

# **Outlier Resistant Model Robust Regression**

by

**Christopher Ashley Assaid**

Dissertation submitted to the faculty of  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

**DOCTOR OF PHILOSOPHY  
IN  
STATISTICS**

Jeffrey B. Birch, Chair  
Raymond H. Myers  
Eric P. Smith  
Clint W. Coakley  
George R. Terrell

April 14, 1997  
Blacksburg, VA

Keywords: Regression, Robust, M-Estimation, Loess  
Copyright 1997, Christopher A. Assaid

# Outlier Resistant Model Robust Regression

Christopher A. Assaid

(ABSTRACT)

Parametric regression fitting (such as OLS) to a data set requires specification of an underlying model. If the specified model is different from the true model, then the parametric fit suffers to a degree that varies with the extent of model misspecification. Mays and Birch (1996) addressed this problem in the one regressor variable case with a method known as Model Robust Regression (MRR), which is a weighted average of independent parametric and nonparametric fits to the data. This paper was based on the underlying assumption of “well-behaved” (Normal) data. The method seeks to take advantage of the beneficial aspects of the both techniques: the parametric, which makes use of the prior knowledge of the researcher via a specified model, and the nonparametric, which is not restricted by a (possibly misspecified) underlying model.

The method introduced here (termed Outlier Resistant Model Robust Regression (ORMRR)) addresses the situation that arises when one cannot assume well-behaved data that vary according to a Normal distribution. ORMRR is a blend of a robust parametric fit, such as M-estimation, with a robust nonparametric fit, such as Loess. Some properties of the method will be discussed as well as illustrated with several examples.

## **Acknowledgements**

I wish to express my deepest appreciation to Dr. Jeffrey B. Birch for his extremely hard work, support, and friendship during this research. Many thanks also to the other members of my committee, Dr. Raymond H. Myers, Dr. Eric P. Smith, Dr. Clint W. Coakley, and Dr. George R. Terrell, for their time and helpful suggestions. I would also like to thank the faculty and staff of the department of statistics for the support and experiences in teaching, consulting, and research. Also to Tim Robinson, who was always there for friendship, collaboration, and encouragement.

I would also like to thank my wife Erica for her love, support, and patience, and my mother and father for their continuing support and prayers. Finally and foremost, I would like to thank God for His grace and mercy throughout this time of my life.

## Table of Contents

	page #
<b>List of Tables</b>	vi
<b>List of Figures</b>	vii
<b>Chapter 1 Introduction and Motivation</b>	
1.A Statement of the Problem	1
1.B Setting	2
1.C Proposed Research	2
<b>Chapter 2 Classical Regression Using Ordinary Least Squares</b>	
2.A Formulation of the Model	4
2.B The HAT Matrix	5
2.C Outliers	7
2.D Outlier Diagnostics in the Linear Regression Model	7
2.E Summary	9
<b>Chapter 3 Robust Parametric Regression</b>	
3.A M, Generalized-M (GM), Bounded-Influence (BI), Mallow's 1-Step (M1S) and Schweppe's 1-Step (S1S) Estimators	11
3.B Example Using M-Regression	15
3.C $\psi$ Functions	16
3.D Method of Solution for M-Estimators: Iterated Reweighted Least Squares	18
3.E Discussion	20
<b>Chapter 4 Nonparametric Regression</b>	
4.A Kernel Regression	21
4.A.1 Nadaraya-Watson Weights	22
4.A.2 r-Nearest Neighbor Weights	23
4.A.3 Gasser- Müller Weights	24
4.B Local Polynomial Regression	25
4.C Bandwidth Selection	26
4.D Example of Local Polynomial Regression	28
<b>Chapter 5 Robust Nonparametric Regression</b>	
5.A Local M-Estimation	33
5.B Locally Weighted Regression and Smoothing Scatterplots (Loess)	36
5.C M-Type Smoothing Splines	37
5.D $L_1$ Nonparametric Regression	38
5.E Practical Method: Robust Local Linear Regression (RLLR)	39
5.F Example of Robust Local Linear Regression	40

<b>Chapter 6</b>	<b>Proposed Methodology</b>	
6.A	Model Robust Regression	42
6.B	Development of Methodology	43
6.C	Example of Outlier Resistant Model Robust Regression	45
<b>Chapter 7</b>	<b>Theoretical Comparisons</b>	
7.A	Introduction	50
7.B	MSE Criteria	51
7.C	Applications of the MSE Formulas	57
7.D	Example	58
7.E	Description of Model	71
7.F	Verification of Bias and Variance Calculations	74
7.G	Fit Diagnostics	80
<b>Chapter 8</b>	<b>A Simulation Study</b>	
8.A	Introduction	82
8.B	Simulation Comparison of Procedures Using Optimal Parameters	82
8.C	Data-Driven Parameter Selection	86
<b>Chapter 9</b>	<b>Multivariate Methodology</b>	
9.A	Introduction	93
9.B	Extension of Methodology to Multivariate Technique	93
9.C	Example	95
<b>Chapter 10</b>	<b>Summary and Future Research</b>	
10.A	Summary	107
10.B	Future Research	108
10.B.1	Automatic Selection Criteria	108
10.B.2	Local Bandwidth/Mixing Parameter	109
10.B.3	Outlier Diagnostics	109
10.B.4	Confidence Intervals	109
10.B.5	Multiple Regression Extension	110
<b>References</b>		111
<b>Appendix</b>		114

