A Semiparametric Approach to Dual Modeling

Timothy J. Robinson[a*], Jeffrey B. Birch[b], and B. Alden Starnes[c]

[a]*Department of Statistics, University of Wyoming,*
Laramie, WY 82071-3332, U.S.A.

[b]*Department of Statistics, Virginia Polytechnic Institute and State University,*
Blacksburg, VA 24061-0439, U.S.A

[c]*Department of Mathematics, Carson-Newman College,*
Jefferson City, TN 37760, U.S.A

Abstract

In typical normal theory regression, the assumption of homogeneity of variances is often not appropriate. When heteroscedasticity exists, instead of treating the variances as a nuisance and transforming away the heterogeneity, the structure of the variances may be of interest and it is desirable to model the variances. Modeling both the mean and variance is commonly referred to as dual modeling. In parametric dual modeling, estimation of the mean and variance parameters are interrelated. When one or both of the models (the mean or variance model) are misspecified, parametric dual modeling can lead to faulty inferences. An alternative to parametric dual modeling is nonparametric dual modeling. However, nonparametric techniques often result in estimates that are characterized by high variability and ignore important knowledge that the user may have regarding the process. We develop a dual modeling approach [Dual Model Robust Regression (DMRR)], which is robust to user misspecification of the mean and/or variance models. Numerical and asymptotic results illustrate the advantages of DMRR over several other dual model procedures.

KEYWORDS: Robustness; Variance modeling; Mixing; Model misspecification; Asymptotic convergence

*

1. Introduction

* Corresponding author. Department of Statistics, University of Wyoming, Dept. 3332
1000 E. University Avenue, Laramie, WY 82071, Tel: 307-766-5108, Email: tjrobin@uwyo.edu

When estimating a regression function, nonparametric techniques such as local polynomial regression tend to be useful when the trend in the data cannot be captured parametrically. However, nonparametric fits are often characterized by too much variability and for a small data set, the estimate tends to fit the data too closely. When there is no replication and heteroscedasticity is present, residuals from the estimated mean are often used to model the response variance. However, in small to moderately sized data sets, if a nonparametric smooth is utilized for the mean and observations are fit too closely, there may not be enough information in the residuals to adequately estimate the variance function. Parametric techniques, on the other hand, provide a less variable estimate since the fits are anchored by the stability of the user supplied model. However, the proposed parametric form may be inadequate over certain regions of the data, resulting in biased fits. Bias in the estimate of the mean results in contaminated residuals with which to estimate the variance function [see Robinson and Birch (2000)].

Einsporn and Birch (1993) propose a semiparametric method for modeling the mean response for assumed constant error variance. Their technique, model robust regression 1 (MRR1), combines parametric and nonparametric fits to the raw data in a convex combination via a mixing parameter, λ. Mays, Birch, and Einsporn (2000) introduced model robust regression 2 (MRR2) as an improvement for modeling the mean with constant variance. Similar to MRR1, MRR2 combines a parametric fit and a nonparametric fit via a mixing parameter; however, the parametric fit is to the raw data whereas the nonparametric fit is to the residuals from the parametric estimate. In both cases, MRR1 and MRR2, the result is an estimated mean response which is robust to parametric model misspecification and has less variation than a purely nonparametric estimate. Mays, Birch, and Starnes (2001), henceforth referred to as MBS, present asymptotic results for the MRR techniques. In this manuscript we extend the MRR techniques to the dual modeling problem where one is not only interested in modeling the underlying mean function but is also interested in

simultaneously estimating the underlying variance function. The robust estimate to the mean is termed *means model robust regression* (MMRR) and the robust variance fit is termed *variance model robust regression* (VMRR).

Holst, Hössjer , Björklund , Ragnarson, and Edner (1996) discuss the use of local polynomial regression for evaluation of the concentration of atmospheric atomic mercury measured with LIDAR [see Sigrist (1994)]. The data are plotted in Figure 1(a). Figure 1(b) displays the MMRR fit to the LIDAR data set. Using the squared MMRR residuals, the VMRR variance fit is then obtained. Figure 1(c) shows the squared MMRR residuals along with the proposed VMRR variance estimate, a local polynomial variance fit, and a parametric variance fit. Advantages of the proposed semiparametric fitting are clearly seen in Figure 1(c). Note that the nonparametric (local polynomial) fit's variability is seen in the 'wiggles' when X is between 650 and 730 and the parametric fit (based on gamma regression) is not flexible enough to capture certain anomalies in the data. The VMRR estimate provides a 'middle of the road' fit to the data as it achieves an optimal balance of bias and variance.

In the next section we present a general model framework for the dual model and then review the various parametric and nonparametric approaches to dual modeling. It is important to note that approaches to dual modeling often depend upon whether or not replication is present. Here, we consider experimental studies which do not involve replication. In Section 3, we provide the details of *dual model robust regression* (DMRR). In Section 4, the asymptotic properties of DMRR are discussed and in Section 5 we discuss data-driven bandwidth and mixing parameter selection. In Section 6 the proposed semiparametric approach is compared to popular parametric and nonparametric competitors via a simulation study.

2. Overview of Dual Model Regression

A common approach to heteroscedastic data is to assume that, like the mean, the variance changes systematically and smoothly with a set of regressors (potentially different from those influencing the mean). In general, the dual model can be expressed as

Means Model: $$y_i = h\left(x_{1i} \ldots x_{ki}\right) + g^{1/2}\left(z_{1i} \ldots z_{li}\right)\varepsilon_i$$

$$= h\left(\mathbf{x}_i^*\right) + g^{1/2}\left(\mathbf{z}_i^*\right)\varepsilon_i$$

Variance Model: $$\sigma_i^2 = g\left(\mathbf{z}_i^*\right)$$

where ε_i is the error term at the i^{th} observation (i = 1, …, n), $\mathbf{x}_i^*$ is a k x 1 vector of means model regressors, and $\mathbf{z}_i^*$ denotes the l x 1 vector of variance model regressors. It is assumed that the ε_i are *i.i.d.* N(0,1). There are several approaches to the analysis of a dual model, depending on whether or not the researcher expresses the mean and/or variance functions in closed form.

Historically, dual modeling has been approached in the following ways:

1. *Parametric Dual Modeling*: A purely parametric approach involves the user specifying functional forms for both the mean and variance functions. In what could be considered an off-the-shelf parametric model, the user assumes a *known* linear model for the process mean $\left[h\left(\mathbf{x}_i^*\right) = \mathbf{x}_i'\boldsymbol{\beta}\text{, where } \mathbf{x}_i' = \left(1\ x_{1i} \ldots x_{ki}\right)\right]$, and a *known* log-linear relationship for the variance $\left[g\left(\mathbf{z}_i^*\right) = \exp\left\{\mathbf{z}_i'\boldsymbol{\theta}\right\}\text{, where } \mathbf{z}_i' = \left(1\ z_{1i} \ldots z_{li}\right)\right]$. Regarding notation, $\boldsymbol{\beta}$ is a (k+1) x 1 vector of means model parameters and $\boldsymbol{\theta}$ is a (l + 1) x 1 vector of variance model parameters. Aitkin (1987) proposed an iterative estimation procedure for this dual model. The iterative analysis begins with an initial ordinary least squares (OLS) fit to the mean and then uses gamma regression to fit an exponential function to the squared OLS residuals. The mean and variance model parameters are re-estimated via estimated weighted least squares (EWLS) and the entire iterative process continues until convergence of the parameter estimates in the means model. At convergence, these mean and

variance model estimates are the maximum likelihood estimates provided that the errors are normally distributed. Valid inferences from such an analysis depend heavily on the assumption that the specified forms of h and g are sufficient across the entire range of the data. If h and g are misspecified, any inferences from the analysis become suspect.

2. *Nonparametric Residual-Based Dual Modeling:* The mean is estimated parametrically or nonparametrically, depending on the researcher's knowledge regarding the underlying form of the means model, and then a nonparametric technique is used to smooth the means model residuals for variance estimation. Specifically, Carroll (1982) uses a parametric model for h while Hall and Carroll (1989) consider both parametric and nonparametric estimation of h and then use kernel regression [see Härdle (1990)] to estimate g. Carroll and Ruppert (1988) and Ruppert, Wand, Holst and Hössjer (1997) discuss the use of bias-corrected local polynomial regression to estimate the underlying variance function, g. Ruppert et. al (1997). point out that local polynomial methods for smoothing are preferable to kernel smoothing because of their automatic boundary bias correction and adaptivity to unequally spaced designs.

3. *Nonparametric Difference-Based Dual Modeling:* Müller and Stadtmüller (1987 and 1993) propose a purely nonparametric estimate of the dual model in which the mean is estimated via kernel regression and then a difference-based estimate of the variance is used. Difference-based variance estimation is appealing because the estimate of the variance does not depend on the estimated mean function. Instead, the variance is estimated using a local polynomial smooth of the squared pseudo-residuals. Pseudo-residuals are constructed as weighted averages of fixed neighborhoods of the observed responses (the y_i's). Pseudo-residuals were originally proposed for homoscedastic data by Gasser, Sroka, and Jennen-Steinmetz (1986).

