A Graphical Analysis of Simultaneously Choosing the Bandwidth and Mixing Parameter for Semiparametric Regression Techniques

Derick L. Rivers
Virginia Commonwealth University

Follow this and additional works at: https://scholarscompass.vcu.edu/etd
Part of the Physical Sciences and Mathematics Commons

Downloaded from
https://scholarscompass.vcu.edu/etd/1896

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.
A GRAPHICAL ANALYSIS OF SIMULTANEOUSLY CHOOSING THE BANDWIDTH AND THE MIXING PARAMETER FOR SEMIPARAMETRIC TECHNIQUES

A Thesis submitted in partial fulfillment of the requirements for the degree of Masters in Mathematical Sciences Concentration in Statistics at Virginia Commonwealth University.

by

DERICK L. RIVERS

B.S. in Mathematics, Morris College, 1995

M.Ed. in Mathematics Education, Cambridge College, 2004

Director: James E. Mays, Associate Professor

Department of Statistical Sciences and Operations Research

Virginia Commonwealth University

Richmond, Virginia

August 2009

Keywords: Bandwidth Selection, Data-driven Selectors, Mixing Parameter Selection, Semiparametric Regression, 3D Plots
Acknowledgements

I would like to express my deepest appreciation to my thesis advisor, Dr. James Mays, whose encouragement, guidance and support from the initial stages to the final level of completion enabled me to develop an understanding of the subject. Without his guidance and persistent help this thesis would not have been possible.

I would also like to offer my regards and blessings to my family, friends, and co-workers who supported me in my quest to achieve this goal. Your words of encouragement and prayers along the way were vital as I endeavored to complete this task.
Contents

1 Introduction .................................................. 1

1.1 Ordinary Least Squares .................................... 4

1.2 Nonparametric Regression .................................. 8

1.2.1 Kernel Regression ........................................ 9

1.2.2 Bandwidth Choice ........................................ 12

1.2.2.1 Cross Validation ..................................... 14

1.2.2.2 Penalizing Functions ................................ 14

1.2.2.3 Plug-In Method ....................................... 15

1.2.2.4 Bootstrapping ......................................... 16

1.2.2.5 PRESS* ............................................... 17

1.2.2.6 PRESS** ............................................. 17

1.2.3 Local Polynomial Regression ............................ 18

1.2.4 Spline Smoothing ......................................... 22

1.3 Semiparametric Regression .................................. 24

1.3.1 Partial Linear Regression (PLR) ......................... 26

1.3.2 Model Robust Regression 1 (MRR1) ...................... 28
2 Previous Research

2.1 MSE Criterion .................................................. 35

2.2 Previous Results Comparing OLS, LLR, PLR, MRR1, and MRR2 ........ 36

2.3 Previous Results Comparing PRESS* and PRESS** ....................... 38

2.4 Previous Comparisons of Data sets ..................................... 39

2.4.1 Examples ...................................................... 39

2.4.1.1 Example 1 .................................................. 40

2.4.1.2 Example 2 .................................................. 42

2.4.1.3 Example 3 .................................................. 43

2.4.2 Selection Criterion Study .......................................... 45

2.4.2.1 Theoretically Optimal Bandwidth and Mixing Parameter .......... 46

2.4.2.2 PRESS* Results ........................................... 48

2.4.2.3 PRESS** Results ........................................... 50

2.4.2.4 PRESS* versus PRESS** Summary ............................. 52

3 Graphical Study

3.1 Graphical Approach of Examples, and AVEMSE Results .................. 53

3.1.1 Summary of Example 1 ........................................... 55

3.1.2 Summary of Example 2 ........................................... 58

3.1.3 Summary of Example 3 ........................................... 61
3.1.4 Summary of AVEMSE Examples ........................................... 62

3.2 Graphical Selection Criterion Study ................................. 65

3.2.1 Graphical PRESS* Results ............................................. 66
  3.2.1.1 Data Set 1 .................................................. 66
  3.2.1.2 Data Set 2 .................................................. 70
  3.2.1.3 Data Set 3 .................................................. 72
  3.2.1.4 Data Set 4 .................................................. 75
  3.2.1.5 Data Set 5 .................................................. 77
  3.2.1.6 Graphical PRESS* Results Summary ......................... 80

3.2.2 Graphical PRESS** Results .......................................... 81
  3.2.2.1 Data Set 1 .................................................. 82
  3.2.2.2 Data Set 2 .................................................. 86
  3.2.2.3 Data Set 3 .................................................. 88
  3.2.2.4 Data Set 4 .................................................. 91
  3.2.2.5 Data Set 5 .................................................. 93
  3.2.2.6 Graphical PRESS** Results Summary ......................... 96

4 Conclusion and Future Research ........................................ 101
List of Figures

3.1 AVEMSE₀ Plot for Example 1, using MRR1 as Fitting Technique . . . . . . 56
3.2 AVEMSE₀ Plot for Example 1, using MRR2 as Fitting Technique . . . . . . 57
3.3 AVEMSE₀ Plot for Example 2, using MRR1 as Fitting Technique . . . . . . 59
3.4 AVEMSE₀ Plot for Example 2, using MRR2 as Fitting Technique . . . . . . 60
3.5 AVEMSE₀ Plot for Example 3, using MRR1 as Fitting Technique . . . . . . 61
3.6 AVEMSE₀ Plot for Example 3, using MRR2 as Fitting Technique . . . . . . 62
3.7 PRESS* Plot for Data Set 1, using MRR1 as Fitting Technique . . . . . . 67
3.8 PRESS* Plot for Data Set 1, using MRR2 as Fitting Technique . . . . . . 69
3.9 PRESS* Plot for Data Set 2, using MRR1 as Fitting Technique . . . . . . 71
3.10 PRESS* Plot for Data Set 2, using MRR2 as Fitting Technique . . . . . . 71
3.11 PRESS* Plot for Data Set 3, using MRR1 as Fitting Technique . . . . . . 73
3.12 PRESS* Plot for Data Set 3, using MRR2 as Fitting Technique . . . . . . 74
3.13 PRESS* Plot for Data Set 4, using MRR1 as Fitting Technique . . . . . . 75
3.14 PRESS* Plot for Data Set 4, using MRR2 as Fitting Technique . . . . . . 76
3.15 PRESS* Plot for Data Set 5, using MRR1 as Fitting Technique . . . . . . 78
3.16 PRESS* Plot for Data Set 5, using MRR2 as Fitting Technique . . . . . . 79
3.17 PRESS** Plot for Data Set 1, using MRR1 as Fitting Technique . . . . . . . 83
3.18 PRESS** Plot for Data Set 1, using MRR2 as Fitting Technique . . . . . . . 85
3.19 PRESS** Plot for Data Set 2, using MRR1 as Fitting Technique . . . . . . . 87
3.20 PRESS** Plot for Data Set 2, using MRR2 as Fitting Technique . . . . . . . 87
3.21 PRESS** Plot for Data Set 3, using MRR1 as Fitting Technique . . . . . . . 89
3.22 PRESS** Plot for Data Set 3, using MRR2 as Fitting Technique . . . . . . . 90
3.23 PRESS** Plot for Data Set 4, using MRR1 as Fitting Technique . . . . . . . 91
3.24 PRESS** Plot for Data Set 4, using MRR2 as Fitting Technique . . . . . . . 92
3.25 PRESS** Plot for Data Set 5, using MRR1 as Fitting Technique . . . . . . . 94
3.26 PRESS** Plot for Data Set 5, using MRR2 as Fitting Technique . . . . . . . 95
List of Tables

1.1 Efficiencies of Kernel functions. .................................................. 10

2.1 Comparison of Example 1 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously) 41
2.2 Comparison of Example 2 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously) 43
2.3 Comparison of Example 3 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously) 45
2.4 Comparison of Optimal AVEMSE and INTMSE (Separately vs Simultaneously) . 47
2.5 Comparison of Fits When Using PRESS* .................................... 49
2.6 Comparison of Fits When Using PRESS** ................................. 51
Abstract

A GRAPHICAL ANALYSIS OF SIMULTANEOUSLY CHOOSING THE BANDWIDTH AND THE MIXING PARAMETER FOR SEMIPARAMETRIC TECHNIQUES

By Derick L. Rivers, Master of Science

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2009.

Director: James E. Mays, Associate Professor, Department of Statistical Sciences and Operations Research

There has been extensive research done in the area of Semiparametric Regression. These techniques deliver substantial improvements over previously developed methods, such as Ordinary Least Squares and Kernel Regression. Two of these hybrid techniques: Model Robust Regression 1 (MRR1) and Model Robust Regression 2 (MRR2) require the choice of an appropriate bandwidth for smoothing and a mixing parameter that allows a portion of a nonparametric fit to be used in fitting a model that may be misspecified by other regression methods. The current method of choosing the bandwidth and mixing parameter does not guarantee the optimal choices in either case. The immediate objective of the current work is to address this process of choosing the optimal bandwidth and mixing parameter and to examine the behavior of these estimates using 3D plots. The 3D plots allow us to examine how the semiparametric techniques: MRR1 and MRR2, behave for the optimal (AVEMSE) selection process when compared to data-driven selectors, such as PRESS* and PRESS**.
It was found that the structure of MRR2 behaved consistently under all conditions. MRR2 displayed a wider range of “acceptable” values for the choice of bandwidth as opposed to a much more limited choice when using MRR1. These results provide general support for earlier findings by Mays et al. (2000).
Chapter 1

Introduction

The basic idea of regression is to use one or more explanatory, or regressor variables, $X_1$, $X_2$, $X_3$, ..., $X_k$, to explain the response of a variable $y$. This regression model is of the form

$$y_i = f(X_1, X_2, \ldots, X_k) + \epsilon_i, \ i = 1, \ldots, n.$$  \quad (1.1)

Regression requires some assumptions: one must assume that the $X_i$ are nonrandom and observed with negligible measurement error, while the $\epsilon_i$, the error terms, are random variables such that $\epsilon_i \sim (0, \sigma^2)$ (homogeneous variance assumption). Myers (1990) explains that the $E(\epsilon_i) = 0$ and $E(\epsilon_i^2) = \sigma^2$. In addition, we assume that the $\epsilon_i$ are uncorrelated from observation to observation. The classical parametric regression estimation method where there is a known linear form for $f$ is called Ordinary Least Squares (OLS). OLS provides
optimal prediction capabilities as long as the model is not misspecified. If \( f \) is misspecified, even over portions of the data, the analysis can be quite misleading.

If the true form of \( f \) is unknown, then an alternative to parametric regression is nonparametric regression. Nonparametric regression techniques rely solely on the data points themselves to determine the regression fit rather than an assumed underlying function, \( f \). There is a plethora of procedures available that address such situations. The two techniques that will be mentioned in the current work are Kernel Regression and Local Polynomial Regression. Eubank (1999) states that parametric methods can be quite dangerous in situations where there is little that is known about the regression function and that the bulk of the information about \( f \) lies within the data.

The idea of the nonparametric regression procedure is to use the points closest to \( \mathbf{x}_0 \) in order to get the best information. For example, consider a point \( \mathbf{x}_0 = (x_{10}, \ldots, x_{k0})' \) where the prediction of \( \mathbb{E}(y) = f(x_0) \) is desired. If \( f \) is a smooth function, then the best information on \( f(x_0) \) should come from the \( y \) values at the locations \( \mathbf{x}_i \) closest to \( \mathbf{x}_0 \). In Kernel Regression, the value of \( f(x_0) \) at \( \mathbf{x}_0 \) is expressed as a weighted sum of the responses, where the weights are dependent upon the distances of the regressors from the point of prediction. The further the distance from \( \mathbf{x}_0 \), the smaller the assigned weight to the observations at that location. A nonparametric fit is obtained once predictions at all of the regressor locations have been found. There isn’t a closed form expression for \( f \), but the obtained information will provide the practitioner a form to study (Mays, 1995).

