Chapter 1 : Introduction and Motivation

1.A Statement of the Problem

The use of regression to describe a response variable of interest as a function of a given set of independent regressor variables, $(X_1, X_2, \ldots X_k)$, is common practice. In general, the i^{th} value of the response variable, y, is modeled as follows:

$$y_i = h(X_{1i}, X_{2i}, \cdots, X_{ki}) + \varepsilon_i, \qquad i = 1, 2, \ldots, n. \qquad (1.A.1)$$

The function h is known as the regression function and ε denotes the random error in the process. For purposes of this research, we will assume that the random errors follow a Gaussian (normal) distribution.

In parametric regression, h is assumed to take on some known parametric form, the parameters of which are estimated using the sample data. A standard assumption in normal theory regression is homogeneity of variance across the data. However, in many applications, the variance is non-constant across the domain of the data and the analysis must be adjusted to account for the variance heterogeneity. In order to account for the non-constant variance (ex. weighted least squares), one must know the values of σ_i^2. However, it is rare that the values of the individual variances are known, hence it becomes necessary to estimate them. A wealth of literature has been devoted to variance estimation.

A method of variance estimation that is currently receiving a great deal of attention is variance modeling. The idea is that process variability changes systematically and often smoothly with a set of predictors. Therefore, similar to modeling the process mean as a function of 'mean' regressors $(X_1, X_2, \ldots X_k)$, the variance is modeled as a function of "variance" regressors, $(Z_1, Z_2, \ldots Z_l)$. When the process mean and process variance are estimated simultaneously, the process is said to be described in terms of a "dual" model.

The parametric regression viewpoint assumes that both the mean and variance functions have known parametric forms and write the dual model as

Means Model : $y_i = h(X_{1i}, X_{2i}, \cdots, X_{ki}; \beta) + g^{1/2}(Z_{1i}, Z_{2i}, \cdots Z_{li}; \theta)\,\varepsilon_i$

Variance Model : $\sigma_i^2 = g(Z_{1i}, Z_{2i}, \cdots, Z_{li}; \theta)$

where h and g are, respectively, the true, underlying process mean and process variance functions and β and θ represent the sets of parameters for each function. If we consider the variance model above as a regression model in which the values of σ_i^2 are taken to be the responses and $Z_1, Z_2, \cdots, Z_l$ are regressors, we notice that the responses are unknown. However, a regression analysis with unknown responses is ludicrous. Solutions to this problem

depend on whether or not replication exists at each of the n data points. More attention will be given to this in Chapter 3 but for purposes of this research, the assumption is that no replication exists. In the case of non-replication and normal errors, the responses in the variance model are typically taken to be the squared residuals from the means model fit. The parameters in the dual model are estimated within the context of an iterated, generalized least squares (GLS) algorithm. In this procedure, the estimate for h is used to estimate g and then the estimate of g is used to update the estimate of h. The procedure continues until convergence of the means model parameters. It is clear that if either or both of the functions, h and g, are misspecified, the quality of their respective inferences suffers.

The dual model has also been approached from a nonparametric viewpoint in which both the mean function h, and variance function g, are assumed to be unknown. There are two main advantages to the nonparametric approaches. First, a nonparametric fit is more flexible than the parametric fit in that it is not confined to a specified form. This enables the nonparametric fit to capture certain "wrinkles" in the data that a parametric model cannot capture. Second, there have been proposals for nonparametric variance estimation which do not first require a fit to the means model. Thus, unlike parametric variance estimation which depends on the means model residuals, the nonparametric variance estimate is not adversely affected if the mean is badly estimated.

Nonparametric procedures, however, are not a panacea in that they too have disadvantages. Since the nonparametric procedures involve no information from the user, they often fit irregular patterns in the data. Secondly, the fact that nonparametric estimates are not anchored by a stable, specified form often causes the fit to be too variable.

If the user can correctly specify parametric models for both the mean and variance, then the parametric approach is deemed best. However, if either or both of the models are misspecified, the nonparametric procedures become more appealing. The problem lies with the fact that the forms of the true underlying models are rarely known and it is difficult to determine the best approach for analysis. Also, in many cases, parametric models do well at capturing the basic structure of the data but there may be important "bumps" or "wrinkles" that the parametric models cannot capture. This suggests the need for a "middle-of-the-road" technique which uses as much of the user's parametric knowledge as possible but has the ability to detect the "bumps" and "wrinkles" that the parametric model fails to capture.

Mays and Birch (1996) and Nottingham and Birch (1996) have demonstrated a semi-parametric method that is effective if model misspecification is present in the one regressor setting. We propose to extend their ideas to the dual modeling setting and offer an approach which is robust to misspecification of the mean and/or variance model. This robust approach utilizes the parametric knowledge from the user but also provides estimates for the mean and variance which are flexible enough to capture the departures from the true models.

This research is organized as follows. In Chapter 2 we discuss estimation of the means model from a parametric as well as a nonparametric viewpoint. We then use these discussions to motivate the development of the ideas of Mays and Birch (1996) and Nottingham and Birch (1996). In Chapter 3, variance estimation is discussed from both parametric and nonparametric viewpoints within the context of dual modeling. Chapter 4 describes the proposed dual model robust regression procedure (DMRR). This procedure offers a dual modeling technique which is robust to means and/or variance model misspecification. In Chapter 5 theoretical expressions for

the bias and variance of the mean and variance fits for the competing dual modeling procedures are derived. In Chapters 6 and 7 DMRR is compared to its parametric and nonparametric counterparts through three examples. Comparisons in Chapters 6 and 7 are based on both the theoretical and simulated integrated mean squared errors for the mean and variance fits of each of the procedures. In Chapter 8 a data-driven technique is proposed for obtaining the bandwidths and mixing parameters used in DMRR. Finally, Chapter 9 outlines some issues which may be considered in future research.