3. Dual Model Robust Regression

The proposed technique extends the MRR procedures (MRR1 and MRR2) of Einsporn and Birch (1993) and Mays, Birch, and Einsporn (2000), respectively, to the unreplicated dual modeling problem. Dual model robust regression (DMRR) provides an estimate of the mean (MMRR) as well as an estimate of variance (VMRR) that is simultaneously robust to a misspecified means model and a misspecified variance model. The MRR procedures are based on the assumption that any regression model can be expressed as

$$\begin{aligned} y_i &= h\left(\mathbf{x}_i^*\right) + \varepsilon_i \\ &= \mathbf{x}_i'\boldsymbol{\beta} + f\left(\mathbf{x}_i^*\right) + \varepsilon_i \end{aligned}$$

where $\mathbf{x}_i'\boldsymbol{\beta}$ is the user-supplied parametric portion (a linear model is used here, though, in general the user-supplied model may be nonlinear) and $f\left(\mathbf{x}_i^*\right)$ denotes a "lack-of-fit" component. The ε_i are assumed to be independent and normally distributed with mean zero and constant variance, σ^2. The MRR1 estimate of the mean is then given in matrix notation as

$$\begin{aligned} \hat{\mathbf{y}}^{(\mathrm{MRR1})} &= \lambda\hat{\mathbf{y}}^{(\mathrm{LLR})} + (1-\lambda)\hat{\mathbf{y}}^{(\mathrm{OLS})} \qquad (3.1) \\ &= \left[\lambda\mathbf{H}^{(\mathrm{LLR})} + (1-\lambda)\mathbf{H}^{(\mathrm{OLS})}\right]\mathbf{y} = \mathbf{H}^{(\mathrm{MRR1})}\mathbf{y} \end{aligned}$$

where $\lambda \in [0,1]$, $\hat{\mathbf{y}}^{(\mathrm{OLS})}$ denotes the *nx1* vector of ordinary least squares (OLS) fits to the data, $\hat{\mathbf{y}}^{(\mathrm{LLR})}$ represents the *nx1* vector of local linear regression (LLR) fits to the data (LLR is used here, though, in general, any nonparametric method utilizing a "smoother" matrix may be used), $\mathbf{H}^{(\mathrm{OLS})}$ is the OLS hat matrix, and $\mathbf{H}^{(\mathrm{LLR})}$ and $\mathbf{H}^{(\mathrm{MRR1})}$ are the "smoother" matrices for local linear regression and MRR1, respectively. The mixing parameter, λ, is chosen to adjust the parametric model fits, which may be inadequate, with an appropriate amount of structure contained in the LLR fit to the data. The choice of λ involves a bias-variance trade-off and is discussed in Section 5.

The MRR2 approach of Mays, Birch, and Einsporn (2000) also combines a parametric fit and a nonparametric fit via a mixing parameter; however, the parametric fit is to the raw data (as in MRR1) while the nonparametric fit is to the residuals from the parametric fit. The *nx1* vector of residuals, **r**, represents the structure in the mean which is not captured by the user specified parametric model. The residuals are fit nonparametrically via LLR resulting in the following vector of smoothed residuals

$$\hat{\mathbf{r}} = \mathbf{H}_r^{(\mathrm{LLR})}\mathbf{r}$$

where $\mathbf{H}_r^{(\mathrm{LLR})}$ is the LLR smoother matrix with the responses being the residuals from the OLS fit to the raw data. The MRR2 estimate of the mean is then obtained by adding a portion, λ, of the LLR smoothed residuals back to the original OLS fit, yielding

$$\begin{aligned}\hat{\mathbf{y}}^{(\mathrm{MRR2})} &= \hat{\mathbf{y}}^{(\mathrm{OLS})} + \lambda\hat{\mathbf{r}}^{(\mathrm{LLR})} \\ &= \left[\mathbf{H}^{(\mathrm{OLS})} + \lambda\mathbf{H}_r^{(\mathrm{LLR})}\left(\mathbf{I} - \mathbf{H}^{(\mathrm{OLS})}\right)\right]\mathbf{y} = \mathbf{H}^{(\mathrm{MRR2})}\mathbf{y}\end{aligned} \qquad (3.2)$$

where $\lambda \in [0,1]$. The choice of λ is discussed in Section 5. MBS show that the MRR1 and MRR2 estimates of the mean possess superior integrated mean squared error properties over both $\hat{\mathbf{y}}^{(\mathrm{OLS})}$ (if the user's model is misspecified) and $\hat{\mathbf{y}}^{(\mathrm{LLR})}$ (the fits obtained by LLR or any other appropriate nonparametric smooth). In the discussion to follow, we extend the MRR approaches to the dual modeling problem. The proposed DMRR technique contributes to the growing list of literature dealing with semiparametric regression. Such references include work by Speckman (1988), Burman and Chaudhuri (1992), Rahman, Gokhale, and Ullah (1997), and Fan and Ullah (1999), Ruppert, Wand, and Carroll (2003), among many others.

3.1 The Model

In developing the DMRR technique, it is helpful to consider the process mean and variance as functions which can be expressed in two components: a user supplied parametric component and a "lack of fit" component which represents the portion of the mean and variance functions which cannot be captured parametrically. We decompose the mean into a parametric linear portion and a portion not captured by the linear specification. In estimating the variance we consider an exponential specification and a portion not captured by this specification. It should be pointed out that DMRR is not limited to these particular parametric specifications but for simplicity of presentation, these functions are explicitly considered here. The true underlying dual model can then be written as follows:

Means Model: $$y_i = h\left(\mathbf{x}_i^*\right) + g^{1/2}\left(\mathbf{z}_i^*\right)\varepsilon_i$$

$$= \mathbf{x}_i'\boldsymbol{\beta} + f\left(\mathbf{x}_i^*\right) + g^{1/2}\left(\mathbf{z}_i^*\right)\varepsilon_i$$

Variance Model: $$\sigma_i^2 = g\left(\mathbf{z}_i^*\right)$$

$$= \exp\left\{\mathbf{z}_i'\boldsymbol{\theta}\right\} + l\left(\mathbf{z}_i^*\right).$$

The lack of fit components are given by $f\left(\mathbf{x}_i^*\right)$ in the means model and $l\left(\mathbf{z}_i^*\right)$ in the variance model and f and l are considered to be unknown (smooth) regression functions. We assume a one-regressor dual model, implying that $\mathbf{x}_i = \left(1\ x_i\ x_i^2\ ...\right)' = \mathbf{z}_i = \left(1\ z_i\ z_i^2\ ...\right)'$. It should be noted that the DMRR technique can easily be extended to situations involving more than one regressor and other types of parametric specifications for the mean and variance.

3.2 Model Robust Mean Estimation

To estimate the means model, we propose *means model robust regression* (MMRR). MMRR is a simple extension of the MRR2 procedure where the OLS mean estimate in MRR2 is replaced with an estimated weighted least squares (EWLS) estimate. Weights are taken to be the

estimated variances at each of the *n* data points (denoted below as the diagonal elements of the matrix $\hat{\mathbf{V}}$). Using local linear regression (LLR) to smooth the EWLS residuals, the *n x* 1 vector of means model robust estimates (MMRR) is then given as:

$$\hat{\mathbf{y}}^{(\text{MMRR})} = \mathbf{H}^{(\text{MMRR})}\mathbf{y}$$

where

$$\mathbf{H}^{(\text{MMRR})} = \left[\mathbf{H}^{(\text{EWLS})} + \lambda_\mu \mathbf{H}_{b_\mu}^{(\text{LLR})}\left(\mathbf{I} - \mathbf{H}^{(\text{EWLS})}\right)\right] .$$

Here, $\mathbf{H}^{(\text{EWLS})} = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}$ and $\mathbf{H}_{b_\mu}^{(\text{LLR})} = \begin{bmatrix} \mathbf{h}_{b_{\mu 1}}^{(\text{LLR})\prime} \\ \vdots \\ \mathbf{h}_{b_{\mu n}}^{(\text{LLR})\prime} \end{bmatrix}$ where

$\mathbf{h}_{b_{\mu i}}^{(\text{LLR})\prime} = \mathbf{x}_i^{(\text{LLR})\prime}\left(\mathbf{X}^{(\text{LLR})\prime}\mathbf{W}^{(\text{LLR})}(\mathbf{x}_i)\mathbf{X}^{(\text{LLR})}\right)^{-1}\mathbf{x}_i^{(\text{LLR})}\mathbf{W}^{(\text{LLR})}(\mathbf{x}_i)$, $\mathbf{X}^{(\text{LLR})} = \begin{bmatrix} \mathbf{x}_1^{(\text{LLR})\prime} \\ \vdots \\ \mathbf{x}_n^{(\text{LLR})\prime} \end{bmatrix}$ is the model matrix

for LLR, with $\mathbf{x}_i^{(\text{LLR})\prime} = (1 \;\; x_i)$, and $\mathbf{W}^{(\text{LLR})}(\mathbf{x}_i) = diag\left(\mathrm{h}_{i1}^{(\text{ker})}, ..., \mathrm{h}_{in}^{(\text{ker})}\right)$ where

$\mathrm{h}_{ij}^{(\text{ker})} = \dfrac{\mathrm{K}\left(\dfrac{\mathrm{x}_i - \mathrm{x}_j}{b_\mu}\right)}{\sum_{j=1}^{n}\mathrm{K}\left(\dfrac{\mathrm{x}_i - \mathrm{x}_j}{b_\mu}\right)}$. The function K(.) denotes a univariate kernel function. The means model mixing parameter is $\lambda_\mu \in [0,1]$ and b_μ denotes a global bandwidth used to smooth the EWLS residuals. Choice of λ_μ and b_μ is discussed in Section 5. The choice of kernel function is not crucial to the performance of the nonparametric estimator [see Simonoff (1996)] and thus, for convenience, we use the simplified Gaussian kernel, $K(u) = e^{-u^2}$ for all nonparametric estimators discussed.

The estimated variances on the diagonal of $\hat{\mathbf{V}}$ will be obtained from the model robust variance estimate, which is discussed in the next subsection. The means model mixing parameter, λ_μ, serves the same purpose as that of λ from the MRR2 procedure in that it increases from 0 to 1 as the amount of model misspecification in the parametric increases. The subscript 'μ' is to distinguish this mixing parameter from the one that will be used in the robust variance estimate.