Nonparametric regression has its drawbacks, since it is completely reliant upon the data
points themselves. It can have more variability when compared to the Ordinary Least Squares technique. Secondly, nonparametric fits may fit to irregular or over exaggerated patterns in the data. The largest drawback of this technique is how best to determine the weighting scheme that will be used.

Semiparametric regression offers solutions to the drawbacks found with parametric and nonparametric regression. This hybrid technique combines both parametric regression and nonparametric regression into one model. The problem at hand is that some semiparametric models require the choice of two parameters: an appropriate bandwidth, $b$ (for smoothing) and a mixing parameter, $\lambda$ (adds appropriate mix of parametric and nonparametric components together). The current work will investigate the simultaneous selection process of these two parameters by the use of 3-D plots. The current work attempts to answer the following questions:

1. Which is more important, $b$ or $\lambda$?

2. How changing $b$ or $\lambda$ will affect the surface?

3. If $b$ is wrong, can $\lambda$ give a “reasonable” fit?

4. If $b$ is correct, what impact does $\lambda$ have?

5. Will data-driven selectors display the same pattern as the “optimal” process?

These questions will be answered by analyzing two semiparametric regression techniques: Model Robust Regression 1 and Model Robust Regression 2. The two data-driven selectors
that will be compared to a theoretically “optimal” selection process will be PRESS* and PRESS**.

1.1 Ordinary Least Squares

Takezawa (2006) states that parametric regression makes use of regression equations such as

\[ y = ax + b, \quad (1.2) \]
\[ y = px^2 + qx + r, \quad (1.3) \]
\[ y = a_1 \sin(b_1 x) + a_2 \cos(b_2 x), \quad (1.4) \]
\[ y = \frac{s_1}{1 + s_2 \exp(s_3 x)}, \quad (1.5) \]

where \( x \) is a predictor, \( y \) is a response variable, and \( a, b, p, q, r, a_1, b_1, a_2, b_2, s_1, s_2, \) and \( s_3 \) are constants called regression coefficients and are unknown parameters. Parametric regression is the estimation of the values of parameters (regression coefficients) using the given data and a weighting scheme based on a chosen parametric model. Regression equations in (1.2) and (1.3) are called linear models because the parameters in these two equations appear in the model in a linear fashion. The derivation of a linear model is called linear regression. In equations (1.4) and (1.5) the parameters do not appear in a linear form. The derivation of these equations is known as nonlinear regression (Takezawa, 2006). Though equation 1.5 may correspond to a generalized linear model via a link function.
Ordinary Least Squares, a method that was exclusively used until the 1970s, assumes a known form for \( f \), where we minimize the sum of the squared residuals, \( \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \) in order to obtain the fitted or predicted values, \( \hat{y}_{ols} \). Myers (1990) states that the model may be rewritten in matrix notation as

\[
y = X\beta + \epsilon, \tag{1.6}
\]

where \( y \) is an \( n \) dimensional response vector and of interest is to estimate the unknown parameters in \( \beta \), which is a \( k + 1 \) dimensional vector. Here \( X \) is a non-singular \( n \times (k + 1) \) model matrix of \( k \) regressors augmented with a column of ones and \( \epsilon \) is a \( n \) dimensional vector of random errors. Once an estimate of \( \beta \), \( \hat{\beta} \), is determined, prediction of the response variable is possible, and then one can obtain the fitted values as \( \hat{y} = X\hat{\beta} \). The fitted values \( \hat{y} = X\hat{\beta} \) can be expressed as

\[
\hat{y} = X(X'X)^{-1}X'y = H^{(ols)}y, \tag{1.7}
\]

where \( H^{(ols)} \) is the Ordinary Least Squares “hat” matrix. The fit for \( y_i \) at location \( X_i \) may be expressed as

\[
y_i^{(ols)} = \sum_{j=1}^{n} h_{ij}^{(ols)} y_i, \tag{1.8}
\]
a weighted sum (for \( j = 1, \ldots, n \)) of the \( n \) observations \( y_j \), where \( h_{i1}^{(ols)}, \ldots, h_{in}^{(ols)} \) are the elements of the \( i^{th} \) row of \( H^{(ols)} \).

**Hat Matrix**

The hat matrix is very important in inferences that are based on Ordinary Least Squares and several of the procedures described in this current work. Myers (1990) defines the properties of the \( HAT \) matrix as follows:

1. \(-1 \leq h_{ij}^{(OLS)} \leq 1\)

2. \( \frac{1}{n} \leq h_{ii}^{(OLS)} \leq 1\)

3. \( \text{trace}(H^{(OLS)}) = \sum_{i=1}^{n} h_{ii}^{(OLS)} = p \) (where \( p = k + 1 \))

4. \( \sum_{j=1}^{n} h_{ij}^{(OLS)} = 1 \) (\( i^{th} \) row sums to 1)

5. \( \text{Var}(\hat{y}_{OLS}) = \text{Var}(H^{(OLS)}y) = \sigma^2 H^{(OLS)} H^{(OLS)} = \sigma^2 H^{(OLS)} \)

6. Residual \( e^{(OLS)} = y - \hat{y}_{(OLS)} = (I - H^{(OLS)})y \)
7. \( \hat{\sigma}^2_{(OLS)} = \frac{\sum e^2_{(OLS)}}{n-p} = \frac{\sum e^2_{(OLS)}}{\text{trace}([I-H_{(OLS)}][I-H_{(OLS)}]^T]} \)

Since the fit for \( y_i \) at \( x_i \) can be expressed as \( \hat{y}^{(OLS)}_i = \sum_{j=1}^{n} h_{ij}^{(OLS)} y_j \), an observation with a large \( h_{ij} \) has a significant influence on the regression fit for \( \hat{y}_i \).

The problem that arises with such techniques as OLS is that this method relies on a weighting scheme that is directly related to the assumed model. In the case of a simple linear regression model (one regressor, \( x \)), each \( h_{ij} \) is calculated by

\[
    h_{ij} = \frac{1}{n} \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},
\]

where \( (x_i - \bar{x}) \) represents the distance that \( x_j \) is from \( \bar{x} \). Data points at location \( x_j \) far from \( \bar{x} \) will have heavy influence on the fit, while data points at location \( x_j \) close to \( \bar{x} \) will have little or no influence on the fitted value, \( \hat{y}^{(OLS)}_i \).

Prediction capability is most likely the ultimate goal, and Ordinary Least Squares will derive optimal results if the model is chosen correctly. However, if the assumed function \( f \) is misspecified, OLS will deliver poor predictions at some data points, and the inferences could be quite misleading. Due to possible model misspecification, even over portions of the data, inferences based on this technique will be inaccurate. When in doubt about the true model, and we need an alternative to OLS where \( f \) is considered to be unknown and we can use a weighting scheme that places more weight on observations close to the point of prediction rather than on observations far away, we can consider Nonparametric Regression (Mays and Birch, 2002).
1.2 Nonparametric Regression

Nonparametric Regression techniques rely solely on the data themselves to determine the regression fit. This technique relaxes the assumption of linearity, substituting the much weaker assumption of a smooth population regression function $f(x)$. The cost of relaxing the assumption of linearity is substantially greater computation and, in some instances, a more difficult to understand result. The gain is potentially a more accurate estimate of the regression function (Fox, 2000).

The current work will only consider the case where there is only a single predictor $x$ in the model. Fox (2000) explains that nonparametric simple regression is useful for two reasons:

1. Nonparametric simple regression is often called scatterplot smoothing, because in typical applications, the method passes a smooth curve through the points in a scatterplot of $y$ against $x$.

2. Nonparametric simple regresion forms the basis, by extension, for nonparametric multiple regression.

Since nonparametric techniques are not dependent upon a known function $f$, problems of misspecification are resolved.
1.2.1 Kernel Regression

A commonly used procedure for obtaining the weighting scheme is known as kernel regression. \( K(u) \), the kernel function, is a decreasing function of \(|u|\). The kernel function may be any continuous probability density function that is bounded and integrates to 1. Fan and Gijbels (1996) list two commonly used kernel functions, the Gaussian kernel, \( K(u) = (\sqrt{2\pi})^{-1} e^{-u^2/2} \), and the ‘symmetrical Beta family’ kernel

\[
K(u) = \frac{1}{\text{Beta}(\frac{1}{2}, \gamma + 1)}(1 - u^2)^\gamma, \gamma = 0, 1, \ldots
\] (1.10)

where the subscript + denotes the positive part, which is assumed to be taken before the exponentiation. Fan and Gijbels (1996) state that the choices \( \gamma = 0, 1, 2 \) and 3 lead to respectively the uniform, the Epanechnikov, the biweight and the triweight kernel functions. Gasser et al. (1979) found the optimal kernel, with respect to Mean Integrated Squared Error, to be the Epanechnikov kernel (Epanechnikov, 1969):

\[
K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1),
\] (1.11)

where \( I \) is the indicator function. Härdle (1990) calculates the efficiencies for several of the commonly used kernel functions. The results are in Table 1.1.

The efficiencies are calculated from the optimal Mean Integrated Squared Error (MISE) of the predicted function \( \hat{f}(x) \) from the Epanechnikov kernel divided by the MISE of the
Table 1.1: Efficiencies of Kernel functions.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>( K(u) )</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epanechnikov</td>
<td>( \frac{3}{4} (1 - u^2) I(</td>
<td>u</td>
</tr>
<tr>
<td>Quartic</td>
<td>( \frac{15}{16} (1 - u^2)^2 I(</td>
<td>u</td>
</tr>
<tr>
<td>Triangular</td>
<td>( (1 -</td>
<td>u</td>
</tr>
<tr>
<td>Guass</td>
<td>( (\sqrt{2\pi})^{-1} e^{-\frac{u^2}{2}} )</td>
<td>0.961</td>
</tr>
<tr>
<td>Uniform</td>
<td>( \left(\frac{1}{2}\right) I(</td>
<td>u</td>
</tr>
</tbody>
</table>

particular kernel of interest. Härdle suggests that the choice of the kernel function should be based on other considerations besides the MISE, such as computational efficiency. Härdle (1990) also claims that the choice of the kernel function is negligible to the fit of the values. Thus, for computational ease we will use the the simplified Guassian (Normal) kernel, which can be expressed as:

\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \quad (1.12)
\]

Kernel fits are given by

\[
\hat{y}_{i}^{(ker)} = \hat{f}(X_i) = \sum h_{ij}^{(ker)} y_j, \text{ for } j = 1, \ldots, n. \quad (1.13)
\]
This may be expressed in matrix notation as

\[ \hat{y}^{(ker)} = H^{(ker)}y, \] (1.14)

where \( H^{(ker)} = h_{ij}^{(ker)} \) is the kernel “hat” matrix. Nadaraya (1964) and Watson (1964) introduced what is now a common way of obtaining the weights \( h_{ij}^{(ker)} \):

\[ h_{ij}^{(ker)} = \frac{K(X_i - X_j)}{\sum_{j=1}^{n} K(X_i - X_j)}, \] (1.15)

where \( b > 0 \) is the bandwidth (smoothing parameter), which determines how rapidly the weights decrease as the distance from the location being fit increases. Other weighting schemes can be found in Priestley and Chao (1972) and Chu and Marron (1991).

This approach assumes that the points, \( x_i \), closest to the point of prediction, \( x_0 \), contain the most information about obtaining the fitted values. This method uses a bandwidth to determine the weights given to any particular observation. The choice of the bandwidth is much more critical than that of the kernel function for obtaining the optimal fit. The fundamental purpose of kernel regression is to determine an appropriate weighting scheme to yield the fitted values, \( \hat{y}_i^{(ker)} = \sum h_{ij}^{(ker)} y_j \), for \( j = 1, \ldots, n \) and where \( h_{ij}^{(ker)} \) are the weights from the \( i^{th} \) row of the kernel hat matrix. The numerator of the kernel hat matrix, \( K \left( \frac{X_i - X_j}{b} \right) \), guarantees that the observations close to \( x_i \) are given more weight than the observations that are farther away. The denominator insures that each row of the hat matrix sums to one, as
in OLS.