## List of Tables

Table		page #
4.1	Bias and variance calculations for Kernel regression using Nadaraya-Watson weights and Gasser-Müller weights, and local linear regression using Nadaraya-Watson weights (Fan)	28
7.1	Optimal AIMSE values for procedures for model in (7.D.1) and error distribution $CN(.1, \sigma_1 = 0.75, \sigma_2 = 3.0)$	59
7.2	Empirical mse values for model in (7.D.1) with error distribution $CN(.1, \sigma_1 = 0.75, \sigma_2 = 3.0)$	71
7.3	Theoretically optimal (AIMSE) and simulated (INTMSE) mse values using the optimal smoothing parameters for the ORMRR estimator for $CN(.1, \sigma_1 = 0.75, \sigma_2 = 3.0)$	76
7.4	Comparison of AIMSE (bold) and INTMSE (non-bold) for all procedures using the optimal smoothing parameters for $CN(.1, \sigma_1 = 0.75, \sigma_2 = 3.0)$	79
8.1	INTMSE values for Loess, ORMRR, M-Regression, and RLLR estimators for $CN(.05, \sigma_1 = 0.75, \sigma_2 = 3.0)$ . Fits are based on theoretically <i>optimal</i> parameter values	84
8.2	INTMSE values for Loess, ORMRR, M-Regression, and RLLR estimators for $CN(.10, \sigma_1 = 0.75, \sigma_2 = 3.0)$ . Fits are based on theoretically <i>optimal</i> parameter values	85
8.3	INTMSE values for Loess, ORMRR, M-Regression, and RLLR estimators for $CN(.25, \sigma_1 = 0.75, \sigma_2 = 3.0)$ . Fits are based on theoretically <i>optimal</i> parameter values	86
8.4	Average bandwidths and mixing parameters selected across simulated data sets via $d_p(\theta)$ as compared with theoretically optimal values	90
8.5	Comparison of INTMSE values: theoretically optimal, simulated using optimal parameters, and simulated using parameters chosen via $d_p(\theta)$	91
9.1	EMSE values for procedures for example model in (9.1) with $\gamma = 0.75$ and error distribution $CN(.10, 1.0, 5.0)$ .	100

## List of Figures

Figure		page #
3.1	Example Using M-Regression	16
3.2	Plot of Huber's $\psi$ Function	17
3.3	Plot of Bisquare $\psi$ Function	18
3.4	Plot of Huber and Bisquare Weight Functions	19
4.1	Example of Local Linear Regression on data set with small variability	29
4.2	Example of Local Linear Regression on data set with large variability	30
4.3	Example of Local Linear Regression on data set with large variability and an outlier	31
5.1	Example of Robust Local Linear Regression on data set with outliers	40
6.1	M-Regression fit to example data with $\sigma = 0.25$ .	46
6.2	RLLR fit to example data with $\sigma = 0.25$ .	47
6.3	ORMRR fit to example data with $\sigma = 0.25$ .	48
6.4	Comparison of ORMRR and Loess fits to example data with $\sigma = 0.25$ .	49
7.1	Plot of true mean function for example model in (7.D.1), along with portion explained by a quadratic model ( $\gamma = 0.75$ )	60
7.2	Plot of nonlinear portion of model in (7.D.1) with $\gamma = 0.75$	61
7.3	Plot of OLS fit (and mean function) for data from model in (7.D.1) with $\gamma = 0.75$	63
7.4	Plot of M-Regression fit (and mean function) for data from model in (7.D.1) with $\gamma = 0.75$	64
7.5	Plot of RLLR fit to residuals from M-Regression fit for data from model in (7.D.1) with $\gamma = 0.75$ , as compared to true amount of remaining structure left in the residuals.	65
7.6	Plot of ORMRR fit (and mean function) for data from model in (7.D.1) with $\gamma = 0.75$	67
7.7	Plot of Loess fit (and mean function) for data from model in (7.D.1) with $\gamma = 0.75$	69
7.8	Plot of RLLR fit (and mean function) for data from model in (7.D.1) with $\gamma = 0.75$	70
7.9	Plot of base model to be used in simulations for varying degrees of model misspecification $\gamma$	73
9.1	Plot of example model with $\gamma = 0$ (no model misspecification for quadratic user's model)	96
9.2	Plot of data generated from example model with $\gamma = 0.75$ and error distribution $CN(0.10, 1.0, 5.0)$	97

9.3	Plot of example model with $\gamma = 0.75$ (moderate amount of model misspecification for quadratic user's model)	99
9.4	Plot of final ORMRR fit to data generated from the example model for $\gamma = 0.75$	101
9.5	Plot of smoothed residuals that resulted from M-Regression fit to data generated from example model for $\gamma = 0.75$	102
9.6	Plot of Loess fit ( $f = 0.5$ ) to data generated from example model for $\gamma = 0.75$	103
9.7	Plot of Loess fit ( $f = 0.25$ ) to data generated from example model for $\gamma = 0.75$	104
9.8	Plot of M-Regression fit to data generated from example model for $\gamma = 0.75$	105