3.3 Model Robust Variance Estimation

To estimate the variance function in the dual model, we propose *variance model robust regression* (VMRR), a residual-based variance estimate. The VMRR procedure can be thought of as a robust extension to the parametric estimate proposed by Aitkin (1987). Assuming a correct specification of the means model, Aitkin estimates the variance model parameters ($\boldsymbol{\theta}$) via joint maximum likelihood. Aitkin showed that this approach is equivalent to a gamma regression in which one regresses the squared residuals from the means fit on some appropriate function of the variance regressors. Similar to Aitkin we choose the exponential function. From standard generalized linear models (GLM) theory [see McCullagh and Nelder (1994)], it follows that the estimated variance model parameters at the s^{th} iteration are given by

$$\hat{\boldsymbol{\theta}}_s^{(\text{GLM})} = \hat{\boldsymbol{\theta}}_{s\text{-}1}^{(\text{GLM})} + \left(\mathbf{Z}'\boldsymbol{\Delta}_{s\text{-}1}\, \mathbf{V}^{-1}_{\mathbf{e}^{2\,(\text{EWLS})},s\text{-}1}\boldsymbol{\Delta}_{s\text{-}1}\, \mathbf{Z}\right)^{-1} \mathbf{Z}'\boldsymbol{\Delta}_{s\text{-}1}\, \mathbf{V}^{-1}_{\mathbf{e}^{2\,(\text{EWLS})},s\text{-}1}\left(\mathbf{e}^{2\,(\text{EWLS})}\text{-}\exp\left\{\mathbf{Z}\hat{\boldsymbol{\theta}}_{s\text{-}1}^{(\text{GLM})}\right\}\right) \quad (3.3)$$

where $\mathbf{Z}$ is the model matrix of variance regressors, $\mathbf{e}^{2\,(\text{EWLS})}$ is the $n \times 1$ vector of squared EWLS residuals from the means model, $\boldsymbol{\Delta}_{s\text{-}1} = diag\left(\delta_1, \delta_2, \cdots, \delta_n\right)$ with $\delta_i = \left[\partial\exp\left(\mathbf{z}_i'\hat{\boldsymbol{\theta}}\right)\big/\partial\left(\mathbf{z}_i'\hat{\boldsymbol{\theta}}\right)\right]_{\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}_{s\text{-}1}} = \exp\left(\mathbf{z}_i'\hat{\boldsymbol{\theta}}_{s-1}\right)$, and $\mathbf{V}_{\mathbf{e}^{2\,(\text{EWLS})},s\text{-}1}$ is the nxn diagonal matrix of variances of the squared EWLS residuals given by $diag\left(2\exp\{2\mathbf{z}_1'\hat{\boldsymbol{\theta}}_{s-1}\}, \ldots, 2\exp\{2\mathbf{z}_n'\hat{\boldsymbol{\theta}}_{s-1}\}\right)$. At convergence, the GLM variance estimate at the i^{th} observation is given by $\hat{\sigma}_i^{2\,(\text{GLM})} = \exp\left\{\mathbf{z}_i'\hat{\boldsymbol{\theta}}^{(\text{GLM})}\right\}$.

If the specified means model is insufficient, then the residuals from the means fit not only contain information regarding the process variance, but they are also contaminated with the lack-of-fit from the parametric means estimate. The result is a variance estimate which contains substantial bias [Robinson and Birch (2000)]. A nonparametric estimate of the mean also presents potential problems for a residual-based variance estimate. In many situations, the nonparametric estimate can fit the data too closely, thus leaving meager residuals with which to model the variance. This manifests itself in an underfit variance model.

The MMRR estimate of the mean is intended not only to offer an estimate of the mean which is robust to model misspecification but also an estimate which is less variable than a purely nonparametric fit. Thus, it seems natural to use the squared MMRR residuals as variance model data. The residuals from the MMRR estimate are given as

$$\mathbf{e}^{(\mathrm{MMRR})} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(\mathrm{EWLS})} - \lambda_{\mu}\, \mathbf{H}_{b_{\mu}}^{(\mathrm{LLR})} \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(\mathrm{EWLS})} \right) . \qquad (3.4)$$

Notice that if there is no misspecification of the means model then λ_{μ} should be zero and the MMRR residuals will simply be the parametric (EWLS) residuals. However, if the specified means model is insufficient, λ_{μ} should be closer to 1, and the MMRR fit will more resemble a purely nonparametric fit.

Like MMRR, the variance model robust regression (VMRR) procedure combines a parametric and nonparametric fit. We use an MRR1 fit for the robust variance estimate since there is nothing to guarantee a positive variance estimate when using an MRR2 fit. Recall that MRR1 involves combining parametric and nonparametric estimates of the data to form the model robust estimate. The VMRR procedure begins by regressing the squared MMRR residuals on an exponential function of the variance model regressors [$exp\{\mathbf{Z}\boldsymbol{\theta}\}$ in this discussion] to obtain a parametric (GLM) fit. We then obtain a purely nonparametric fit to the squared MMRR residuals

via LLR. The VMMR estimate is then a convex combination of the GLM and LLR fits via a variance model mixing parameter λ_σ. The *nx1* vector of VMRR variance estimates is given by

$$\hat{\boldsymbol{\sigma}}^{2\,(\mathrm{VMRR})} = \lambda_\sigma \mathbf{H}_{b_\sigma}^{(\mathrm{LLR})} \mathbf{e}^{2(\mathrm{MMRR})} + (1-\lambda_\sigma)\exp\left\{\mathbf{Z}\hat{\boldsymbol{\theta}}^{(\mathrm{GLM})}\right\} \tag{3.5}$$

where $\mathbf{H}_{b_\sigma}^{(\mathrm{LLR})} = \begin{bmatrix} \mathbf{h}_{1b_\sigma}^{(\mathrm{llr})\prime} \\ \vdots \\ \mathbf{h}_{nb_\sigma}^{(\mathrm{llr})\prime} \end{bmatrix}$ with $\mathbf{h}_{ib_\sigma}^{(\mathrm{llr})\prime} = \mathbf{z}_i^{(\mathrm{llr})\prime}\left(\mathbf{Z}^{(\mathrm{llr})\prime}\mathbf{W}^{(\mathrm{llr})}(\mathbf{z}_i)\mathbf{Z}^{(\mathrm{llr})}\right)^{-1}\mathbf{z}_i^{(\mathrm{llr})}\mathbf{W}^{(\mathrm{llr})}(\mathbf{z}_i)$, $\mathbf{Z}^{(\mathrm{llr})} = \begin{bmatrix} \mathbf{z}_1^{(\mathrm{llr})\prime} \\ \vdots \\ \mathbf{z}_n^{(\mathrm{llr})\prime} \end{bmatrix}$ is the model matrix for LLR, with $\mathbf{z}_i^{(\mathrm{llr})\prime} = (1 \;\; z_i)$, and $\mathbf{W}^{(\mathrm{llr})}(\mathbf{z}_i) = diag\left(\mathrm{h}_{i1}^{(\mathrm{ker})}, ..., \mathrm{h}_{in}^{(\mathrm{ker})}\right)$ where $\mathrm{h}_{ij}^{(\mathrm{ker})} = \dfrac{\mathrm{K}\left(\dfrac{z_i - z_j}{b_\sigma}\right)}{\sum_{j=1}^{n}\mathrm{K}\left(\dfrac{z_i - z_j}{b_\sigma}\right)}$. As was the case for the MMRR estimate, $K(.)$ denotes the simplified Gaussian kernel. Note that when estimating both the mean and variance nonparametrically, a different kernel function may be used for the mean than for the variance. The variance model mixing parameter is $\lambda_\sigma \in [0,1]$ and b_σ denotes a global bandwidth used to smooth the variance residuals from the GLM variance fit. Choice of λ_σ and b_σ is discussed in Section 5.

3.4 DMRR Algorithm

Since the mean and variance estimates are interdependent (means model estimates require estimated variances and the data for the variance model are the squared means model residuals), it seems logical that estimation of the dual model should take place within a single, iterative algorithm. DMRR combines the robust means fit (MMRR) with the robust variance fit (VMRR) in a single, iteratively reweighted least squares algorithm. The algorithm is given as follows:

1. Let $\hat{\mathrm{V}} = \mathrm{diag}\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_n^2\right)$ where $\hat{\sigma}_i^2 = 1$, i=1,…,n.

2. Using EWLS, obtain the parametric estimate of the means model : $\hat{y}_i^{(EWLS)} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(EWLS)} = \mathbf{x}_i' \left(\mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{X} \right)^{-1} \mathbf{X}' \hat{\mathbf{V}}^{-1} \mathbf{y}$, i=1,…,n. Let $\hat{\mathbf{y}}^{(EWLS)}$ denote the nx1 vector of EWLS fits.

3. Form the residuals from the fit found in Step 2 , $\mathbf{e}^{(EWLS)} = \left(\mathbf{y} - \hat{\mathbf{y}}^{(EWLS)} \right)$, and perform local linear regression on this set of residuals, obtaining $\hat{r}_{i_\mu} = \mathbf{h}_{i_{b_\mu}}^{(llr)'} \mathbf{e}^{(EWLS)}$, where $\mathbf{h}_{i_{b_\mu}}^{(llr)'}$ is the i^{th} row of $\mathbf{H}_{b_\mu}^{(LLR)}$ and $\mathbf{e}^{(EWLS)}$ is the *nx1* vector of EWLS residuals.

4. Obtain the MMRR fit to the means model, written as: $\hat{\mathbf{y}}^{(MMRR)} = \hat{\mathbf{y}}^{(EWLS)} + \lambda_\mu \hat{\mathbf{r}}_\mu$ where the i^{th} element of $\hat{\mathbf{r}}_\mu$ is $\hat{r}_{i_\mu}$ from Step 3.

5. Form the squared residuals from the MMRR fit to the mean, obtaining : $e_i^{2\,(MMRR)} = \left(y_i - \hat{y}_i^{(MMRR)} \right)^2$, i=1,…,n. Let $\mathbf{e}^{2(MMRR)}$ denote the nx1 vector of squared MMRR residuals.