1.2.2 Bandwidth Choice

The function of the bandwidth \( b \) is to control the amount of weight (averaging) given to each of the observations. Rice (1984) and Chiu (1991) conclude that the choice of bandwidth is crucial in estimating \( f(x) \). Varying the bandwidth of the kernel estimator controls the smoothness of the curve and determines how quickly the weights decrease for observations away from where we are attempting to predict. Although increasing the bandwidth helps decrease the variance of the fit, it also causes an increase in the bias term. When the bandwidth is too large (close to the range of the \( x \) values), we can have spikes or dips in the data because even points far from the location being fit are given similar weights to those at points close by. The result is a curve that is underfit, or oversmoothed. At the other extreme, choosing a smaller bandwidth helps to decrease bias, but increases the variance term because more weight is placed on the point of prediction itself. The resulting curve is said to be overfit, or undersmoothed. Hence the choice of bandwidth must be a compromise between variance and bias. Since, \( MSE = Variance + Bias^2 \), a natural criterion used for selection of bandwidth would involve the minimization of some type of mean squared error (MSE). Three measures of MSE presented by Härdle (1990) are as follows: integrated squared error (ISE), conditional average squared error (CASE), and a discrete form of ISE called average squared error (ASE). These measures are defined as follows:
\[ \text{ASE} = d_A(b) = \frac{\sum_{j=1}^{n} \left( \hat{f}_b(x_j) - f(x_j) \right)^2 w(x_j)}{n}, \]  
\[ \text{ISE} = d_I(b) = \int \left( \hat{f}_b(x_j) - f(x_j) \right)^2 w(x) g(x) dx, \]  
\[ \text{CASE} = d_C(b) = E[d_A(b) \mid x_1, \ldots, x_n], \]  

where \( \hat{f}_b \) is the kernel estimate of \( f \), \( g(x) \) is the density of the \( X \)'s (equals 1 if \( X \)'s are fixed), and \( w(x) \) is a nonnegative weight function that is present to reduce boundary effects on rates of convergence. Härdle (1990) found that \( w(x) \) was not significantly influential in the choice of \( b \), thus \( w(x) \) will be taken as a constant value for simplicity.

Due to its computational simplicity, ASE will be used in the current work as a performance criterion for choosing the bandwidth. Härdle (1990) states that one must estimate ASE or an equivalent measure (up to some constant) by a smoothing parameter selection process, and hopefully the smoothing parameter that minimizes this estimate is also a good estimate for the smoothing parameter that minimizes the ASE itself. The portion of ASE that must be estimated is what will be the criterion for selecting the bandwidth. The prediction error, \( p(b) \), involving the sum of squared errors,

\[ p(b) = \frac{SSE}{n} = \frac{\sum_{j=1}^{n} \left( y_j - \hat{f}_b(x_j) \right)^2 w(x_j)}{n}, \]

was found to be a good estimate. As shown in Härdle (1990), \( \frac{SSE}{n} \) is a biased estimate of ASE and tends to overfit the data by choosing the smallest possible bandwidth. Three
techniques that are proposed by Härdle (1990) to find an unbiased estimate of (ASE) are the leave-one-out (cross validation), penalizing functions and plug-in methods.

1.2.2.1 Cross Validation

The cross validation method is based on making the fitted value at location \( x_i \) and the corresponding observed value independent by omitting the \( i^{th} \) observation at the location \( x_i \) being fit. Independence is achieved through removing the data points, \( y_i \), one at a time and then obtaining fitted values, \( \hat{y}_{i,-i} \), through a weighted sum of the remaining observations. The cross validation in essence is the PRESS (prediction error sum of squares) statistic that is commonly used in parametric regression. For further explanation of the PRESS statistic see Myers (1990). This method is defined as follows:

\[
PRESS = CV(b) = \frac{\sum_{j=1}^{n} [y_i - \hat{y}_{i,-i}]^2 w(x_i)}{n}.
\]  

(1.20)

1.2.2.2 Penalizing Functions

The second method proposed by Härdle (1990) for choosing \( b \) based on MSE uses penalizing functions to adjust \( p(b) \) so that small values for \( b \) are less likely to be chosen. This method accomplishes the penalizing for small values of \( b \) by the way of a penalizing function \( \Xi(u) \), which is increasing in \( u \). The general idea of the function is to adjust the biased prediction error, \( p(b) \), of equation (1.19) by
\[ n^{-1}b^{-1}K(0) \left( \frac{y_j - \hat{f}_b(x_j)}{\hat{f}_b(x_j)} \right), \quad (1.21) \]

where \( \hat{f}_b(x_j) \), as described by Härdle (1990) is the Rosenblatt-Parzen kernel density estimator of the marginal density of \( X \) at the value \( x_j \). The equation is modified to

\[ G(b) = n^{-1}\sum_{j=1}^{n} \left( y_j - \hat{f}_b(x_j) \right)^2 \Xi \left( n^{-1}b^{-1}K(0) \right), \quad (1.22) \]

### 1.2.2.3 Plug-In Method

The third method for selecting \( b \) proposed by Härdle (1990) is the “Plug-In” procedure.

The plug-in methods approximate the bias of an estimate \( \hat{f} \) as a function of the unknown \( f \) through a Taylor expansion,

\[ f(x) = f(b) + f'(b)(x-b) + \frac{f''(b)}{2!}(x-b)^2 + \frac{f'''(b)}{3!}(x-b)^3 + \cdots + \frac{f^{(n)}(b)}{n!}(x-b)^n \ldots (1.23) \]

Loader (1999) states that the “optimal” \( b \) minimizes this estimated measure of fit (i.e. expression for estimate for ASE) There is some evidence by Loader (1999) that indicates plug-in estimates tend towards oversmoothing (bandwidth too large) and have large biases if the initial bandwidth has been misspecified.
1.2.2.4 Bootstrapping

The method of bootstrapping is another technique that is considered in the process of bandwidth selection. The bootstrap procedure as defined by Efron and Tibshirani (1993) was developed for constructing sampling distributions empirically from the data at hand. To obtain a bootstrap sample from an original sample of size $n$, one draws many (B) samples, each of size $n$, with replacement. We choose $\theta$ (such as the median) as the parameter of interest and one can obtain an estimate $\hat{\theta}$ of $\theta$ from the original sample. From the original sample, resample with replacement a bootstrap sample of size $n$ to obtain an estimate $\theta^*$ of $\theta$ from each of the B bootstrap samples. The bootstrap principle is that the observed distribution of the $\theta^*$'s approximates the true distribution of $\hat{\theta}$. This concludes that the distribution of the $\theta^*$'s can be used to gain insight about the true behavior of the estimate $\hat{\theta}$. For a brief discussion of bootstrapping applied to actual data, see Stine (1989). Faraway (1990) uses bootstrapping as the method of choosing the bandwidth in kernel regression and describes how to form the estimate of the MSE so that it is consistent with the true MSE. Since bootstrapping is a highly computer intensive procedure, this technique for bandwidth selection will not be studied any further in this current work. More detail can be found in Faraway (1990).
1.2.2.5 \( \text{PRESS}^* \)

In the current work, two of the aforementioned techniques for bandwidth selection are combined to give a selection criterion. Mays and Birch (1996) developed an adaptation to the traditional PRESS statistic, where the denominator imposes a penalty for choices of small bandwidths.

\[
\text{PRESS}^* = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,n-i})^2}{n - \text{tr}(\mathbf{H}^{(ker)})} = \frac{\text{PRESS}}{n - \text{tr}(\mathbf{H}^{(ker)})}
\]  

(1.24)

For example, suppose we are fitting at the point \( x_i \). As the bandwidth gets smaller, the \( h_{ii}^{(ker)} \) get larger and so \( \text{tr}(\mathbf{H}^{(ker)}) \) also increases, but \( n - \text{tr}(\mathbf{H}^{(ker)}) \) gets smaller, thereby penalizing for small bandwidths. Mays and Birch (2002) indicated that while PRESS* did protect against choices of bandwidths too small, the tendency was to choose bandwidths that were too large, often yielding bandwidths greater than or equal to 1 when scaling the regressor to be between 0 and 1.

1.2.2.6 \( \text{PRESS}^{**} \)

In response to the shortcoming of PRESS*, a new statistic, PRESS** was proposed by Mays and Birch (2002) that incorporates a comparable penalty for choices of bandwidth too large.

\[
\text{PRESS}^{**} = \frac{\text{PRESS}}{n - \text{tr}(\mathbf{H}^{(ker)}) + (n - 1) \frac{\text{SSE}_{\text{mean}} - \text{SSE}_b}{\text{SSE}_{\text{mean}}}}
\]

(1.25)
Since this new penalty term approaches 0 as \( b \) approaches 1 and approaches \( n - 1 \) as \( b \) approaches 0, it imposes a penalty for choosing bandwidth values too large that is comparable to the penalty term, \( n - tr(H^{(ker)}) \), included for small bandwidths (Mays and Birch, 2002).

### 1.2.3 Local Polynomial Regression

Kernel regression has the drawback that it inaccurately fits the regression curve at the boundaries of the data. For example, when predicting \( y_0 \) at a point \( x_0 \) close to the extreme left boundary, the only points available to aid in prediction are \( x_0 \) and those observations to the right of \( x_0 \). Should the data show an increasing trend then all the observations to the right of \( x_0 \) will correspond to \( y \)-values greater than or equal to \( y_0 \). This will yield a prediction of \( y_0 \) which is too high. A second drawback of kernel regression is the possible increase in both bias and variance when there is significant curvature found in the true regression function. The third drawback is if the \( x \)'s have a non-uniform distribution, it can fail to capture the true trend because of limitations in the data. One technique for dealing with the problems of kernel regression is Local Polynomial Regression, LPR, developed by Cleveland (1979).

Local Polynomial Regression (LPR) is an adaptation of iterated weighted least squares (IWLS) (Hardle, 1990) based on an initial kernel fit of the data. Local polynomial regression uses weighted least squares (WLS) regression to fit a \( d^{th} \) degree polynomial, 

\[
y_i = a + b_1(x_i - x_0) + b_2(x_i - x_0)^2 + \cdots + b_d(x_i - x_0)^d + e_i, \quad (d \neq 0)
\]

to data. The initial kernel regression fit is used to determine the weights assigned to the observations. Kernel regression, as
described previously, is just a special form of local polynomial regression with \( d = 0 \). Mays (1995) noted local polynomial regression addresses the problem of potentially inflated bias and variance in the interior of the \( x \)'s if the \( x \)'s are nonuniform or if substantial curvature is present in the underlying, though undefined, regression function (a problem common to simpler weighting schemes such as kernel regression).

