Chapter 1

Introduction and Motivation

1.A Statement of the Problem

Recently, regression methods (both parametric and nonparametric) that are robust to outliers have become popular, and the utilization of these techniques in a method that is robust to model misspecification could prove to be a very useful and widely applicable tool for statistical data analysis. The efficacy of such a technique stems from the presence of random and unpredictable data in real world applications. The traditional assumptions that lead to the use of classical methods (model correct, normal errors with mean zero and constant variance) are generally not completely satisfied, if not violated altogether.

Real world data tend to be composed of some fraction that is of little or no use. This reality can occur in a number of ways, including measurement error, transcription error, or some unknown or uncontrollable phenomenon that occurred at the time of recording. Factors beyond the control of the researcher that cause data to tend toward extreme values led to the exploration of so-called robust methods that provide a means of identifying and dealing with extreme observations.

As a general rule, the robust techniques that currently exist provide a means of identification of the outliers, and an automated means of downweighting these observations. The robust *nonparametric* methods include only weak assumptions, but their reliance on data driven techniques results in a more variable fit than their parametric counterparts. The robust parametric procedures have more stable predicted values, but these methods are very model dependent and rely heavily on the assumption of a *correctly* specified model.

In general, it is impossible to know the true underlying model from which a dataset originates unless the data are contrived. Generally, the best the researcher can hope for is that the specified model has a form similar to that of the true underlying model. This results in a dilemma.

Should the analyst use the specified model, which may or may not be sufficient, or use a nonparametric technique that does not rely on a specified model form but results in more variable fitted values?

1.B Setting

In the current research, it will be assumed that there is only one independent variable x, and that other variables which may be in the specified model are polynomial functions of x. The multiple regressor case will be discussed and illustrated briefly in Chapter 9. We will also assume that the model is linear in its parameters, and thus can be written in the form $\mathbf{y} = \mathbf{X}^{P}\beta + \varepsilon$, where $\mathbf{X}^{P}$ is the parametric model matrix. The error structure will be assumed to be symmetric and heavy-tailed, such as errors generated from a contaminated normal distribution, denoted $\mathrm{CN}(\pi, \sigma_1, \sigma_2)$. That is, ε_i has mean 0, variance σ_1^2 with probability 1-π, and has variance σ_2^2 with probability π (note that, in general, $\sigma_1^2 < \sigma_2^2$ and π is relatively small).

1.C Research

The goal of this research is to first make use of a method which has been developed to utilize the advantages of both parametric and nonparametric techniques, which results in the elimination of the need to choose from one of these procedures. We then ultimately extend this method to handle the situation in which the data are not necessarily well-behaved and may contain extreme observations. The method will be termed Outlier Resistant Model Robust Regression (ORMRR), a method that is robust to a misspecified parametric model, and also robust to outlying observations.

The procedure takes on a form which makes use of residuals to obtain a fit that is both model and outlier robust. Each of these is based on the groundbreaking model robust techniques developed first by Einsporn and Birch (1988) and (1993), and further developed by Mays and Birch (1996) and Nottingham and Birch (1996).

Chapter 2 briefly summarizes some of the assumptions and methodology of classical regression analysis, and highlights some of its shortcomings. Chapter 3 will introduce robust parametric methods in the one variable regression setting. Chapter 4 will cover some of the latest methodology for nonparametric regression methods, and Chapter 5 will extend these methods to modern nonparametric methods that are robust to outliers. Chapter 6 will cover the proposed methodology in depth along with some examples. Chapter 7 develops (and verifies by way of simulations) the asymptotic bias and variance calculations for the competing methods. In addition, comparisons are offered between the optimal fits obtained from the competing procedures. Chapter 8 offers comparisons of the procedures using simulations, and details and evaluates a data-driven parameter selection criterion that we propose be used in conjunction with ORMRR. Chapter 9 briefly describes the extension of ORMRR to the multiple regressor setting. An example is offered along with comparisons with the competing procedures. Chapter 10 summarizes the research and describes some topics of future research.