6. The GLM model for estimating the variance is: $\phi\left(e_i^{2\,(MMRR)} \right) = \mathbf{z}_i' \boldsymbol{\theta}$ where $\phi(.)$ is the link function (assumed log link here). The fitted values are then given by: $\hat{\sigma}_i^{2(GLM)} = \exp\left\{ \mathbf{z}_i' \hat{\boldsymbol{\theta}}^{(GLM)} \right\}$, i=1,…,n

7. Perform local linear regression on the set of squared MMRR residuals, obtaining $\hat{\sigma}_i^{2(LLR)} = \mathbf{h}_{i_{b_\sigma}}^{(llr)'} \mathbf{e}^{2(MMRR)}$ where $\mathbf{h}_{i_{b_\sigma}}^{(llr)'}$ is the i^{th} row of $\mathbf{H}_{b_\sigma}^{(LLR)}$, i=1,…,n.

8. Obtain the VMRR estimates of variance which are written as: $\hat{\sigma}_i^{2\,(VMRR)} = \lambda_\sigma \hat{\sigma}_i^{2\,(LLR)} + (1 - \lambda_\sigma) \hat{\sigma}_i^{2\,(GLM)}$, i=1,…,n, and where $\lambda_\sigma \in [0,1]$ is the variance model mixing parameter.

9. Return to step 2 with $\hat{V} = diag\left(\hat{\sigma}_1^{2(VMRR)}, \cdots, \hat{\sigma}_n^{2(VMRR)} \right)$.

10. Cycle through steps 2 - 9 until convergence of the means model parameters.

The flexibility that DMRR affords to the researcher is due to the mixing parameters λ_μ and λ_σ. By varying the values of λ_μ and λ_σ, DMRR can successfully accommodate various states of the user's proposed parametric model ranging from both models (mean and variance) being correctly specified, to processes in which only one model is correctly specified, to situations where neither model has been correctly specified. Herein lies the contribution of DMRR. Current dual modeling techniques have been developed under the assumption that the researcher either has *complete* confidence (parametric approach) in the forms of one or both of the specified models or the researcher has *no* confidence (nonparametric approach) in an explicit specification of one or both of the models. This research contends that there are many cases in which the researcher's state of knowledge is not binary. Rather, the researcher is confident that a specified parametric function may be appropriate across part of the data but realizes that there may be certain characteristics in the data which cannot be captured parametrically. DMRR seeks to utilize as much of the researcher's parametric knowledge as possible while still allowing for specific deviations in the data to be captured. In the next section we detail the asymptotic characteristics of DMRR and then in Section 5 we present data-driven methods for bandwidth and mixing parameter selection.

4. Dual Model Asymptotic Results

Asymptotic results for MRR were presented in MBS. They derived the theoretically optimal mixing parameter, λ, and gave the rates of convergence for MRR1 and MRR2 estimators. Their results are extended here to the dual model. In particular, using the methods of MBS it is straightforward to show that the theoretically optimal mixing parameter for the means model (using MRR2), denoted by λ_μ^*, is

$$\lambda_{\mu}^{*} = \frac{\left\langle \hat{\mathbf{r}}_{\mu}, \boldsymbol{\psi} - \hat{\mathbf{y}}^{(\mathrm{EWLS})} \right\rangle}{\left\| \hat{\mathbf{r}}_{\mu} \right\|^{2}}$$

and the theoretically optimal mixing parameter for the variance model (using MRR1), denoted by λ_{σ}^{*}, is

$$\lambda_{\sigma}^{*} = \frac{\left\langle \hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})} - \hat{\boldsymbol{\sigma}}^{2(\mathrm{LLR})}, \hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})} - \boldsymbol{\psi} \right\rangle}{\left\| \hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})} - \hat{\boldsymbol{\sigma}}^{2(\mathrm{LLR})} \right\|^{2}}$$

where $\boldsymbol{\psi}$ is an *nx1* vector representing the true mean or variance function at each of the *n* observations, $\hat{\boldsymbol{\sigma}}^{2\,(\mathrm{GLM})}$ is the *nx1* vector of parametric fits to the squared MMRR residuals and $\hat{\boldsymbol{\sigma}}^{2\,(\mathrm{LLR})}$ is the *nx1* vector of LLR fits to the squared MMRR residuals. For notational convenience, if $\mathbf{h}_1$ and $\mathbf{h}_2$ are any two functions of $\mathbf{x}_i^{'}$, we define the inner product as

$$\left\langle \mathbf{h}_1, \mathbf{h}_2 \right\rangle = n^{-1} \sum_{i=1}^{n} \mathbf{h}_1\left(\mathbf{x}_i\right) \mathbf{h}_2\left(\mathbf{x}_i\right),$$

and the norm as

$$\left\| \mathbf{h}_1 \right\|^{2} = \left\langle \mathbf{h}_1, \mathbf{h}_1 \right\rangle, \text{ with } \left\| \mathbf{h}_1 \right\| = \left(\left\langle \mathbf{h}_1, \mathbf{h}_1 \right\rangle \right)^{0.5}.$$

Estimation of λ_{μ}^{*} and λ_{σ}^{*} is discussed in the next section.

Under the six assumptions (denoted A1 – A6) and four requirements (denoted by R1-R4) given in MBS and Starnes (1999), the following two theorems give asymptotic convergence rates for the means and variance estimates after one iteration of the DMRR algorithm. The proofs for these theorems as well as definitions of all terms (to include the nonparametric convergence rate), are given in the Appendix.

Theorem 1: If A1 – A6 hold, then

$$\left\|\hat{\lambda}_{\mu}^{*}\hat{\mathbf{r}}_{\mu}+\hat{\mathbf{y}}^{(\mathrm{EWLS})}-\boldsymbol{\psi}\right\|=\begin{cases}O_P\left(\gamma_n\right) & \text{if } \lim_{n\to\infty}\left(\delta_n\right)\neq 0\\ O_P\left(n^{-.5}\right) & \text{if } \lim_{n\to\infty}\left(\delta_n\right)=0\end{cases}$$

Theorem 2: If A1 – A6 and R1 –R4 hold, then

$$\left\|\hat{\lambda}_{\sigma}^{*}\hat{\boldsymbol{\sigma}}^{2(\mathrm{LLR})}+(1-\hat{\lambda}_{\sigma}^{*})\hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})}-\boldsymbol{\psi}\right\|=\begin{cases}O_P\left(\gamma_n\right) & \text{if } \lim_{n\to\infty}\left(\delta_n\right)\neq 0\\ O_P\left(n^{-.5}\right) & \text{if } \lim_{n\to\infty}\left(\delta_n\right)=0\end{cases}.$$

Thus, DMRR obtains the best asymptotic convergence rate available regardless of whether the user's parametric model is correct or incorrect. Clearly the convergence rate for the entire one iteration DMRR estimate will converge as quickly as the slowest of the two component estimates. Thus, DMRR, with the choice of mixing parameters given above, has the same asymptotic properties as MRR in the single means model case. Terms and proofs for these results are given in the appendix.

5. Data-Driven Bandwidth and Mixing Parameter Selection

For DMRR to be implemented in practice, it is important that there exist data-driven methods for bandwidth and mixing parameter selection. In this section, we outline a cross-validation method for bandwidth selection as well as explicit methods for choosing the mixing parameters for MMRR and VMRR. We begin with a discussion regarding bandwidth selection.

5.1 Bandwidth Selection

The smoothness of the nonparametric estimate is controlled by the bandwidth, *b*. When the bandwidth is too small, the nonparametric fit is too variable and when the bandwidth is too large, the nonparametric estimate suffers from bias. Hence, the choice of bandwidth involves a bias-variance trade-off. The literature is rich with bandwidth selection methods [see for example Härdle (1990), Härdle et al. (2004)]. Since the choice of bandwidth involves a bias-variance trade-

off, bandwidth selection is generally done so as to minimize an optimality criteria such as mean squared error (MSE). MBS introduce a penalized cross-validation technique, PRESS**, for choosing an appropriate bandwidth. The approach selects the bandwidth as the value of b which minimizes PRESS**, where PRESS** is defined as

$$\text{PRESS**} = \frac{\text{PRESS}}{n - trace\left(\mathbf{H}^{(\text{LLR})}\right) + \left[n - \left(k+1\right)\right]\dfrac{SSE_{max} - SSE_b}{SSE_{max}}}$$

where SSE_{max} denotes the largest error sum of squares over all possible bandwidths, SSE_b is the error sum of squares associated with a particular bandwidth, b; k is the number of regressors, and the prediction error sum of squares, PRESS, is given by:

$$\text{PRESS} = \sum_{i=1}^{n}\left(y_i - \hat{y}_{i,-i}\right)^2$$

where $\hat{y}_{i,-i}$ denotes the estimated mean response obtained by leaving out the i^{th} observation when estimating at location $\mathbf{x}_i$. MBS show that PRESS** performs well by guarding against very small and very large bandwidths.

Recall for DMRR that there are two sets of data, one for the means model and one for the variance model. Applying the expression for PRESS** to the model robust fit to the mean, the nonparametric fit is performed on the residuals from the EWLS parametric fit. Consequently, the 'y's' for the PRESS statistic are the EWLS residuals which are formed in Step 3 of the DMRR algorithm. The $\mathbf{H}^{(\text{LLR})}$ matrix in the PRESS** expression would then be $\mathbf{H}_{b_\mu}^{(\text{LLR})}$ from equation (3.4). Applying the expression for PRESS** to the robust fit to the variance, the nonparametric fit is performed on the squared MMRR residuals. Consequently, the 'y's' for the PRESS statistic are the squared MMRR residuals which are formed in Step 5 of the DMRR algorithm. The $\mathbf{H}^{(\text{LLR})}$ matrix

in PRESS** would then be $\mathbf{H}_{b_\sigma}^{(\mathrm{LLR})}$ from equation (3.5). The PRESS** statistic is used for all bandwidth selection of nonparametric fitting in the simulation study presented in the next section.