Consider fitting \( y_i \) at the point \( x_j \). First, a kernel fit is obtained for the entire data set in order to obtain the kernel hat matrix, \( W_{(ker)} \),

\[
W_{(ker)} = \begin{bmatrix}
w_1'(ker) \\
\vdots \\
w_n'(ker)
\end{bmatrix}
\]

where \( w_i'(ker) \) is the \( i \)th row of \( W_{(ker)} \) (same as \( H_{(ker)} \) in section 1.2.1). The kernel hat matrix can be used to obtain the fitted values for \( y_i \) at location \( x_i \) via

\[
\hat{y}_{(ker)}(x_i) = \sum_{j=1}^{n} w_{ij}(ker) y_j = w_i'(ker)y
\] (1.26)

where the \( w_{ij}(ker) \), \( j = 1, 2, \ldots, n \) are the \( n \) elements of the \( i \)th row of the \( W_{(ker)} \). The kernel weights \( w_{ij}(ker) \) give weights to \( y_i \) based on the location \( x_j \) from \( x_i \). With local polynomial regression though, the \( w_{ij}(ker) \), for a fixed \( i \), become the weights to be used in weighted least squares regression. It is important to note that these distinct weights vary with changing \( i \) (Mays, 1995).
As described by Mays (1995), the diagonal weight matrix for the local polynomial regression, used for fitting at $x_i$, can be written as:

$$W_{\text{diag}(x_i)} = \text{diagonal}(w_i(\text{ker})) = \begin{bmatrix} w_{i1}(\text{ker}) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_{in}(\text{ker}) \end{bmatrix}$$

Following the procedures of weighted least squares, the estimated coefficients for the local polynomial regression fit are then found via

$$\hat{\beta}_{i(\text{LPR})} = (X'W_{\text{diag}(x_i)}X)^{-1}X'W_{\text{diag}(x_i)}y$$

(1.27)

where $X$ is the $X$ matrix from the local polynomial regression determined by the degree $d$ of the polynomial, with the $i^{th}$ row defined as $x'_i$. Thus, provided $(X'W_{\text{diag}(x_i)}X)^{-1}$ exists, the fit at $x_i$ is obtained as:

$$\hat{y}_{i(\text{LPR})} = x'_i\hat{\beta}_{i(\text{LPR})} = x'_i(X'W_{\text{diag}(x_i)}X)^{-1}X'W_{\text{diag}(x_i)}y = w'_{i(\text{LPR})}y$$

(1.28)

As Mays (1995) noted, the $n$ fitted values can be expressed in matrix notation as

$$\hat{y} = W_{(\text{LPR})}y$$

(1.29)
where

\[
W_{LPR} = \begin{bmatrix}
w'_1(LPR) \\
\vdots \\
w'_n(LPR)
\end{bmatrix}.
\]

Further development of this technique is provided in Cleveland (1979) and Hastie and Loader (1993). Authors generally agree that for the majority of cases, a first order fit (local linear regression) is an adequate choice for \(d\). Local linear regression is suggested by Cleveland (1979) to balance computational ease with the flexibility to reproduce patterns that exist in the data. Nonetheless, local linear regression may fail to capture sharp curvature if present in the data structure (Mays 1995). In such cases, local quadratic regression (\(d = 2\)) may be needed to provide an adequate fit. Most authors agree there is usually no need for polynomials of order \(d > 2\) (Mays 1995). A bandwidth as large as possible is recommended by Cleveland (1979) since increasing the bandwidth increases the smoothness of the regression curve and minimizes variability without compromising the structure of the data. Since LPR protects against boundary bias as well as other problems associated with kernel regression, it is used in the current research and has been validated for LLR by Mays and Birch (2002).
1.2.4 Spline Smoothing

Spline smoothing is a nonparametric regression method that quantifies the trade-off between

1. the goodness-of-fit to the data

2. the smoothness of the estimated curve without too much rapid local variation.

Splines are defined as piece-wise polynomial functions that fit together at knots. The regression curve $\hat{f}_\delta(x)$ is obtained by minimizing the penalized residual sum of squares

$$S(f) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \delta \int_a^b \{f''(x)\}^2 dx$$ (1.30)

where $f$ is a twice-differentiable function on $[a,b]$, and $\delta$ represents the rate of exchange between residual error and roughness of the curve $f$ (Aydin, 2008).

The first term in equation (1.30) is the residual sum of squares and it penalizes the lack-of-fit. The second term, which is weighted by the parameter $\delta > 0$, is the roughness penalty, which penalizes the curvature of the function $f$. The parameter $\delta$ controls the trade-off between smoothness, as measured by $\int_a^b \{f''(x)\}^2 dx$, and the goodness of fit to the data, as measured by $\sum_{i=1}^{n} (y_i - f(x_i))^2$. Silverman (1984) explains that the larger the value of $\delta$, the more the data will be smoothed to produce the curve estimate.

The function $\hat{f}(x)$ that minimizes equation (1.30) is a natural cubic spline with knots at the distinct observed values of $x$ (Fox, 2002). A spline is considered a cubic spline when the first and second derivatives are continuous at the knots. Natural splines place two additional knots at the ends of the data, and constrain the function to be linear beyond these points. Fox (2002) states that although this result seems to imply that $n$ parameters are required
(when all \(x\)-values are distinct), the roughness penalty imposes additional constraints on the solution, typically reducing the equivalent number of parameters for the smoothing spline substantially, and preventing \(\hat{f}(x)\) from interpolating the data. Fox (2002) states that it is common to select the smoothing parameter \(\delta\) indirectly by setting the equivalent number of parameters for the smoother.

Aydin (2008) defines \(f = (f(x_1), \ldots, f(x_n))\) to be the vector of values of function \(f\) at the knot points \(x_1, \ldots, x_n\). The smoothing spline estimate \(\hat{f}_\delta\) of this vector or the fitted values for the data \(y = (y_1, \ldots, y_n)'\) is given by

\[
\hat{f}_\delta = \begin{bmatrix} \hat{f}_\delta(x_1) \\ \vdots \\ \hat{f}_\delta(x_n) \end{bmatrix} = (S_\delta) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad \text{or} \quad \hat{f}_\delta = S_\delta y
\]

where \(S_\delta\) is the hat matrix defined by Eubank (1999) as

\[
X(X'X + n\delta\Omega)^{-1}X'
\] (1.31)

and \(\Omega\) is defined as

\[
\left\{ \int_0^1 x_i^m(t)x_j^m(t)dt \right\}_{i,j=1,n}.
\] (1.32)

\(\hat{f}_\delta\) is a natural cubic spline with knots at \(x_1, \ldots, x_n\) for a fixed smoothing parameter \(\delta > 0\), and \(S_\delta\) is a well-known positive-definite (symmetrical) smoother matrix which depends on \(\delta\) and the knots points \(x_1, \ldots, x_n\), but not on \(y\) (Aydin, 2008).
Selection of $\delta$

This parameter, $\delta$, may be selected by visual trial and error, picking a value that balances smoothness against the precision of the data. More formal methods of selecting smoothing parameters typically try to minimize the MSE of the fit, either by approximating the MSE by way of “plug-in” estimates or by some form of cross-validation (Fox, 2002). Aydin (2008) used R and S-Plus programs for the choice of the smoothing parameter, $\delta$. They each chose the $\delta$ using either ordinary or generalized cross validation methods.

Local polynomial regression (LLR specifically) is used as the nonparametric fitting technique in this work. Nonparametric regression (whether it be LLR, spline regression, or other options) has its drawbacks, since it is completely reliant upon the individual data values. It can have more variability in model fitting when compared to the ordinary least squares technique. Secondly, the choice of bandwidth is quite complicated, but is still important to the fit of the nonparametric technique. Because of the disadvantages of this technique, interest has been given to methods that combine both parametric regression and nonparametric regression into one model; this is known as Semiparametric Regression.

1.3 Semiparametric Regression

There have been developments in regression techniques that combine the parametric and nonparametric procedures mentioned previously. This collaboration of techniques allows the
practitioner to incorporate any known information about the parametric model as well as identify any deviations in the data from this model. Semiparametric regression is summarized as being able to use a parametric fit based on certain regressors and a nonparametric fit on the other regressors (Speckman, 1988). See Härdle(1990) for a brief introduction to semiparametric approaches and the various forms that they can take. Semiparametric models combine components of parametric and nonparametric, keeping the easy interpretability of the former and retaining the flexibility of the latter. The semiparametric model can be expressed as

$$y_i = z_i'\beta + f(t_i) + \epsilon_i, \ i = 1, 2, 3, \ldots, n,$$

where $\beta$ is a vector of unknown regression coefficients and $f$ is an unknown, smooth regression function. The response variable, $y$, depends on a parametric component $z$, in addition to nonparametric component, $t$ (Speckman, 1988). Such models are generally referred to as partial linear models due to the linear structure of the parametric portion of the model. Three (3) semiparametric techniques for fitting such models are as follows: Partial Linear Regression (PLR), Model Robust Regression 1 (MRR1), and Model Robust Regression 2 (MRR2).
1.3.1 Partial Linear Regression (PLR)

The partial linear model can be expressed as:

\[
y_i = \mathbf{x}_i' \beta + f(t_i) + \epsilon_i, \quad i = 1, 2, \ldots, n,
\]

(1.34)

where \( \mathbf{x}_i' \) are fixed known \( 1 \times k \) vectors, \( \beta \) is a vector of unknown parameters, \( t_i \) is the \( i^{th} \) value of the regressor variable \( T \), and \( f \) is an unknown and smooth regression function. In vector-matrix form we write model (1.34) as

\[
y_{(PLR)} = \mathbf{X}\beta + \mathbf{f} + \mathbf{\epsilon}
\]

(1.35)

with \( \mathbf{y} \) the response vector, \( \mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')' \), \( \mathbf{f} = (f(t_1), \ldots, f(t_n))' \) and \( \mathbf{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \) (Eubank, 1999).

One must obtain estimates of the unknown parameters for \( \beta \) and \( f \), \( \hat{\beta} \) and \( \hat{f} \), respectively in order to estimate the response variable, \( \hat{y} \). Speckman (1988) obtained these estimates by supposing that \( f \) could be parametrized as \( f = (f(t_1), \ldots, f(t_n))' = \mathbf{T}\gamma \), where \( \mathbf{T} \) is a \( n \times q \) matrix of full rank and \( \gamma \) is an additional parameter vector. Speckman (1988) assumed that the unit vector \( (1, \ldots, 1)' \) is in the span of \( \mathbf{T} \), but not of \( \mathbf{X} \) in order for the \( n \times (p + q) \) matrix \( (\mathbf{X}, \mathbf{T}) \) to have full rank. Model (1.35) as proposed by Eubank (1999) is now written in matrix notation as
\[ y_{(PLR)} = X\beta + T\gamma + \epsilon \] (1.36)

By taking the derivative of \( \epsilon^2 = (y - X\beta - T\gamma)'(y - X\beta - T\gamma) \), first with respect to \( \beta \) and then with respect to \( \gamma \), and then setting these equations equal to zero, one can obtain the following normal equations for Model (1.36):

\[ X'X\beta = X'(y - T\gamma), \] (1.37)
\[ T\gamma = P_T(y - X\beta), \] (1.38)

where \( P_T = T'(T'T)^{-1}T' \) denotes a hat matrix.

