Chapter 9: Summary and Future Research

Dual model robust regression (DMRR) appears to provide an excellent alternative to the traditional parametric and nonparametric approaches to the dual modeling problem. Theoretical results presented in Chapters 6 and 7 suggest that, with respect to its parametric and nonparametric counterparts, DMRR improves the average mean squared error for both the mean and variance fits. This chapter will provide a summary of the current research and then propose several items which are worthy of consideration in future research.

9.A Summary

Dual model regression techniques are generally incorporated when the researcher feels that both the mean and variance of a particular process change systematically and smoothly with sets of regressor variables (one set influencing the mean and another influencing the variance). The parametric approach to dual modeling assumes that the underlying mean and variance functions can be adequately expressed through parametric mathematical models. A major problem associated with parametric dual modeling is that its success depends heavily on the user's ability to correctly specify the form of the underlying mean and variance functions. If the specified functions are inadequate for even part of the data, the resulting mean and variance fits will be characterized by bias (due to model inadequacy) and inferences will become adversely affected.

Unlike its parametric competitor, nonparametric dual modeling techniques are not susceptible to the user's misspecification of the underlying mean and variance models. The nonparametric techniques only assume that the underlying mean and variance are determined by unknown, smooth functions. Thus, no models are specified by the user and the fits for both the mean and variance rely solely on the data. Without the stability of specified models for the mean and variance, the nonparametric fits are often too jagged and fit too closely to irregular trends in the data (fits are highly variable).

DMRR has been developed with the intention of providing a "hybrid" approach to the dual modeling problem. The intent of DMRR is to use as much of the user's parametric information regarding the process as possible while still allowing for important deviations from the parametric portion to be captured in the final fit. This is accomplished by mixing parametric and nonparametric fits in both the mean and variance estimates. Thus, by allowing for an augmentation to the user's specified parametric model, DMRR seeks to avoid the bias problems encountered in parametric dual modeling. In taking the user's parametric knowledge into consideration, DMRR seeks to provide more stability (less variance) to the final fit than would occur if the user relied solely on nonparametric techniques.

The performance of DMRR, in relation to the traditional parametric and nonparametric dual modeling techniques, was compared for three basic scenarios. These scenarios were as follows: 1. The researcher correctly specifies the form of the underlying variance function but the specified means model is insufficient. 2. The researcher correctly specifies the form of the means model but the variance model is under-specified

by leaving out an important term. 3. The user incorrectly specifies the forms of both the underlying mean and variance functions. In all three cases, DMRR outperformed its parametric and nonparametric competitors with respect to the integrated mean squared error values of both the mean and variance estimates. The following is a list of the more important observations from Chapters 6, 7 and 8:

1. When the means model prescribed by the researcher is less than adequate for even part of the data, parametric dual modeling is highly ineffective due to the bias in the means fit. The nonparametric procedures (residual-based and difference-based) provide improvements over the parametric technique when the means model is misspecified but the fits tend to quite variable. DMRR provides substantially lower MMIMSE and VMIMSE values than the other dual modeling procedures for all degrees of means model misspecification considered and DMRR's performance improves as *n* increases.
2. If the researcher moderately misspecifies the means function ($\gamma \geq 5.0$), plots of the residuals from the means fit (an OLS means fit) may suggest heteroscedasticity in the data set when in fact the variance is constant across the domain of the data. The use of a model robust fit to the means function will more than likely eliminate the possibility of using dual modeling techniques in the presence of constant variance. This concept can be further explored through more simulation study. It should be noted that the use of DMRR in the presence of constant variance produces consistently lower MMIMSE and VMIMSE values than ordinary least squares (OLS) except when $\gamma = 0.0$.

3. Parametric dual modeling is less sensitive to misspecification of the variance model than it is to misspecification of the means model. Nevertheless, variance misspecification does adversely affect the performance of parametric dual modeling. The variance model robust component of DMRR, known as variance model robust regression (VMRR), appropriately mixes the parametric and nonparametric fits when the user misspecifies the form of the variance function. That is, when the user misspecifies the true form of the underlying variance function, the optimal variance model mixing parameter (λ_{σ_o}) is 1, or close to 1 and when the true form is correctly specified by the user, λ_{σ_o} is 0, or close to 0. Due to the effectiveness of VMRR, the dual model robust regression technique offers lower MMIMSE and VMIMSE values than the other dual modeling procedures in the case of variance model misspecification.