5.2 Mixing Parameter Selection

Sample estimators of λ_μ^*, the means model mixing parameter, and λ_σ^*, the variance model mixing parameter are given by

$$\frac{< \hat{\mathbf{r}}_\mu, \mathbf{y} - \hat{\mathbf{y}}^{(\mathrm{EWLS})} >}{\left\| \hat{\mathbf{r}}_\mu \right\|^2},$$

and

$$\frac{< \hat{\boldsymbol{\sigma}}^{2\,(\mathrm{GLM})} - \hat{\boldsymbol{\sigma}}^{2\,(\mathrm{LLR})}, \hat{\boldsymbol{\sigma}}^{2\,(\mathrm{GLM})} - \mathbf{e}^2 >}{\left\| \hat{\boldsymbol{\sigma}}^{2\,(\mathrm{GLM})} - \hat{\boldsymbol{\sigma}}^{2\,(\mathrm{LLR})} \right\|^2},$$

respectively. In the next section, we compare DMRR to some natural parametric and nonparametric competitors for estimating the dual model via a simulation study.

6. Simulation Study

In this section we present graphical and numerical comparisons of DMRR to traditional parametric and nonparametric dual modeling techniques in four scenarios: 1. The functional forms of both the underlying mean and variance models are correctly specified; 2. The functional form of the underlying variance function is correctly specified but the form of the mean is misspecified; 3. The functional form of the mean is correctly specified but the form of the variance model is misspecified; 4. Functional forms of both the mean and variance are misspecified. Comparisons among the methods will be based on the Monte Carlo simulated integrated mean squared error values for the mean and variance estimates (SIMSE(M) and SIMSE(V) respectively) given by

$$SIMSE(M) = \frac{\sum_{i=1}^{500} asem}{500}, \text{ with } asem = \frac{\sum_{j=1}^{1000} \left(E\left(y_j\right) - \hat{y}_j \right)^2}{1000}$$

and

$$SIMSE(V) = \frac{\sum_{i=1}^{500} asev}{500}, \text{ with } asev = \frac{\sum_{j=1}^{1000} \left(\sigma_j^2 - \hat{\sigma}_j^2 \right)^2}{1000}$$

where 'asem' and 'asev' denote the average squared error for the mean and variance fits, respectively, across 1000 locations (uniformly spaced) in the x-space for each of 500 simulated data sets. Also regarding notation, $E\left(y_j\right)$ and σ_j^2 are the true underlying mean and variance values at location x_j, respectively. The estimates of the mean and variance at location x_j, are $\hat{y}_j$ and $\hat{\sigma}_j^2$, respectively. The five competing methods are as follows: 1. DMRR, 2. the parametric approach of Aitken (denoted 'PAR'), 3. a nonparametric approach using LLR for estimating the mean and the pseudo-residual approach of Müller and Stadtmüller (1987 and 1993) (denoted 'NPAR1'), 4. a nonparametric approach using LLR for both the mean and variance estimates (denoted 'NPAR2'), and 5. the bias-corrected approach of Ruppert et al. (1997) (denoted as 'NPAR3'). Note that in all three nonparametric cases (NPAR1, NPAR2, and NPAR3) the mean estimates are identical as the techniques only differ in their estimation of the variance functions.

6.1 Simulation Scenarios 1 and 2

In this sub-section, we consider simulating data to explore comparisons among the five methods for scenarios 1 and 2 from above: 1. Both the mean and variance models are correctly specified by the user and 2. The form of the variance model is correctly specified but the mean is not. Data for this subsection as well as subsection 6.2 will be generated based on the following underlying dual model:

$$y_i = 2(x_i - 5)^2 + 5\,x_i + \gamma \sin\left(\frac{\pi\,(x_i - 1)}{2.25}\right) + g^{1/2}(z_i)\,\varepsilon_i, \qquad (6.1)$$

$$\sigma_i^2 = g(z_i) = \exp\{3.125 - 1.25x_i + 0.125x_i^2\}. \qquad (6.2)$$

We assume that x = z (the same variable influencing the mean also influences the variance). We also assume independent errors with the i^{th} error, $\varepsilon_i \sim N(0,1)$. Regarding notation γ denotes the misspecification parameter. Note that as the value of γ increases from 0 the adequacy of a specified quadratic model for the mean deteriorates. For the parametric approach, we assume that the user specifies a quadratic means model and an exponential variance model with both a linear and quadratic effect. This specification is also assumed for the parametric portion of the DMRR approach. The true underlying mean function for γ = 5.0 is plotted in Figure 2(a). Three sample sizes are considered (n = 20, 40, and 60) and five values of γ, ranging from the mean model correctly specified when γ =0, to varying degrees of misspecification, γ =2.5, 5.0, 7.5, and 10.0). In all examples, the x-values are taken at evenly spaced locations from 0 to 10, and are scaled to be between 0 and 1 when choosing the bandwidths and mixing parameters.

The simulated integrated mean squared error values for the estimated mean and variance functions, SIMSE(M) and SIMSE(V), respectively, are provided in Table 1 with minimum values across the five methods indicated in bold. Numbers in the tops of each cell denote SIMSE(M) whereas numbers in the bottom of each cell denote SIMSE(V). Note that when $\gamma = 0$ both the mean and variance functions are correctly specified in the parametric (PAR) approach. As a result, PAR performs uniformly best (mean and variance estimates) while $\gamma = 0$ across all sample sizes. As γ increases in magnitude from 0, one would expect that the performance of PAR would deteriorate for both the mean and variance estimates. The mean estimate for PAR deteriorates because the quadratic parametric model becomes less adequate and the variance estimate of PAR deteriorates since the data used for the variance model involves the residuals from the estimated mean. The

impact of means model misspecification can be seen for a particular data set (*n=40*) in Figures 2(b) and 2(c) for $\gamma = 5.0$. Note that the parametric estimate of the mean in 2(b) misses the bumps caused by the addition of the sine function. DMRR and the LLR estimates of the mean, on the other hand, capture the true structure of the underlying function. Although the variance model is correctly specified by PAR, when the residuals are contaminated with lack-of-fit, the specified variance function is not only trying to fit the variance structure but also the lack-of-fit from the mean. The failure of PAR for $\gamma = 5.0$ is visualized in Figure 2(c) as the PAR estimate of variance is seen to be extremely biased. Comparing Figure 2(d) to 2(c) we see that the DMRR and NPAR methods do reasonably well at capturing the structure of the underlying variance function for this particular data set. Note for γ >5.0, the PAR approach to dual modeling is uniformly worse than all other methods. The only exception is the difference-based variance estimate, NPAR1 when γ =5.0 and n=20. The difference-based variance estimate (NPAR1) performs better with larger sample sizes but performs least effectively among the nonparametric methods considered here.

As the parametric mean becomes more misspecified, γ >0, the proposed dual model robust methodology (DMRR) performs uniformly more superior than any of its competitors. Among the purely nonparametric estimates, the local linear smooth of the mean and local linear smooth of the squared mean residuals (NPAR2), performs uniformly best. The bias adjusted method of Ruppert et. al. (1997) (NPAR3) becomes more competitive with NPAR2 as the sample size increases. Note that the most pronounced difference between NPAR2 and DMRR occurs when there is small ($\gamma = 2.5$) to moderate ($\gamma = 5.0$) degrees of misspecification in the mean. When $\gamma = 2.5$, the DMRR SIMSE(M) is 14.9% smaller than NPAR2 for *n*=20, 13.9% smaller than NPAR2 for *n=40*, and 11.6% smaller than NPAR2 for *n=60*. Regarding the variance estimate, the DMRR SIMSE(V) is 4.8%, 9.4%, and 16% smaller than NPAR2 for *n* = 20, 40, and 60, respectively. The improved performance of the DMRR variance estimate for larger sample sizes is likely do to the improved

performance of the parametric variance fit with larger sample sizes (note the performance of the PAR variance estimate across all values of γ as n increases).

The improvement that DMRR offers over the purely parametric approach increases as γ increases in magnitude. The improvement that DMRR offers over its purely nonparametric competitors is greatest with small and moderate degrees of mean misspecification. The greater the misspecification in the mean, the closer the DMRR estimate is to the NPAR2 estimate. Recall that the philosophy which drives the DMRR approach is that of explaining as much of the underlying structure as possible parametrically and then augmenting with a nonparametric smooth to capture 'anomalies'. Consequently, as the user-supplied parametric function becomes less and less adequate, the DMRR fit becomes almost entirely comprised of the nonparametric estimate, thus reducing the degree of difference between DMRR and NPAR2. In the next subsection, we consider misspecification in both the mean and variance.

6.2 Simulation Scenarios 3 and 4

In this sub-section, we consider simulating data to explore comparisons among the five methods for scenarios 3 and 4 above: 3. The functional form of the mean is correctly specified but the form of the variance model is misspecified. 4. Functional forms of both the mean and variance are misspecified. Data will be generated based upon the same functional forms as given in (6.1) and (6.2). We again assume that x = z (the same variable influencing the mean also influences the variance). We also assume independent errors with the i^{th} error, $\varepsilon_i \sim \mathrm{N}(0,1)$. For the parametric approach, we still assume a quadratic means model is specified but now we assume that an exponential variance model is specified with only a linear main effect. Thus, the linear predictor in the variance model is misspecfied. This specification is also assumed for the parametric portion of the DMRR approach. Similar to the previous section, the mean becomes more misspecified as γ increases from zero. Three sample sizes are considered (n = 20, 40, and 60) and five degrees of

means misspecification are studied (γ = 0 (correct specification), 2.5, 5.0, 7.5, and 10.0). Simulated integrated mean squared error values for the mean and variance estimates, SIMSE(M) and SIMSE(V), are provided in Table 2.