Green et al. (1985) obtained the estimates of \( \beta \) and \( T\gamma \) by substituting for \( T\gamma \) in equation (1.37) with equation (1.38) and solving for \( \beta \) to obtain

\[ \hat{\beta} = (X'(I - P_T)X)^{-1}X'(I - P_T)y. \] (1.39)

Green et al. also proposed replacing \( P_T \) with a “smoother” \( M \) to obtain the final estimates. Taking this smoother to be the kernel hat matrix \( H^{(ker)} \) from kernel smoothing defines the Green-Jennison-Scheuert (GJS) estimates as follow:

\[ \hat{\beta}_{GJS} = (X'(I - H^{(ker)})X)^{-1}X'(I - H^{(ker)})y \] (1.40)
\[ \hat{f}_{GJS} = T\hat{\gamma} = H^{(ker)}(y - X\hat{\beta}_{GJS}) \] (1.41)

Partial residuals for \( X \) and \( y \) (after adjusting for the nonparametric component of \( X \)) are constructed using kernel regression, and the estimate for \( \beta \) is found by regressing the
Derick L. Rivers  
Semiparametric Regression

partial residuals \( \tilde{y} = (I - H_p^{(\ker)})y \) on \( \tilde{X} = (I - H_p^{(\ker)})X_p \), where \( X_p \) is the usual “\( X \)” matrix without a column of ones. The estimate for \( f \) is found using a nonparametric fit to the residuals from the parametric fit. Mays et al. (2000) obtain estimates for \( \beta \) and \( f \) simultaneously as:

\[
\hat{\beta}_{(PLR)} = (\tilde{X}'\tilde{X})^{-1} \tilde{X}'\tilde{y} \tag{1.42}
\]

\[
\hat{f}_{(PLR)} = H_p^{(LPR)}(y - X_p\hat{\beta}_{(PLR)}), \tag{1.43}
\]

\( \tilde{X} \) and \( \tilde{y} \) are the partial residuals for \( X \) and \( y \) (after adjusting for the nonparametric component of \( X \)), \( y - X_p\hat{\beta}_{(PLR)} \) are residuals from the parametric fit \( \hat{\beta}_{(PLR)} \), and \( H_p^{(LPR)} \) is the hat matrix from a local polynomial fit to the residuals (Mays et al., 2000). There are several drawbacks when using this procedure. First, since both parametric and nonparametric fits are found simultaneously, the computations involved can be somewhat complex. Also, because of the elimination of the intercept term, the parametric portion always crosses the \( y \)-axis at 0, yielding it unsuitable when used without the nonparametric portion. In addition, even when a parametric fit would be suitable, thereby resulting in a smooth curve, the nonparametric fit is always included, which may lead to an increase in variance of the final fit.

1.3.2 Model Robust Regression 1 (MRR1)

Model Robust Regression 1 (MRR1) was developed by Einsporn and Birch (1993) to address the problems found in PLR. This technique uses a mixing parameter, \( \lambda \), to combine
the parametric and nonparametric fits to the raw data. Fitted values from MRR1 can be represented by:

\[
\hat{y}_{(MRR1)} = \lambda \hat{y}_{(LLR)} + (1 - \lambda) \hat{y}_{(OLS)},
\]

(1.44)

where \( \hat{y}_{(LLR)} \) are the fitted values obtained using local linear regression, \( \hat{y}_{(OLS)} \) are the ordinary least squares fitted values, and \( \lambda \) increases from 0 to 1 as the amount of misspecification in the model increases (Einsporn and Birch, 1993). If the parametric model gives an adequate fit, then the mixing parameter, \( \lambda \), should be close to 0. A \( \lambda \approx 0 \) would yield a fit primarily based on Ordinary Least Squares regression. Consequently, if the model is severely misspecified, then \( \lambda \) should be close to 1. A \( \lambda \approx 1 \) would yield a fit primarily based on the nonparametric fit, Local Linear Regression. If the model is somewhat adequate but still fails to capture the entire structure of the data, then \( \lambda \) is in the middle of the interval \([0, 1]\). This choice would allow for a portion of the nonparametric fit to enter the final fit in an attempt to capture the previously unexplained data. The purpose of \( \lambda \) is to combine the parametric and nonparametric fits in the most efficient proportions.

In terms of hat matrices, MRR1 can be expressed as

\[
\hat{y}_{(MRR1)} = \lambda H^{(LLR)} y + (1 - \lambda) H^{(OLS)} y
\]

\[
= [\lambda H^{(LLR)} + (1 - \lambda) H^{(OLS)}] y
\]

\[
= H^{(MRR1)} y.
\]

(1.45)
MRR1 combines separate fits for the nonparametric and parametric portions at each \( X_i \), selecting a value between them for the final estimate of \( y_i \). Mays et al. (2001) state that if locations exist in the data where the two fits are either both too high or both too low, then MRR1 has no way to correct for insufficient fits. When developed, this fitting technique was based on choosing the bandwidth and mixing parameter separately; the current research will investigate the selection of these two parameters simultaneously and will examine the behavior/properties of these estimates through the help of 3-dimensional plots.

Choosing \( \lambda \)

One of the problems found with the Partial Linear Regression (PLR) procedure was the increase in variance due to always including the nonparametric portion of the model when a parametric fit would be suffice (Mays et al., 2000). In addition, if too much reliance is placed on the parametric portion of the model without adequately accounting for variations in the data, the model may suffer an increase in the bias term. When \( \lambda \) is chosen too large (too much of the nonparametric fit is used) the final model has a large variance term. Consequently, when \( \lambda \) is chosen too small (not enough of the nonparametric fit is included) the final model has an increased bias term. Since the goal of choosing \( \lambda \) involves both variance and bias concerns (like the bandwidth), some criterion involving the minimization of MSE should be considered.

Pickle et al. (2008) state that the choice of mixing parameter \( \lambda \) is similar to the choice of
Derick L. Rivers  

Semiparametric Regression

bandwidth in LLR, and involves a bias-variance trade-off. In the current work, the mixing parameter will be chosen using the same criteria that was used to select the bandwidth - AVEMSE, PRESS*, and PRESS**. Mays et al. (2001) derived the following data driven expression for the asymptotically optimal value of the mixing parameter, λ, for MRR1:

\[ \hat{\lambda}_{opt}^{(MRR1)} = \frac{\langle \hat{y}_{-i}^{(LLR)} - \hat{y}_{-i}^{(OLS)}, y - \hat{y}^{(OLS)} \rangle}{\| \hat{y}_{-i}^{(LLR)} - \hat{y}^{(OLS)} \|^2}, \]  

(1.46)

where the \(i^{th}\) observations of \(\hat{y}_{-i}^{(LLR)}\) and \(\hat{y}_{-i}^{(OLS)}\) are \(\hat{y}_{i,-i}^{(LLR)}\) and \(\hat{y}_{i,-i}^{(OLS)}\), respectively. The values \(\hat{y}_{i,-i}^{(LLR)}\) and \(\hat{y}_{i,-i}^{(OLS)}\) denote the LLR and OLS estimates, respectively, obtained by leaving out the \(i^{th}\) observation when estimating at \(x_i\). The notation \(\langle \; \rangle\) represents the inner product and \(\| \; \|\) represents the standard \(L_2\) (Euclidean) norm. Mays et al. (2001) also presented the convergence properties of this estimator. This asymptotically optimal selector is not used in the current work on simultaneous choice of bandwidth and λ, but could be studied in future work.

1.3.3 Model Robust Regression 2 (MRR2)

To address the issues found with MRR1, Mays et al. (2000) developed a third semiparametric procedure called Model Robust Regression 2 (MRR2). Instead of using the raw data to determine the nonparametric portion of the fit, MRR2 uses the residuals from the parametric fit. Mays (1995) found this new technique to eliminate some of the bias problems associated with MRR1 when both the parametric and nonparametric portions are either too high or
low. The final model can expressed as:

\[ \hat{y}_{(MRR2)} = \hat{y}_{(OLS)} + \lambda \hat{r}, \tag{1.47} \]

where \( \hat{y}_{(OLS)} \) are the OLS fitted values, \( \hat{r} \) are the fitted values to the residuals from the parametric fit using LLR, and \( \lambda \) is a mixing parameter that adds back a portion of the residual fit. In terms of hat matrices, MRR2 can be expressed as

\[
\hat{y}_{(MRR2)} = H^{(OLS)}y + \lambda H^{(LLR)}r
= [H^{(OLS)} + \lambda H^{(LLR)}(I - H^{(OLS)})]y
= H^{(MRR2)}y. \tag{1.48}
\]

Pickle et al. (2008) note that MRR1 and MRR2 are similar, in that they both combine a parametric fit and a nonparametric fit by the use of a mixing parameter; however they differ in that the MRR2 nonparametric fit is to the residuals from the parametric fit. The vector of residuals (\( r \)) represents the structure in the data that is not captured by the specified parametric model. Local Linear Regression is used to nonparametrically fit the vector of residuals, and results in a vector of smoothed residuals as follow:

\[ \hat{r} = H^{(LLR)}_r r \tag{1.49} \]

where \( H^{(LLR)}_r \) is computed similarly to that of the LPR smoother matrix in equation (1.28)
but with the response variable being the residuals from the OLS fit to the raw data.

With MRR2, the smoothing parameter, $\lambda$, increases as the amount of misspecification increases, but is not necessarily proportionate like that of MRR1. In MRR2 the value of $\lambda$ indicates the amount of correction needed from the residual fit, as noted by Pickle et al. (2008). Notice that $\lambda = 1$ actually represents only a partial contribution from the residual fit to the final fit as the coefficient for the parametric fit is always 1. Mays et al. (2001) derived the data driven expression for the asymptotically optimal value of the mixing parameter, $\lambda$, for MRR2 as follows:

$$\hat{\lambda}_{opt}^{(MRR2)} = \frac{\langle \hat{r}, y - \hat{y}^{(OLS)} \rangle}{\|\hat{r}\|^2}. \quad (1.50)$$

As with MRR1, the selection criteria of AVEMSE, PRESS*, and PRESS** will be used in the current work to select the bandwidth and mixing parameter (simultaneously) for MRR2. Previous simulation studies and comparisons of the five fitting techniques (OLS, LLR, PLR, MRR1, and MRR2) indicated that overall MRR2 resulted in the best performance diagnostics and was able to maintain bias and variance properties comparable to those obtained by OLS when there was no model misspecification as well as to LLR when there existed a large amount of model misspecification (Mays and Birch, 2002).
Chapter 2

Previous Research

The previous research conducted by Mays et al. (2000) was centered around evaluating the performance of the various bandwidth selectors for several example data sets by comparing MSE values of the data-driven fits to those of the optimal fits from using the optimal bandwidth, $b_0$, and the optimal mixing parameter, $\lambda_0$. Mays (1995) compared the optimal fits for many generated data sets, while showing consistent significant potential improvement of PLR and MRR over the individual procedures of OLS and LLR. These optimal fits were initially considered when selecting $b$ and $\lambda$ separately. Similar research was done to consider these values when $b$ and $\lambda$ were chosen simultaneously (Burton, 2002). The research mentioned in this chapter focuses on whether simultaneous selection provides appropriate bandwidths and mixing parameters when compared to a selection process that chooses these values separately.
2.1 MSE Criterion

As previously mentioned, an obvious choice for comparing fitting techniques and choosing the bandwidth and mixing parameter for nonparametric and semiparametric procedures should involve some type of Mean Squared Error, \[ \text{MSE} = (\text{bias})^2 + \text{variance} \] criterion. Mays et al. (2000) derived the formulas for bias(\(\hat{y}\)) and var(\(\hat{y}\)) for each of the fitting procedures considered in the current work. These MSE formulas were used to compute both the average MSE (AVEMSE) over the specific data points and an approximate integrated MSE (INTMSE) across the entire range of the data in order to choose the bandwidth and mixing parameter and compare the fitting procedures. Mays and Birch (2002) obtained the MSE at each point in the range of the data and then these MSE values were averaged across all of the data points to obtain the AVEMSE. This AVEMSE value is then minimized to separately find the “optimal” bandwidth, \(b_0\), and “optimal” mixing parameter, \(\lambda_0\), for the fitting technique of interest. The INTMSE value is used to evaluate the overall fits of the procedures. Mays and Birch (2000) minimized the AVEMSE, using only the data points, instead of the INTMSE because this mirrors the information available for data-driven selection methods, as mentioned in Härdle (1990).

In the previous research, the bandwidth and mixing parameter for each procedure were taken to be fixed at these optimal \(b_0\) and \(\lambda_0\) and found separately. The current work expands on the findings of Mays et al. (2000) and similar work, so AVEMSE is used to determine the optimal bandwidths and mixing parameters and INTMSE is used as the main diagnostic for the comparison of fits.
However, while previous research has been considered when choosing $b$ and $\lambda$ separately versus simultaneously, of interest in the current work is if a graphical approach using 3-Dimensional plots will help us determine whether the “optimal” values truly deliver the best diagnostic (i.e. lower INTMSE, lower PRESS*, and lower PRESS** values) and provide us with information to study the properties of these parameters at values other than just the single “optimal” values.

