	0.0599	0.0015	2.5028
bisquare	4.685	NEQ	0.0328	0.0320	0.0008	2.5028

Weight Function	Tuning Constant	Variable	SAS p-values	Macro p-values	SAS - Macro	(SAS-Macro) / Macro*100%
huber	1.345	Intercept	0.0274	0.0271	0.0003	1.2820
huber	1.345	LFG	0.6985	0.6915	0.0070	1.0128
huber	1.345	GAP	0.0072	0.0076	-0.0005	-6.2109
huber	1.345	EQP	0.0000	0.0000	0.0000	-72.0156
huber	1.345	NEQ	0.0151	0.0154	-0.0003	-2.1652
bisquare	4.685	Intercept	0.0106	0.0113	-0.0007	-6.2806
bisquare	4.685	LFG	0.5772	0.5701	0.0071	1.2423
bisquare	4.685	GAP	0.0038	0.0044	-0.0006	-13.8923
bisquare	4.685	EQP	0.0000	0.0000	0.0000	-80.8574
bisquare	4.685	NEQ	0.0069	0.0076	-0.0007	-9.4088

### WISHLIST FOR PROC ROBUSTREG

Proc RobustReg has 10 options for weight functions in M estimation, including Huber, bisquare, Andrews, Cauchy, fair, Hampel, logistic, median, Talworth and Welsch. However, SAS lacks the weight function derived from the generalized F-distribution, which the macro has.

The weight functions derived from the generalized F-distribution as discussed in McKean and Sievers (1989) include asymmetric distributions (positively and negatively skewed), symmetric ones, and distributions of either very light tails or very heavy ones. If we know in advance that the error distribution is skewed and/or heavy-tailed to a certain extent, using a proper generalized F-distribution weight function to match the shape of the error distribution may improve the

efficiency of estimation and hypothesis test. The shape of the error distribution can be captured roughly by its kurtosis and skewness, which are determined by the degrees of freedom ( $m_1$  and  $m_2$ ) of the generalized F-distribution.

The macro can first take in the residuals of a preliminary least squares analysis, to get  $m_1$  and  $m_2$ , the estimated degrees of freedom of the generalized F-distribution. It then fits the original data by using  $m_1$  and  $m_2$  as the tuning constants of the weight function derived by the generalized F-distribution.

SAS claimed that "... for the weight functions, we used a popular reference. There might be other weight functions discussed in some other papers. But for robust regression and outlier detection, it is the tuning constant(s) which play a more important role than the weight function itself. ..."

Currently "LSMEANS", "ESTIMATE" and "CONTRAST" statements are not available in Proc RobustReg, but SAS said these are on their list of things to do for later versions.

### **A BUG IN PROC ROBUSTREG**

If you write "model y = gender dose" in Proc RobustReg, where gender is a class variable, SAS will give you incorrect results for the standard errors and p-values of parameter estimates. You must put all the continuous variables first, then the class variables. This bug will be fixed in version 9.1.

### **CONCLUSION**

For M-estimation with weight function of bisquare and Huber, the results produced by these two procedures are numerically close to each other.

Although the macro does not provide comprehensive options and functionalities as does Proc RobustReg, it does provide a tool to possibly improve the efficiency of parameter estimation and hypothesis testing by matching the shape of the error distribution with the generalized F-distribution, when we know in advance that the error distribution is skewed and/or heavy-tailed. It would be nice if Proc RobustReg offered a similar option.

Also "LSMEANS", "ESTIMATE" and "CONTRAST" statements will be useful and handy tools if they're implemented in the future version of Proc RobustReg.

### **REFERENCES**

Colin Chen (2002), "Robust Regression and Outlier Detection with the ROBUSTREG Procedure", SUGI 27.

Zaman, A., Rousseeuw, P.J., and Orhan, M. (2001), Econometric Applications of High-Breakdown Robust Regression Techniques, *Economics Letters*, 71, 1-8.

J. W McKean and G. L. Sievers (1989). Rank scores suitable for analysis of linear models under asymmetric error distributions. *Technometrics*, 31:207-218.

### **ACKNOWLEDGMENTS**

The author would like to thank John K. Troxell, Bruce S. Binkowitz and Thomas P. Capizzi, Biostatistics and Research Decision Sciences, Merck & Co., Inc., for their help, suggestions, and review.

### **CONTACT INFORMATION**

Your comments and questions are valued and encouraged. Contact the author at:

Shuang Lu  
Merck & Co., Inc.  
RY34-A472  
P.O. Box 2000  
Rahway, NJ 07065-0900  
Tel: (732) 594-0502  
Fax: (732) 594-6075  
Email: [sean\\_lu@merck.com](mailto:sean_lu@merck.com)

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.