When there is no misspecification in the mean, PAR offers the best estimate of the mean in terms of SIMSE(M). Although PAR is superior to the other methods for estimating the mean when $\gamma = 0$, its SIMSE(M) values are larger than they were when the variance model was correctly specified (Table 1). As was observed in Table 1, the PAR estimates of both the mean and variance deteriorate as γ increases in magnitude from zero. The SIMSE(V) values for PAR are largely greater in Table 1 (correctly specified variance) than in Table 2 (incorrectly specified variance). Although this may appear counter-intuitive, it is important to remember that for $\gamma > 0$, the specified variance model is not only attempting to fit the variance structure but also the lack-of-fit from the poorly estimated mean. Consequently, the functional form of the variance plus lack-of-fit may be better described by a linear predictor involving x than a linear predictor involving x^2.

As was observed in Table 1, the DMRR estimate of the mean is superior to all other methods in terms of SIMSE(M) when $\gamma > 0$. Comparisons of the estimated variance models are slightly more complicated for the scenarios considered in Table 2. For nearly all situations in Table 2, there is negligible difference between the local linear smooth (NPAR2) and DMRR in terms of SIMSE(V). The reason for this relates again to the philosophy of the DMRR estimate. The DMRR estimate's performance is based upon the user supplying a parametric form that explains a reasonable amount of structure in the underlying functional form. For the results in Table 2, the user has supplied a linear predictor that is linear in x when the actual variance function involves a linear predictor that is quadratic in x. As a result, the parametric specification is poor. When the user-supplied parametric form is poor, the DMRR and local linear smooth (NPAR2) will tend to perform equally as well. This is evident in Table 2 as NPAR2 and DMRR vary only by an amount

that could be ascribed to simulation error. It is important to note, however, that in practice one never knows the extent to which a parametric model explains the underlying structural form. In situations where the specified parametric function is highly inadequate, the DMRR approach, while computationally more intensive, is generally no worse than the local linear smooth. In situations in which a parametric model can explain most of the underlying structure except for certain anomalies, the DMRR approach will tend to perform better than both the parametric approach and the local linear smooth.

7. Summary and Conclusions

We have developed a model robust approach to dual modeling. Dual modeling involves the simultaneous modeling of both the mean and variance. Most methods of variance estimation (except for the difference-based approach of Müller and Stadtmüller [1987 and 1993] when there is no replication available require residuals from the mean fit in order to fit the variance function. Consequently, it is imperative for the user to fit the mean well in order to increase the likelihood of accurately estimating the variance. Traditionally nonparametric approaches for estimating the mean and variance simultaneously have been considered when the user cannot adequately describe the forms of the mean and variance functions with a parametric model. The proposed dual model robust methodology (DMRR) offers a 'hybrid' approach to dual modeling by combining a parametric fit with a portion of a local polynomial smooth via appropriately chosen mixing parameters. Computational expressions are provided for the choice of mixing parameters. The DMRR approach offers particular advantages over its nonparametric counterparts when the user can describe much of the underlying functional forms via parametric models. The DMRR approach is more beneficial than its parametric counterpart when there are certain anomalies in the data which

cannot be captured parametrically. In practice one never knows the true degree of model misspecification and as a result, we recommend the DMRR approach. In this manuscript we have established asymptotic superiority of DMRR over local polynomial regression and in a small simulation study, we have demonstrated the advantages that DMRR can offer over its parametric and nonparametric counterparts in small sample settings.

REFERENCES

Aitkin, M., Modelling variance heterogeneity in Normal regression using GLIM, *Appl. Statist.* **36**, (1987), 332-339.

Burman, P. and Chaudhuri, P., A Hybrid Approach to Parametric and Nonparametric Regression, *Technical Report No. 243*, Division of Statistics, University of California-Davis,Davis, CA, USA, (1992).

Carroll, R.J., Adapting for Heteroscedasticity in Linear Models, *The Annals of Statistics*, **10**, (1982), 1224-1233.

Carroll, R. J.., and Ruppert, D., *Monographs on Statistics and Applied Probability 33, Transformation and Weighting in Regression*. London: Chapman and Hall, (1988).

Einsporn, R.L. and Birch, J.B., Improving Predictions by Combining Least Squares and Kernel Regression Prediction Weights. *Virginia Tech, Department of Statistics, Technical Report, #93-11*, (1993) 21 pages.

Fan, J. and Ullah, A., Asymptotic Normality of a Combined Regression Estimator, *Journal of Multivariate Analysis*, **71**, (1999), 191-240.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C., Residual Variance and Residual Pattern in Nonlinear Regression, *Biometrika* **73**, (1986), 625-633.

Hall, P. and Carroll, R.J. , Variance Function Estimation in Regression: the Effect of Estimating the Mean, *J.R. Statist. Soc. B* **51**, (1989), 3-14.

Härdle , W., *Applied Nonparametric Regression,* Chapman and Hall, (1990).

Härdle , W., Muller, M., Sperlich, S., and Werwatz A., *Nonparametric and Semiparametric Models*. Springer Verlag, (2004).

Holst, U., Hössjer, O., Björklund, C., Ragnarson, P., and Edner, H., Locally Weighted Least Squares Kernel Regression and Statistical Evaluation of LIDAR Measurements, *Environmetrics*, **7**, (1996), 401-416.

Levine, M. Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: A possible approach. *Computational Statistics & Data Analysis*. 50 (2006), 3405-3431

Mays, J., Birch, J. B., and Starnes, B. A., Model Robust Regression: Combining Parametric, Nonparametric and Semiparametric Methods, *Journal of Nonparametric Statistics*, **13**, (2001), 245-277.

Mays, J. E. , Birch, J. B. and R. L. Einsporn., An overview of Model Robust Regression,

Journal of Statistical Computation and Simulation. Vol. 66, (2000) pp. 79-100.

McCullagh, P., and Nelder, J.A., *Generalized Linear Models*, Second Edition, Chapman and Hall, (1994).

Müller, H. and Stadtmüller, U., Variable Bandwidth Kernel Estimators of Regression Curves, *Annals of Statistics*, **15**, (1987), 182-201.

Müller, H. and Stadtmüller, U., On variance function estimation with quadratic forms, *Journal of Statistical Planning and Inference* **35**, (1993), 213-231.

Rahman, M., Gokhale, D. V. and Ullah, A., A Note on Combining Parametric and Nonparametric Regression, *Communications in Statistics-Simulation and Computation*, **26**, (1997), 519-529.

Robinson T.J. and Birch J.B., Model Misspecification in Parametric Dual Modeling", *Journal of Statistical Computation and Simulation*, Vol. **66**, (2000), 113-126.

Ruppert, D., Wand, M.P., Holst, U. and Hössjer , O., Local Polynomial Variance-Function Estimation, *Technometrics*, **39**, (1997), 262-273.

Ruppert, D., M.R. Wand, and R. J. Carroll. Semiparametric Regression, Cambridge University Press, (2003).

Sigrist, M. (ed.), *Air Monitoring by Spectroscopic Techniques* (Chemical Analysis Series, Vol. **127**), New York: Wiley, (1994).

Simonoff, J. S. *Smoothing Methods in Statistics*. Springer Verlag (1996).

Speckman, P., Kernel Smoothing in Partial Linear Models, *Journal of the Royal Statistical Society B*, **50**, (1988), 413-436.

Starnes, B.A., "Asymptotic Results for Model Robust Regression", Unpublished dissertation, Virginia Polytechnic Institute and State University, Blacksburg, VA., (1999).

Wackerly, D. D., R. L. Scheaffer and W. Mendenhall III. Mathematical Statistics with Applications, Duxbury Press, (2002).

n=20 γ	PAR	NPAR1	NPAR2	NPAR3	DMRR
0	**0.765** ***28.376***	2.788 *304.812*	2.788 *33.161*	2.788 *41.829*	1.563 *29.362*
2.5	3.908 *37.538*	2.974 *369.447*	2.974 *33.689*	2.974 *42.283*	**2.529** ***32.039***
5.0	14.726 *305.975*	3.397 *542.214*	3.397 *36.181*	3.397 *51.674*	**3.069** ***33.667***
7.5	26.450 *928.407*	3.648 *680.427*	3.648 *35.704*	3.648 *54.310*	**3.446** ***35.272***
10	45.813 *2615.06*	3.791 *959.97*	3.791 *34.430*	3.791 *53.776*	**3.600** ***33.951***
n=40 γ					
0	**0.371** ***14.469***	1.520 *69.455*	1.520 *23.648*	1.520 *30.205*	0.916 *20.356*
2.5	3.344 *25.701*	1.712 *74.092*	1.712 *23.600*	1.712 *30.186*	**1.474** ***21.371***
5.0	13.915 *243.133*	1.925 *86.288*	1.925 *24.624*	1.925 *33.210*	**1.783** ***21.924***
7.5	25.362 *763.418*	2.101 *104.870*	2.101 *24.833*	2.101 *36.103*	**1.986** ***21.871***
10	44.694 *2330.40*	2.235 *111.148*	2.235 *25.733*	2.235 *35.600*	**2.184** ***22.802***
n=60 γ					
0	**0.239** ***10.882***	1.127 *60.433*	1.127 *19.930*	1.127 *27.913*	0.676 *16.304*
2.5	3.153 *21.760*	1.227 *57.530*	1.227 *19.142*	1.227 *25.873*	**1.085** ***16.068***
5.0	13.661 *229.518*	1.359 *50.830*	1.359 *19.357*	1.359 *25.810*	**1.284** ***16.944***
7.5	25.088 *735.464*	1.440 *63.580*	1.440 *20.844*	1.440 *27.655*	**1.392** ***17.391***
10	44.427 *2269.2*	1.612 *69.041*	1.612 *21.110*	1.612 *28.701*	**1.570** ***18.040***