### 2.2 Previous Results Comparing OLS, LLR, PLR, MRR1, and MRR2

As mentioned in section (1.3.2), the MRR1 procedure involves separate fits to the raw data, nonparametric (LLR) and parametric (OLS), and then adds these two fits via the mixing parameter, $\lambda$. The PLR technique described in section (1.3.1) uses a parametric fit to the partial residuals and then adds back a nonparametric fit to the residuals from the parametric fit. MRR2 as described in section (1.3.3), uses a parametric fit to the raw data and then adds back a portion of a nonparametric fit to the residuals. In the previous research, MRR2 displayed the best results as a fitting technique. This conclusion was due to the fact that the MRR1 procedure had one major disadvantage, it uses the raw data to determine the parametric and nonparametric fits. Mays et al. (2001) state that if locations exist in the data where the two fits are either both too high or both too low, then MRR1 has no way to overcome this problem. Partial Linear Regression (PLR) is able to combat this problem.
since this technique uses the residuals to determine the nonparametric portion of the model. However, if the parametric (OLS) fit is adequate, PLR results in an inflated variance term because it always involves the use of the entire nonparametric fit to the residuals.

The Model Robust Regression 2 (MRR2) technique was able to correct the aforementioned problems of MRR1 and PLR since it uses the residuals from the raw data to determine the nonparametric portion. The mixing parameter, \( \lambda \), determines the proportion of the nonparametric fit (i.e., residuals) to add back into the final model. Mays et al. (2000) found that when model misspecification did not exist, the three semiparametric techniques: PLR, MRR1, and MRR2, resulted in diagnostics similar to those of OLS. The previous research from Mays et al. (2000) also showed that when there was a significant amount of misspecification in the model, where the nonparametric technique, LLR, would be appropriate, the diagnostics showed that the three semiparametric techniques previously mentioned were as good or better as those obtained from LLR.

The main interest was models with small to moderate misspecification. Mays et al. (2000) showed that PLR, MRR1, and MRR2 all out-performed either of the parametric (OLS) and nonparametric (LLR) techniques. In conclusion, the MRR2 technique showed the most promise and the best overall diagnostics when compared to the other semiparametric techniques. The results of the above studies are based on an optimal MSE criterion. Further details can be found in Mays (1995) and Mays et al. (2000). The following section will describe the optimality criterion and give detailed diagnostics from separately versus simultaneously choosing the appropriate bandwidth and mixing parameter.


### 2.3 Previous Results Comparing PRESS* and PRESS**

The theoretical MSE formulas from research conducted by Mays et al. (2000) do not depend on the actual responses, $y's$. Which data-driven technique to use was the next concern proposed by Mays et al. (2000). In Chapter 1, sections (1.2.2.5) and (1.2.2.6), PRESS* and PRESS**, respectfully, were introduced as data-driven selectors for the bandwidth, $b$. These two techniques are also used as selectors for the mixing parameter, $\lambda$. Mays and Birch (1996) introduced this adaption to the traditional PRESS statistic, which imposes a penalty on the prediction error, $p(b)$, as shown in equation (1.19). PRESS* was developed to protect against the problems of cross validation techniques overfitting.

The PRESS* statistic favors larger values for $b$ and smaller values for $\lambda$, thereby protecting against fits that may be too variable. It was found by Mays and Birch (2002) that PRESS* often yields bandwidths that are too large. PRESS* was often minimized by selecting $b = 1$, even though this selection would not be appropriate. To correct the problem of choosing large bandwidths, Mays and Birch (2002) used a graphical approach where the value of $b$ was chosen at the point of the first local minimum or at a point where the PRESS* curve started to flatten out. When PRESS* selected what would be considered an appropriate value for $b$, it was also observed that it chose an appropriate value for $\lambda$.

As mentioned in Chapter 1, section (1.2.2.6), Mays and Birch (2002) introduced PRESS**. This new selector contains an additional penalty in the denominator of PRESS*. This added value penalizes PRESS* for choosing large values of $b$. With this added penalty, PRESS** should choose smaller values for $b$ and larger values for $\lambda$. It was observed that PRESS**
corrected the problem of choosing $b = 1$, but it also resulted in bandwidths too small since it prevented the denominator from going to 0 for very small values of $b$. To correct the problem of choosing bandwidths too small, Mays and Birch (2002) used the same graphical approach as described for PRESS* where the value of $b$ was chosen at a point where the PRESS** curve begins to level off. It was shown that PRESS** outperformed PRESS* as a data-driven selector for the bandwidth and the mixing parameter.

### 2.4 Previous Comparisons of Data sets

This section will provide comparisons of the generated data sets that were first examined by Mays (1995) and then continued by Burton (2002), when choosing the bandwidth and mixing parameter separately and similarly when choosing $b$ and $\lambda$ simultaneously. The INTMSE was used to compare the performance of PRESS* and PRESS** in the previous work. An additional goal was to see if finding $b$ and $\lambda$ simultaneously would have improvements over selecting them separately.

#### 2.4.1 Examples

The same data sets previously considered by Mays (1995) involved choosing $b$ and $\lambda$ separately and then later emphasis was placed on choosing these values simultaneously (Burton, 2002). The data were generated from three different models using OLS to determine the parametric fit and LLR for the nonparametric fit. It was possible to obtain the optimal
bandwidth and optimal mixing parameter that would minimize the AVEMSE, since the underlying model, \( f(x) \), was known. The main concern of the later research was whether there were any improvements that could be made to the theoretical MSE value when selecting \( b_0 \) and \( \lambda_0 \) simultaneously for MRR1 and MRR2. Since the fits for OLS, LLR, and PLR did not involve \( \lambda \), the diagnostics for these fits were from Mays (1995). Also of interest was whether one of the data-driven selectors, PRESS* or PRESS**, would demonstrate improvements and result in lower values for AVEMSE and INTMSE when choosing \( b_0 \) and \( \lambda_0 \) simultaneously compared to when choosing them separately.

### 2.4.1.1 Example 1

The first example was motivated by the tensile strength data found in Montgomery and Peck (1992). This data set displays a strong quadratic structure, but contains what was called a “interesting peak that would not be captured by a typical parametric fit” (Mays et al., 2000). The underlying model is of the form:

\[
y = 2(X - 5.5)^2 + 5X + 3.5\sin\left(\frac{\pi(X - 1)}{2.25}\right) + \epsilon \tag{2.1}
\]

at ten equally spaced X-values from 1 to 10, where \( \epsilon \sim N(0, 16) \). The sine function introduces a slight deviation from a quadratic model and the multiplier 3.5 may be thought of as the amount of misspecification if using a quadratic model to obtain the fit. A nonparametric fit would capture the deviations in the structure but it would fail at taking advantage of the true quadratic structure and yield a larger variance term in the fit. Mays (1995) verifies that an OLS fit yields results with a high bias term and fails to capture the model’s misspecification,
while a nonparametric fit using LLR results in a fit with a high variance. Overall, the fits derived from the MRR1, MRR2, and PLR procedures show significantly better performance diagnostics than those obtained from OLS and LLR.

The performance diagnostics used for Example 1 above include $df_{model} = \text{trace(Hat matrix)}$, Sum of Squares Error (SSE), PRESS and INTMSE, which are all desired to be small. For this data set, $b$ and $\lambda$ were chosen as the optimal $b_0$ and $\lambda_0$ that minimized AVEMSE. The premier diagnostic was INTMSE, which is based in the theoretical MSE formulas and does not depend on the particular $y$-data generated from equation (2.1), and is the average of the MSE calculations at 1000 evenly spaced locations across the range of the data, (Mays et al., 2000). Table (2.1) shows the results of the compared diagnostics when choosing $b_0$ and $\lambda_0$ separately vs simultaneously.

Table 2.1: Comparison of Example 1 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously)

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$\lambda_0$</th>
<th>$df_{model}$</th>
<th>SSE</th>
<th>PRESS</th>
<th>INTMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>3</td>
<td></td>
<td>74.24</td>
<td>135.03</td>
<td>9.42</td>
<td></td>
</tr>
<tr>
<td>LLR</td>
<td>0.115</td>
<td>6.23</td>
<td>22.63</td>
<td>198.73</td>
<td>8.67</td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.153</td>
<td>5.28</td>
<td>26.25</td>
<td>179.61</td>
<td>7.60</td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.115</td>
<td>0.503</td>
<td>4.67</td>
<td>37.97</td>
<td>111.35</td>
<td>7.66</td>
</tr>
<tr>
<td></td>
<td>0.087</td>
<td>0.445</td>
<td>5.11</td>
<td>32.83</td>
<td>119.15</td>
<td>7.44</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.152</td>
<td>0.713</td>
<td>4.61</td>
<td>36.08</td>
<td>106.75</td>
<td>7.57</td>
</tr>
<tr>
<td></td>
<td>0.099</td>
<td>0.513</td>
<td>5.09</td>
<td>32.06</td>
<td>114.30</td>
<td>7.33</td>
</tr>
</tbody>
</table>
and \( \lambda_0 \) separately versus simultaneously for the Model Robust procedures, along with the performance diagnostics for the other three fitting techniques. The values of the diagnostics using the simultaneous choices of \( b_0 \) and \( \lambda_0 \) are in bold text.

When choosing the bandwidth and mixing parameter separately or simultaneously, it was shown that MRR2 maintains its place as a superior fitting technique. When choosing the bandwidth and mixing parameter simultaneously for the MRR2 technique, the INTMSE value was lower than when chosen separately. A smaller INTMSE value was also observed for MRR1 as well. Also of interest were the lower values for \( b_0 \) and \( \lambda_0 \), where the smaller bandwidth indicates that more weight was placed on the observations close to \( x_0 \), the point of prediction. The corresponding mixing parameter prevented the variance term from being over inflated because it halted the amount of the nonparmetric portion being added back into the final model.

### 2.4.1.2 Example 2

The second example used by Mays (1995) was generated from the following model:

\[
y = 5\sin(2\pi X) + \epsilon
\]  

at twenty equally spaced \( X \)-values from 0 to 1 by 0.05, where \( \epsilon \sim N(0, 1) \). In this case if the true underlying model was unknown, a cubic model would generally be specified. Table (2.2) gives the performance diagnostics for the five fitting procedures when choosing \( b_0 \) and \( \lambda_0 \) separately compared to simultaneously. When choosing simultaneously, MRR1 and MRR2 both had smaller bandwidths and mixing parameters, while the INTMSE values for both
Table 2.2: Comparison of Example 2 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously)

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$\lambda_0$</th>
<th>$df_{model}$</th>
<th>SSE</th>
<th>PRESS</th>
<th>INTMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td></td>
<td></td>
<td>4</td>
<td>3.13</td>
<td>6.80</td>
<td>0.298</td>
</tr>
<tr>
<td>LLR</td>
<td>0.086</td>
<td></td>
<td>8.05</td>
<td>1.24</td>
<td>2.87</td>
<td>0.322</td>
</tr>
<tr>
<td>PLR</td>
<td>0.141</td>
<td></td>
<td>5.87</td>
<td>0.97</td>
<td>2.94</td>
<td>0.250</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.086</td>
<td>0.479</td>
<td>5.94</td>
<td>1.24</td>
<td>3.64</td>
<td>0.244</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.084</td>
<td>0.478</td>
<td>6.03</td>
<td>1.21</td>
<td>3.68</td>
<td>0.243</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.140</td>
<td>0.961</td>
<td>5.70</td>
<td>1.12</td>
<td>3.95</td>
<td>0.245</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.126</td>
<td>0.842</td>
<td>5.84</td>
<td>1.10</td>
<td>3.89</td>
<td>0.245</td>
</tr>
</tbody>
</table>

remained approximately the same.