4. DMRR appears to also be robust to the situation in which the user misspecifies the forms of both the mean and variance functions as it produces consistently lower MMIMSE and VMIMSE values than its parametric and nonparametric competitors in this scenario.

5. Based on 300 Monte Carlo repetitions, the asymptotic formulas derived in Chapter 5 for the bias and variance of the parametric, residual-based nonparametric, and DMRR estimates of the mean and variance functions appear to be valid, especially as *n* increases. The asymptotic formulas for the bias and variance of the nonparametric difference-based variance estimate appear to grossly underestimate the true variability of the difference-based variance fit.

6. The weighted PRESS** procedure, which is used to determine the optimal bandwidths and mixing parameters for DMRR, has been shown to be effective in terms of mean squared error relative efficiencies for the examples considered in this research. It should be noted however, that the weighted PRESS** tends to underestimate the bandwidth and mixing parameters used in the VMRR fit.

Based on the summaries mentioned above, it is apparent that DMRR offers an improvement to the existing parametric and nonparametric approaches to the dual modeling problem. In the next three sections, various avenues are proposed that we feel are worthy of future research.

9.B Multiple Regression

The current research has been performed within the framework of the one-regressor setting. Thus, an obvious extension to this research would be study regarding higher dimensional applications in multiple regression. The key to this extension will be extending the nonparametric portions of the DMRR mean and variance fits to multiple regression. Recall that the nonparametric fits are local linear regression fits which are based on kernel weights. The kernel weights at a point x_i are determined by the distance measure $(x_i - x_j)$. The extension of the kernel weighting scheme to higher dimensions is a simple one in that the kernel weights would now be based on a multidimensional distance measure such as $\left\| \mathbf{x}_i - \mathbf{x}_j \right\|$.

9.C Model Robust Generalized Estimating Equations

The use of generalized estimating equations (GEE's) (Liang and Zeger (1986)) has become a common technique in many fields to account for correlation structure existing in the data. The theory of GEE's is based on the user specifying a 'working correlation' matrix in which residuals from the fit to the raw data are used to estimate the variance and covariance parameters. Since the residuals from the fit to the raw data are used to estimate the variance and covariance parameters, it is pertinent that the user specifies the correct model for the raw data. Model misspecification likely leads to the same type of problems encountered in parametric dual modeling when the mean and/ or variance models have been misspecified. Thus, an area of study that appears interesting is the impact of model misspecification on inferences made in GEE data analysis.

9.D DMRR Inference

The focus of this research has been to provide model robust estimates of both the underlying means function and the underlying variance function. Now that the methodology has been developed for obtaining such estimates, it is desirable to have a measure of the accuracy and precision of these estimates. This can be accomplished by deriving expressions for the confidence intervals on the mean and variance fits. The

derivation of these expressions should be relatively straightforward. Inherent in the development of these expressions is the need for distributional assumptions regarding the individual mean and variance estimates.

9.E Data-Driven Selection of Bandwidths and Mixing Parameters

The most important components of dual model robust regression are the choices of the bandwidths (b_μ and b_σ) and the mixing parameters (λ_μ and λ_σ). Incorrectly choosing the values of these parameters can adversely affect the DMRR estimates of the mean and variance functions. In Chapter 8, the cross-validation technique known as WPRESS** was shown to be a promising candidate for bandwidth and mixing parameter choice. The literature is rich with numerous other methods for bandwidth selection in nonparametric regression and it is conceivable that some of these methods could be adapted for bandwidth and mixing parameter selection for the DMRR procedure. Another type of bandwidth selection that has received a great deal of attention in the literature is the use of local bandwidths in nonparametric function estimation. The idea is to smooth more (large bandwidth) for areas of the data in which there is large variability and to smooth less (small bandwidth) for areas in which the variance is small. This idea could feasibly be extended to DMRR through use of both local bandwidths and local mixing parameters. It should be noted however that local bandwidths appear to work well in large sample settings but the general consensus is that global bandwidths are more appropriate when dealing with small samples. It is clear that there are many avenues which should be explored for future research in this area of data-driven bandwidth and mixing parameter selection.