Table 1: Mean model misspecified, variance model correctly specified. Simulated integrated mean squared error values for the five methods with SIMSE(M) on top, SIMSE(V) on bottom. Best values over the five methods in bold for each γ.

n=20 γ	PAR	NPAR1	NPAR2	NPAR3	DMRR
0	**1.029** ***30.177***	2.653 *280.804*	2.653 *33.801*	2.653 *43.407*	1.480 *31.553*
2.5	3.945 *38.392*	2.977 *345.403*	2.977 *34.347*	2.977 *45.639*	**2.525** ***33.538***
5.0	14.512 *233.914*	3.319 *518.584*	3.319 *33.795*	3.319 *49.730*	**3.060** ***33.711***
7.5	25.991 *705.696*	3.621 *616.949*	3.621 *35.713*	3.621 *54.640*	**3.361** ***33.843***
10	45.192 *2116.77*	3.573 *874.973*	3.573 *35.667*	3.573 *56.956*	**3.360** ***34.287***
n=40 γ					
0	**0.538** *27.706*	1.539 *73.868*	1.539 *22.830*	1.539 *29.867*	0.857 ***21.466***
2.5	3.302 *37.145*	1.717 *81.052*	1.717 *24.245*	1.717 *31.830*	**1.500** ***23.999***
5.0	13.857 *213.890*	1.944 *84.787*	1.944 ***24.771***	1.944 31.676	**1.833** 24.821
7.5	25.242 *673.375*	2.107 *92.200*	2.107 ***25.330***	2.107 *34.605*	**1.993** *25.371*
10	44.404 *2110.34*	2.217 *117.615*	2.217 ***25.508***	2.217 *36.259*	**2.138** *26.173*
n=60 γ					
0	**0.406** *27.036*	1.170 *63.793*	1.170 *19.691*	1.170 *27.631*	0.726 ***18.570***
2.5	3.093 *36.291*	1.205 *66.988*	1.205 *20.556*	1.205 *29.875*	**1.029** ***19.225***
5.0	13.633 *211.858*	1.424 *60.667*	1.424 *21.322*	1.424 *30.355*	**1.336** ***20.882***
7.5	24.972 *673.468*	1.445 *63.473*	1.445 *20.050*	1.445 *27.498*	**1.386** ***19.980***
10	44.166 *2037.90*	1.607 *64.682*	1.607 *20.964*	1.607 *28.451*	**1.572** ***20.912***

Table 2: Mean and Variance Models Misspecified. Simulated integrated mean squared error values for the five methods with SIMSE(M) on top, SIMSE(V) on bottom. Best values over the five methods in bold for each γ

[*Robinson, Birch, and Starnes*]
Figure 1: 1(a) Lidar Data, 1(b) MMRR Fit,
1(c) VMRR, Parametric, and Nonparametric Fits to Squared MMRR Residuals

Figure 1(a)

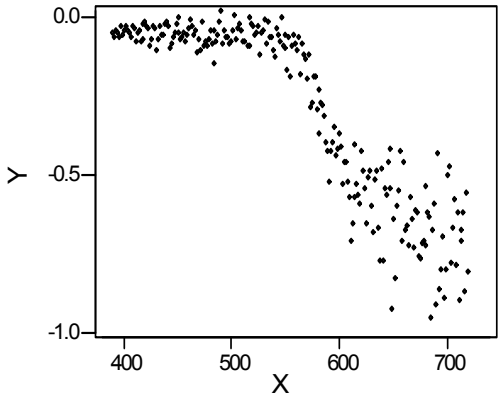
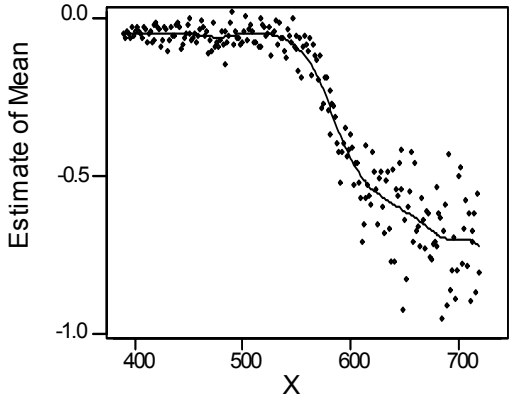
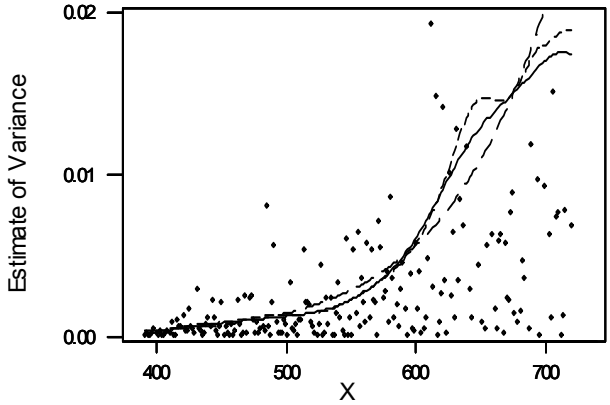


Figure 1(b)



MMRR (Solid)

Figure 1(c)



VMMR (Solid), Parametric (Long Dash), Nonparametric (Short Dash)

[*Robinson, Birch, and Starnes*]
Figure 2 : Example on p. 16
2(a) True Mean Function, 2(b) MMRR, EWLS, and LLR Fits,
2(c) True Variance Function and GLM Estimate of Variance,
2(d) VMRR, Nonparametric Residual Based, and Difference Based Variance Estimates

Figure 2(a)

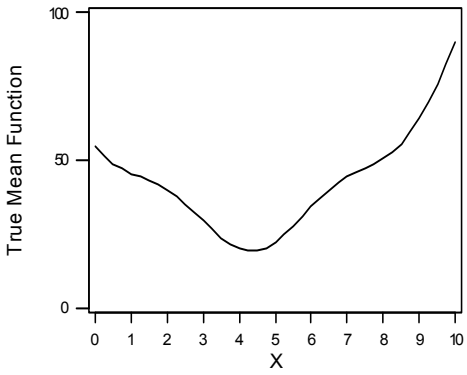
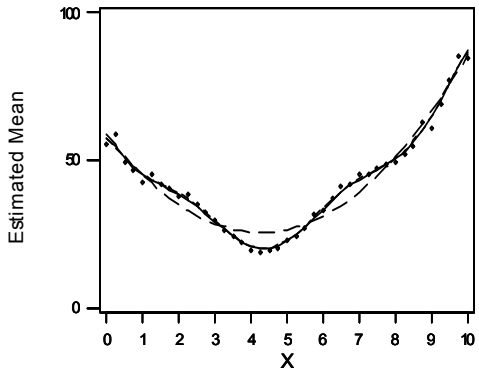
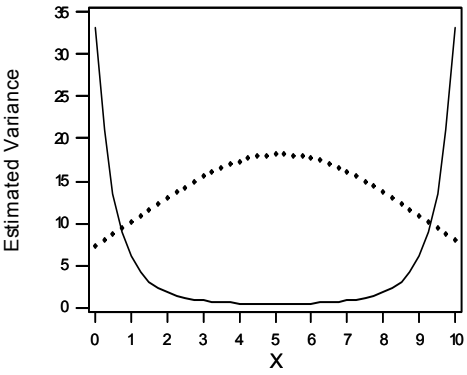


Figure 2(b)



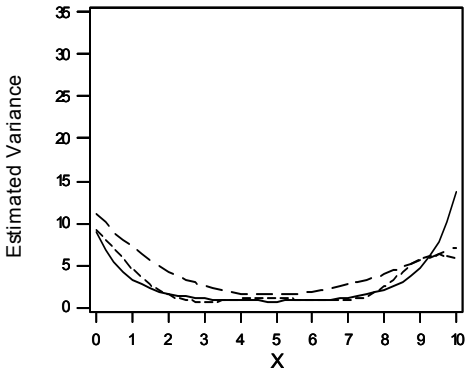
MMRR (Solid), EWLS (Dash), LLR Fit (Dotted)

Figure 2(c)



True Variance Function (Solid), GLM Estimated Variance (Dotted)

Figure 2(d)



VMRR (Solid), Diff (Long Dash), Res (Short Dash)

Appendix: Proofs of Theorems 1 and 2

We prove the asymptotic results in this appendix. We will begin with some conventions. The term "asymptotic" means that the number of observations increases without bound, and the manifestation

of these observations in the X space increases uniformly in C, the predictor space. Note that the predictor space is one dimensional. Consequently, here, and in requirements **R1, R3** and **R4**, and assumption **A1**, we are concerned only with the x_i (and z_i) terms of $\mathbf{x}_i$ (and $\mathbf{z}_i$) on p. 8. Thus, we have a fixed effects structure, as $n \to \infty$, over a uniform design. To that end, we need to introduce some convenient and useful notation. We will use $\mathscr{L}$ to mean $\lim_{n\to\infty}$ (called the "limfinity" symbol).

For the purpose of this section we will consider the first iteration *only* of any part of the DMRR procedure being performed at each *n*. This is because an iterative procedure does not involve any additional observations and the machination for our results is the increasing uniform density of observations, *not* the refinement that takes place in an iterative process with the same *n* observations. If the iterative process produces a better estimate after a number of iterations, then the overall estimate can do no worse than that of the first iteration.