2.4.1.3 Example 3

The last example used by Mays (1995) is based on the same model as Example 1, see Equation (2.1). In the case of Example 3, two observations are taken at each of the ten evenly spaced values over the interval $X = [1, 10]$. This is done under the assumption that two observations at each point should provide added information pertaining to the “true” model, therefore resulting in better fits for the fitting techniques: OLS, LLR, PLR, MRR1, and MRR2. It was found by Mays (1995) that the overall fits from the semiparametric techniques surpass those of the individual parametric and nonparametric fits of OLS and LLR. The lower INTMSE
values for the MRR2 and PLR procedures are results of taking two observations at each
data point (Mays, 1995). The diagnostics for the five fitting techniques when choosing the
bandwidth and mixing parameter separately versus simultaneously are provided in Table
(2.3). Note again that $b_0$ and $\lambda_0$ are smaller when choosing them simultaneously versus
separately. Also, three of the performance diagnostics; SSE, PRESS, and INTMSE, are
noticeably lower for both the MRR1 and MRR2 procedures when $b_0$ and $\lambda_0$ are chosen
simultaneously. Initial comparisons of the INTMSE values for the three examples point
out that choosing the bandwidth and mixing parameter simultaneously results in smaller
bandwidths, while the smaller mixing parameter, $\lambda$, prevents an increase in the variance term.
It is concluded that key performance diagnostics for the fitting techniques can be improved
or further minimized when selecting the bandwidth and mixing parameter simultaneously.
Table 2.3: Comparison of Example 3 (Choosing $b_0$ and $\lambda_0$ Separately vs Simultaneously)

<table>
<thead>
<tr>
<th></th>
<th>$b_0$</th>
<th>$\lambda_0$</th>
<th>$df_{model}$</th>
<th>SSE</th>
<th>PRESS</th>
<th>INTMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>3</td>
<td></td>
<td></td>
<td>205.44</td>
<td>264.62</td>
<td>7.37</td>
</tr>
<tr>
<td>LLR</td>
<td>0.099</td>
<td>7.08</td>
<td>96.41</td>
<td>211.99</td>
<td>4.96</td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.118</td>
<td>6.27</td>
<td>107.93</td>
<td>207.1</td>
<td>4.4</td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.099</td>
<td>0.686</td>
<td>5.8</td>
<td>116.14</td>
<td>212.08</td>
<td>4.69</td>
</tr>
<tr>
<td></td>
<td><strong>0.08</strong></td>
<td><strong>0.613</strong></td>
<td><strong>6.18</strong></td>
<td><strong>102.2</strong></td>
<td><strong>200.53</strong></td>
<td><strong>4.57</strong></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.119</td>
<td>0.879</td>
<td>5.83</td>
<td>115.83</td>
<td>209.83</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td><strong>0.097</strong></td>
<td><strong>0.742</strong></td>
<td><strong>6.1</strong></td>
<td><strong>106.16</strong></td>
<td><strong>202.53</strong></td>
<td><strong>4.38</strong></td>
</tr>
</tbody>
</table>

2.4.2 Selection Criterion Study

The intent of this section is to determine if the data-driven selectors, PRESS* or PRESS**, would perform better when the bandwidth and mixing parameters are chosen simultaneously or separately. To evaluate the performance of the selectors, five data sets were considered. Data set 1 is the same as the data used in Example 1, in section (2.4.1.1). Data set 2 is actually Data set 1 without the error term, ($\epsilon \sim N(0, 16)$). For Data set 2 the error term is $\epsilon \sim N(0, 1)$ when necessary for certain calculations, but the raw data is from the underlying model without the error term. Data set 3 is the same as Example 3, recalling that Example 3 is actually Example 1 with two observations taken at each $X$-value, as in section (2.4.1.3). Data set 4 is from Example 2 from section (2.4.1.2). The last data set is generated from the
model:

\[ y = 2(10X - 5.5)^2 + 50X + 3.5\sin(4\pi X) + \epsilon \]  

(2.3)

at twenty-one evenly spaced \( X \) values on the interval \([0, 1]\), and \( \epsilon \sim N(0, 1) \).

### 2.4.2.1 Theoretically Optimal Bandwidth and Mixing Parameter

Table (2.4) shows the optimal AVEMSE and INTMSE values for the five data sets used by Mays et al. (2000) for each of the fitting techniques of interest. The given results were achieved by choosing the bandwidth and mixing parameter separately and then these fitting techniques were run while choosing them simultaneously. Notice that the AVEMSE\(_0\) and INTMSE\(_0\) values are very similar where the INTMSE\(_0\) values are generally lower. The optimal AVEMSE and INTMSE values obtained when choosing simultaneously are as good as, or better, than those obtained when the bandwidth and mixing parameter are chosen separately. Recall that the AVEMSE was found by averaging the MSE’s of the fitted values at specific points in the data, thereby creating a MSE criterion that is totally dependent on the given data. Mays et al. (2000) provide the derivations of the MSE formulas used for finding \( b_0 \) and \( \lambda_0 \). Both the AVEMSE and INTMSE sufficiently estimate the true MSE value, and performance of the fitting techniques can be determined regardless of which criterion is used. The computations of the AVEMSE value rely strictly upon the data points, which gives it the ability to provide a better comparison for the data-driven criterion, PRESS* or PRESS**, since they too also rely on the data points. Since, \( b_0 \) and \( \lambda_0 \) are both chosen by minimizing the AVEMSE, it is possible that the optimal INTMSE value calculated is not
the minimal INTMSE possible.

Table 2.4: Comparison of Optimal AVEMSE and INTMSE (Separately vs Simultaneously)

<table>
<thead>
<tr>
<th></th>
<th>Separate AVEMSE</th>
<th>Separate INTMSE</th>
<th>Simultaneous AVEMSE</th>
<th>Simultaneous INTMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LLR</td>
<td>9.848</td>
<td>9.848</td>
<td>8.672</td>
<td>8.672</td>
</tr>
<tr>
<td>MRR1</td>
<td>8.341</td>
<td><strong>8.147</strong></td>
<td>7.658</td>
<td><strong>7.435</strong></td>
</tr>
<tr>
<td>MRR2</td>
<td>8.386</td>
<td><strong>8.091</strong></td>
<td>7.573</td>
<td><strong>7.329</strong></td>
</tr>
<tr>
<td>PLR</td>
<td>8.711</td>
<td>8.711</td>
<td>7.604</td>
<td>7.604</td>
</tr>
<tr>
<td><strong>Data 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>5.385</td>
<td>5.385</td>
<td>5.585</td>
<td>5.585</td>
</tr>
<tr>
<td>LLR</td>
<td>0.922</td>
<td>0.922</td>
<td>0.986</td>
<td>0.986</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.912</td>
<td><strong>0.905</strong></td>
<td>1.012</td>
<td>1.023</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.873</td>
<td><strong>0.872</strong></td>
<td>0.845</td>
<td><strong>0.844</strong></td>
</tr>
<tr>
<td>PLR</td>
<td>0.875</td>
<td>0.875</td>
<td>0.848</td>
<td>0.848</td>
</tr>
<tr>
<td><strong>Data 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>7.485</td>
<td>7.485</td>
<td>7.373</td>
<td>7.373</td>
</tr>
<tr>
<td>LLR</td>
<td>5.575</td>
<td>5.575</td>
<td>4.964</td>
<td>4.964</td>
</tr>
<tr>
<td>MRR1</td>
<td>5.066</td>
<td><strong>4.945</strong></td>
<td>4.693</td>
<td><strong>4.566</strong></td>
</tr>
<tr>
<td>MRR2</td>
<td>4.909</td>
<td><strong>4.842</strong></td>
<td>4.415</td>
<td><strong>4.381</strong></td>
</tr>
<tr>
<td>PLR</td>
<td>4.997</td>
<td>4.997</td>
<td>4.404</td>
<td>4.404</td>
</tr>
<tr>
<td><strong>Data 4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.340</td>
<td>0.340</td>
<td>0.298</td>
<td>0.298</td>
</tr>
<tr>
<td>LLR</td>
<td>0.352</td>
<td>0.352</td>
<td>0.322</td>
<td>0.322</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.271</td>
<td>0.271</td>
<td>0.244</td>
<td><strong>0.243</strong></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.277</td>
<td>0.277</td>
<td>0.245</td>
<td>0.245</td>
</tr>
<tr>
<td>PLR</td>
<td>0.281</td>
<td>0.281</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td><strong>Data 5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>5.003</td>
<td>5.003</td>
<td>4.786</td>
<td>4.786</td>
</tr>
<tr>
<td>LLR</td>
<td>0.566</td>
<td>0.565</td>
<td>0.527</td>
<td>0.526</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.566</td>
<td><strong>0.565</strong></td>
<td>0.527</td>
<td><strong>0.526</strong></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.503</td>
<td>0.503</td>
<td>0.467</td>
<td><strong>0.466</strong></td>
</tr>
<tr>
<td>PLR</td>
<td>0.487</td>
<td>0.487</td>
<td>0.451</td>
<td>0.451</td>
</tr>
</tbody>
</table>
2.4.2.2 PRESS* Results

The results in Table (2.5) show the results for the five data sets using PRESS* as the selector for the bandwidth and mixing parameter. The AVEMSE values when choosing separately and simultaneously were seemingly close to the optimal value, but there was some room for improvement. For Data set 1, the simultaneous AVEMSE\(_0\) value for the MRR2 technique was smaller than the AVEMSE\(_0\) value when \(b\) and \(\lambda\) were chosen separately. In conclusion, it was noted that PRESS* continued to perform well when choosing simultaneously on those data sets where it performed well when choosing separately, as seen in Mays et al. (2000). Simultaneous selection of the bandwidth and mixing parameter showed some improvement in the AVEMSE value for the majority of the data sets. It was found that more appropriate bandwidths and mixing parameters were chosen as compared to choosing them separately. There were however, a few instances (Data set 2 in particular) where PRESS* resulted in fits that were much worse than optimal (whether \(b\) and \(\lambda\) were chosen separately or simultaneously).
Table 2.5: Comparison of Fits When Using PRESS*

Optimal Values in Bold

<table>
<thead>
<tr>
<th></th>
<th>Separate</th>
<th>Simultaneous</th>
<th>Separate</th>
<th>Simultaneous</th>
<th>Separate</th>
<th>Simultaneous</th>
<th>AVEMSE</th>
<th>AVEMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td>9.885</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>9.848</td>
<td></td>
</tr>
<tr>
<td>LLR</td>
<td>11.297</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.158</td>
<td>1</td>
<td>0.298</td>
<td>0.032</td>
<td>9.145</td>
<td>9.557</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.36</td>
<td>0.191</td>
<td>0.288</td>
<td>0.394</td>
<td>9.82</td>
<td>9.898</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.152</td>
<td>0.099</td>
<td>0.713</td>
<td>0.513</td>
<td>8.386</td>
<td>8.091</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.472</td>
<td></td>
<td></td>
<td></td>
<td>9.866</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.153</td>
<td></td>
<td></td>
<td></td>
<td>8.711</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>5.385</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.365</td>
<td>5.386</td>
</tr>
<tr>
<td>LLR</td>
<td>3.292</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.922</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.133</td>
<td>0.74</td>
<td>0</td>
<td>0.004</td>
<td>5.385</td>
<td>5.385</td>
<td>5.385</td>
<td>5.386</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.065</td>
<td>0.059</td>
<td>0.057</td>
<td>0.922</td>
<td>0.912</td>
<td>0.905</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.155</td>
<td></td>
<td></td>
<td></td>
<td>2.286</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.073</td>
<td></td>
<td></td>
<td></td>
<td>0.875</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>7.485</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.695</td>
<td>5.575</td>
</tr>
<tr>
<td>LLR</td>
<td>3.530</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.088</td>
<td>0.067</td>
<td>0.566</td>
<td>0.554</td>
<td>5.009</td>
<td>4.983</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.099</td>
<td>0.08</td>
<td>0.086</td>
<td>0.813</td>
<td>5.066</td>
<td>4.945</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.155</td>
<td></td>
<td></td>
<td></td>
<td>5.285</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.119</td>
<td></td>
<td></td>
<td></td>
<td>4.997</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.471</td>
<td>0.352</td>
</tr>
<tr>
<td>LLR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.049</td>
<td>0.05</td>
<td>0.996</td>
<td>0.995</td>
<td>0.468</td>
<td>0.463</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.086</td>
<td>0.084</td>
<td>0.479</td>
<td>0.478</td>
<td>0.271</td>
<td>0.271</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.141</td>
<td></td>
<td></td>
<td></td>
<td>0.281</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.058</td>
<td></td>
<td></td>
<td></td>
<td>0.418</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>5.003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.066</td>
<td>5.066</td>
</tr>
<tr>
<td>LLR</td>
<td>0.566</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR1</td>
<td>0.051</td>
<td>0.051</td>
<td>1</td>
<td>1</td>
<td>0.566</td>
<td>0.565</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MRR2</td>
<td>0.052</td>
<td>0.052</td>
<td>1</td>
<td>1</td>
<td>0.566</td>
<td>0.565</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLR</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
<td>0.51</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td>0.487</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