We shall momentarily label either *nx1* vector of parametric fits ($\hat{\mathbf{y}}^{(EWLS)}$ or $\hat{\boldsymbol{\sigma}}^{2(GLM)}$) by $\hat{\boldsymbol{\psi}}_p$, and either *nx1* vector of nonparametric fits ($\hat{\mathbf{r}}_\mu$ or $\hat{\boldsymbol{\sigma}}^{2(LLR)}$) by $\hat{\boldsymbol{\psi}}_{np}$, where $\boldsymbol{\psi}$, the unknown true regression function vector, is defined on p. 15. We will denote by δ_n, the distance between $\boldsymbol{\psi}$, and the parametric family of continuous (on C) regression models under consideration evaluated at the *n* observations, where

$$\delta_n = \inf\{\|\boldsymbol{\psi} - \boldsymbol{\psi}_\mathbf{p}(\boldsymbol{\beta})\|: \ \boldsymbol{\beta} \in R^{k+1}\}. \qquad (A.1)$$

If the infimum is attained at a particular $\boldsymbol{\beta}$, we will designate this as $\boldsymbol{\beta}^*$ and write

$$\delta_n = \| \boldsymbol{\psi} - \boldsymbol{\psi}_\mathbf{p}(\boldsymbol{\beta}^*)\|.$$

Similarly, we set γ_n as a distance measure for the nonparametric estimate,

$$\gamma_n^2 = E(\| \hat{\boldsymbol{\psi}}_{np} - \boldsymbol{\psi}\|^2) \text{ or } E(\| \hat{\boldsymbol{\psi}}_{np} - (\boldsymbol{\psi} - \boldsymbol{\psi}_\mathbf{p}(\boldsymbol{\beta}^*)\|^2)$$

according to whether we're using MRR1 or MRR2 respectively. Note that the distance is taken using the optimal parametric fit. Even though the nonparametric estimate is dependent upon the convergence of the parametric estimate to its optimal form (given above), the nonparametric rate of convergence is still based upon the sample size. Thus, even if the nonparametric estimate has a different target function for each sample of size *n*, it will reach that goal with a degree of accuracy based on the number of observations. We subsequently show that for the DMRR estimates, the rate of convergence for the parametric estimates will exceed that of the nonparametric estimates. As in

MBS, the above notation is appropriate for the MRR type estimates involved here. In Starnes (1999) it is seen that the nonparametric rate of convergence is dependent on the bandwidth.

Since the nonparametric estimates meet the criteria given in MBS, Burman and Chaudhuri (1992), and Starnes (1999) for theorems 1 and 2, and the first iteration means estimate uses the usual parametric linear regression, we want to show that the parametric rate of convergence holds for the GLM estimate in the variance estimate. The parametric rate of convergence is important because in the variance estimate we utilize a more general type of parametric modeling (nonlinear function). In the results that follow, we will show that the generalized parametric estimate will not require any specific form for the mean response as a function of the independent variable (as in GLM), except that it is a continuous function of both $\mathbf{x}$ and $\boldsymbol{\beta}$.

For the moment we will once again label the generalized parametric estimate by $\hat{\boldsymbol{\psi}}_{\mathrm{p}}$, with $\boldsymbol{\psi}_{\mathbf{p}}$ defined as some function of the parameter vector $\boldsymbol{\beta}$ as in subsequent equation A.1. Notice that we do not require $\boldsymbol{\psi}_{\mathbf{p}}$ to be linear in $\boldsymbol{\beta}$, so that we are now operating with the generalized non-linear model, which contains the family of generalized linear models. Additionally, we will make the following four requirements for the parametric estimate.

R1. The observed responses are independent (but not necessarily identically distributed), and are functions of x_i, values that are fixed uniformly on C = (a,b) with $0 < a < b < \infty$.

R2. $V(\mathbf{x}_i)$ (the variance at each $\mathbf{x}_i$) is bounded both from above and below.

R3. For every $n > p+1$, the matrix D is of full rank $(p+1)$ where D is a matrix of partial derivatives with elements

$$d_{i,j} = \frac{\partial \boldsymbol{\psi}_{\mathrm{p}}(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \beta_j}, \text{ where } i = 1,\ldots,n, \text{ and } j = 0,\ldots,p.$$

Note, that this implies that there is no β_j (for $j = 0, \ldots, p$) such that

$$\frac{\partial \boldsymbol{\psi}_{\mathrm{p}}(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j} \equiv 0, \text{ for } x \text{ in C}.$$

R4. For $j = 0, \ldots, p$, $\dfrac{\partial \boldsymbol{\psi}_{\mathrm{p}}(\mathbf{x}, \boldsymbol{\beta})}{\partial \beta_j}$ is continuous and bounded for both x and $\boldsymbol{\beta}$ on C = (a,b) and R^{p+1} respectively.

Define

$$\boldsymbol{S}_{Gn} = n^{-1} \sum_{i=1}^{n} \left(w_i \frac{\partial \boldsymbol{\psi}_{\mathrm{p}}(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left(\frac{\partial \boldsymbol{\psi}_{\mathrm{p}}(\mathbf{x}_i; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right)^{\mathrm{T}} \right)$$

and

$$\boldsymbol{S}_G = \mathscr{L} \boldsymbol{S}_{Gn},$$

where w_i is a weighting constant related to $V(\mathbf{x}_i)$. We are interested in whether or not this matrix exists and is defined and finite (i.e. both $\det(\boldsymbol{S}_G^{-1})$ and $\det(\boldsymbol{S}_G)$ are nonzero and finite). The following lemma gives the convergence rate for almost any type of non-linear parametric regression estimate formed by using an IRLS iteration scheme [see Carroll and Ruppert (1988)]. Its proof may be found in Starnes (1999).

Lemma 1 Assuming requirements R1-R4 above S_{Gn} has a finite limit asymptotically; that is, S_G^{-1} is defined and finite, where

$$\boldsymbol{S}_G^{-1} = (\mathscr{L} \boldsymbol{S}_{Gn})^{-1}.$$

Carroll and Ruppert (1988) and Bishop, Feinberg and Holland (1975) provide the machinery for Lemma 2.

Lemma 2 Assuming requirements R1-R4 above, the parametric regression estimate, $\hat{\boldsymbol{\psi}}_{\mathrm{p}}$, obtained through IRLS (with a n^{-5} consistent starting estimate for $\boldsymbol{\beta}$), has the property

$$\left\| \hat{\boldsymbol{\psi}}_{\mathrm{p}} - \boldsymbol{\psi}_{\mathrm{p}} \right\|^2 = O_P(n^{-1}).$$

Now that we have established a rate of convergence for a global family of parametric estimates (which include the GLM parametric estimates used in DMRR), we must recall that the parametric estimate in question (from step 6, p. 13) was utilized in an MRR1 framework. We have complementary portions of the mixing parameter used in both nonparametric and parametric parts. In this case, it is not only necessary to show that the parametric convergence rate is appropriate, but

also that the parametric estimate, $\hat{\sigma}^{2\,(\mathrm{GLM})}$, meets the standards of A1 of MBS, the first of the aforementioned 6 assumptions. This assumption is used for the parametric part of any of the MRR estimates and is given as follows.

A1. There exists a function W_1 of two variables which is defined and bounded on C × C, where C is the predictor space, and

$$\left\| \hat{\boldsymbol{\psi}}_{\mathrm{p}}(\hat{\boldsymbol{\beta}},.) - \boldsymbol{\psi}_{\mathrm{p}}(\boldsymbol{\beta}^*,.) - n^{-1}\sum_{i=1}^{n} W_1(.,x_i)\varepsilon_i \right\| = O_P(n^{-1}) \qquad \text{(A.2)}$$

Burman and Chaudhuri (1992) state that this condition is trivially satisfied (the norm is equal to zero) if $\boldsymbol{\psi}$ is linear in $\boldsymbol{\beta}$ and continuous in the second argument. Notice that in steps two and six of the DMRR algorithm (p. 13) both parametric estimates are continuous in all components **x** and **z** respectively. Also notice that for the means model, the parametric estimate is linear in $\boldsymbol{\beta}$, but that the variance parametric estimate is a function of a linear predictor of $\boldsymbol{\beta}$.

According to Starnes (1999) (pp. 96 – 98), the GLM variance parametric estimate, $\hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})}$, satisfies assumption A1 if we assume that the variance is at least bounded. In this instance, the squared residuals resulting from the means model (step 4, p. 13), have an approximate Gamma distribution, which has a finite variance ($\delta\varsigma^2$, where $\delta, \varsigma \in \mathrm{R}^+$) according to Wackerly, Scheaffer and Mendenhall (2002). The essence of the verification is that we utilize a Taylor Series expansion on the function of the linear predictor to form an appropriate weight function, W_1 to satisfy A.2. The result is the same functional rate of convergence in either parametric case (traditional or GLM). So in both the means and variance estimates, the parametric component satisfies assumption A1.

Finally, then, the overall convergence rates for the one iteration mean and variance estimates necessary for DMRR follow the MRR results given in MBS. So that $\left\| \hat{\lambda}_{\mu}^{*}\hat{\mathbf{r}}_{\mu} + \hat{\mathbf{y}}^{(\mathrm{EWLS})} - \boldsymbol{\psi} \right\|$ …

$$= \mathrm{O}_P(\gamma_n), \text{ if } \mathscr{L}\delta_n \neq 0, \text{ and}$$
$$= \mathrm{O}_P(n^{-.5}), \text{ if } \mathscr{L}\delta_n = 0;$$

and $\left\| \hat{\lambda}_{\sigma}^{*}\hat{\boldsymbol{\sigma}}^{2(\mathrm{LLR})} + (1-\hat{\lambda}_{\sigma}^{*})\hat{\boldsymbol{\sigma}}^{2(\mathrm{GLM})} - \boldsymbol{\psi} \right\|$ …

$$= \mathrm{O}_P(\gamma_n), \text{ if } \mathscr{L}\delta_n \neq 0, \text{ and}$$
$$= \mathrm{O}_P(n^{-.5}), \text{ if } \mathscr{L}\delta_n = 0.$$

The proofs of these results may be found in MBS. //.