2.4.2.3 PRESS** Results

Table (2.6) shows the results for the five data sets using PRESS** as the selector for the bandwidth and mixing parameter, including AVEMSE values for the data sets when choosing separately compared to choosing $b$ and $\lambda$ simultaneously. When choosing the bandwidth and mixing parameter simultaneously it is possible to obtain lower AVEMSE values, but PRESS** only maintained these improvements with Data sets 2 and 5. Both the MRR1 and MRR2 procedures resulted in appropriate values for $b$ and $\lambda$ for Data set 2 using the simultaneous selection process (where PRESS* yielded very poor fits). The simultaneous selection process for Data sets 1 and 3 yielded smaller bandwidths and larger mixing parameter values as well as slightly higher AVEMSE values.

In conclusion, slightly smaller bandwidths and larger mixing parameter seem to be the trend for PRESS**. These results would yield a final fit that would include more of the nonparametric fit as well as possibly overfitting the data. Looking at all of the examples overall, however, PRESS** results in fits that maintain the benefits of the model robust procedure (MRR2 in particular). This includes both cases of selecting $b$ and $\lambda$ separately or simultaneously. Even for the worst performance (Data set 1, with simultaneous selection) the diagnostics are not greatly different from those of the optimal fits. This could not be said about PRESS*. One goal of the current research is to investigate in more detail why we are seeing some of these trends in the fits.
Table 2.6: Comparison of Fits When Using PRESS**

<table>
<thead>
<tr>
<th></th>
<th>Data 1</th>
<th>Data 2</th>
<th>Data 3</th>
<th>Data 4</th>
<th>Data 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Separate</td>
<td>Simultaneous</td>
<td>Separate</td>
<td>Simultaneous</td>
<td>Separate</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>b</td>
<td>Lambda</td>
<td>Lambda</td>
<td>AVEMSE</td>
</tr>
<tr>
<td>OLS</td>
<td>5.885</td>
<td></td>
<td></td>
<td></td>
<td>5.385</td>
</tr>
<tr>
<td>LLR</td>
<td>0.668</td>
<td></td>
<td></td>
<td></td>
<td>0.922</td>
</tr>
<tr>
<td>MRR1</td>
<td>0.12</td>
<td>0.105</td>
<td>0.08</td>
<td>0.08</td>
<td>0.099</td>
</tr>
<tr>
<td>MRR2</td>
<td>0.155</td>
<td>0.145</td>
<td>0.191</td>
<td>0.081</td>
<td>0.119</td>
</tr>
<tr>
<td>PLR</td>
<td>0.153</td>
<td>0.152</td>
<td>0.099</td>
<td>0.713</td>
<td>0.152</td>
</tr>
</tbody>
</table>

Optimal Values in Bold
2.4.2.4 PRESS* versus PRESS** Summary

From the previous research of the five fitting techniques, Mays et al. (2000) indicate that MRR2 performed the best. PRESS** is the data-driven technique that yields the most appropriate choices for the bandwidth and mixing parameter when choosing them separately. PRESS** also yielded smaller values for AVEMSE. MRR2 still maintains its top position over the other four fitting techniques when choosing the bandwidth and mixing parameter simultaneously. Data sets 3, 4, and 5 reveal that PRESS* resulted in the most appropriate bandwidths and mixing parameters as well as lower AVEMSE values for both the MRR1 and MRR2 techniques. PRESS** outperformed PRESS* for Data set 2 in the separate selection process as well the simultaneous selection process for choosing $b$ and $\lambda$. Since, PRESS** tends to combine small bandwiths with large mixing parameters, the hypothesized advantages exhibited by Mays et al. (2000) may not be as evident when choosing $b$ and $\lambda$ simultaneously.

The remainder of this work presents a detailed study of the properties of $b$ and $\lambda$ to determine what may be leading to some of these inconsistent results. Focus will be on the performance of all $b$ and $\lambda$ combinations when choosing optimal values for the parameters (using AVEMSE), and observing if the same trends are evident when using the data-driven selectors of PRESS* and PRESS**. Also of interest is to determine if fit diagnostics can be made more consistent by focusing more on the selection of the bandwidth or on the selection of the mixing parameter (or both).
Chapter 3

Graphical Study

3.1 Graphical Approach of Examples, and AVEMSE Results

The purpose of this chapter is to provide computer generated 3-D plots that were created using PROC IML in the SAS (Release 9.2) programming language (SAS Institute Inc. 2008) that will assist in analyzing the behavior of the selection process for choosing $\lambda$ and $b$ simultaneously. The analysis is only based upon comparing MRR1 and MRR2 since these two regression procedures require the use of a bandwidth and a mixing parameter for scatterplot smoothing. Each procedure will be analyzed by using the previously mentioned examples from Chapter 2. An attempt will be made to answer the question of which is more important, the choice of the bandwidth or the choice of $\lambda$?
One would suspect that these two semiparametric techniques would handle model mis-specification quite similarly, but would achieve this differently because of the the idea of the mixing parameter. This suspicion may arise due to the fact that $\lambda$ plays different roles in each of the techniques. The purpose of $\lambda$ in the MRR1 procedure is to find the most efficient proportion of the nonparametric and parametric fits. If the mixing parameter is too small we would expect to see an increase in the bias and if the mixing parameter is too large there will exist an increase in the variance. MRR2 uses $\lambda$ as a means of fixing what the parametric fit misses. With MRR2 the value of $\lambda$ should increase as the model misspecification increases, but is not proportionate like that of the MRR1 technique. The value of the mixing parameter represents the amount of correction needed from the residual fit.

Both MRR1 and MRR2 require the use of the bandwidth $b$ to control the amount of weight given to each observation. In the case of MRR1, the bandwidth is used for fitting the observations by way of LLR, where as MRR2 uses this smoothing parameter to fit the residuals. Previous research by Rice (1984) and Chiu (1991) found that the choice of $b$ was critical for nonparametric estimation. The choice of $b$ must be a compromise between variance and bias, because if the bandwidth is too large the result will be a curve that is underfit, thus a high bias term. Inversely, if the value of $b$ is too small the resulting curve will be overfit, thus an increase in the variance term. As mentioned in sections (1.2.2) and (2.1), the MSE is a natural criterion for this selection process. In the current work, the AVEMSE will be the criterion for selecting the optimal bandwidth and the mixing parameter. The use of
the AVEMSE value in the 3-D plots should allow us to investigate the trade-off between the bias and the variance for both $b$ and $\lambda$, as well as how $\lambda$ is used to correct the model misspecifications.

The 3-D plots will allow us to answer several questions about how the optimal choices of $b$ and $\lambda$ are essential in obtaining the optimal AVEMSE. The graphical study will be used to analyze how changing the bandwidth or changing $\lambda$ will have an effect on the surface (i.e., what causes the most change or instability in the plot?). An analysis of the plots will allow us to observe whether changes are drastic or gradual for different values of $b$ or $\lambda$. For the wrong value of $b$, is there any way $\lambda$ can give a “reasonable” fit or is the wrong bandwidth a disaster? For the “correct” value of $b$, how much impact does changing $\lambda$ have? In section (3.2) we will compare the structure from using AVEMSE to the structure from using PRESS* or PRESS**. This comparison will afford us the opportunity to investigate if the consequences of choosing an inappropriate bandwidth or mixing parameter is more extreme with one procedure than the other.

### 3.1.1 Summary of Example 1

Example 1 was generated from the underlying model of

$$y = 2(X - 5.5)^2 + 5X + 3.5\sin\left(\frac{\pi(X - 1)}{2.25}\right) + \epsilon$$

at ten equally spaced $X$-values from 1 to 10, where $\epsilon \sim N(0, 16)$. In the 3-D plot of AVEMSE vs $\lambda \times b$ in Figure 3.1 we use MRR1 to fit the model. We notice that where $\lambda$ is considered
to be small and $b$ increases to 1, this surface is flat because when $\lambda = 0$ we are only using OLS to fit the data. For large values of $b$ the surface shows more change, showing that we should use a small $\lambda$ to avoid an oversmoothed fit.

The OLS fit in this case would typically be quadratic since we are attempting to fit an underlying quadratic model with a sine function added to it. We also notice that when the value of $b$ is near 0 and $\lambda = 1$ there is a slight curl in the graph. This upturned portion where $\lambda$ is near 1 and $b$ is small is accredited to the nonparametric LLR fit and represents the increase in the variance, where the desired curve is overfitted. Small incremental increases in the bandwidth value seem to decrease the amount of variance and levels out the curve, but if we overstep the optimal bandwidth value then $\lambda$ has no way of correcting this other than taking on the value of 0 to give the OLS fit.

The AVEMSE is at its optimal value of 8.15 when $b_0 = 0.087$ and $\lambda_0 = 0.445$. The final model has $df_{model} = 5.11$ and a PRESS of 119.15. If we choose $b$ near optimal, then it appears
that the choice of $\lambda$ is not crucial; a poor choice of $b$ means that we must be very careful in choosing $\lambda$. For MRR1, we could always choose $\lambda = 0$ to give the OLS fit, but would lose any benefits from the model robust procedure.

The MRR2 procedure is also used to fit the data for a comparison of the simultaneous selection process. In Figure 3.2 we see that when the value of $b$ is small we have severe amounts of curvature along the axis for $\lambda$. AVEMSE is imposing a penalty for small values for the choice of $b$. When $\lambda = 0$, the surface along the bandwidth axis represents the OLS fit to the data. The choice of $b$ is key in this example; if the bandwidth is chosen too small the MSE will be very large unless the amount of the residual fit that is added back is close to zero (giving OLS). When the bandwidth is large, this results in a flat surface that may represent an OLS fit to the data. When $b$ is close to the optimal, there is a wide range of choices of $\lambda$ that provide a much lower MSE than OLS.

The simultaneous selection process resulted in $b_0 = 0.099$ and $\lambda_0 = 0.513$. In Figure 3.2
the derived AVEMSE₀ value of 8.09 exists underneath the upturned portion. The value of λ₀ indicates the amount of the residual fit that is added back into the model to correct the misspecification. As suspected, MRR2 has a lower \( df_{model} \) and PRESS than that of MRR1. These values are 5.09 and 114.3 respectfully. MRR2 generally delivers a flat surface that describes the worst case scenario of quadratic OLS fit to the data. MRR2 had drastic changes around small values of \( b \), but even the peak is not as high as many of the AVEMSE values for MRR1 (see Figure 3.1, and note that the values on the “MSE” axis are very different for the two figures). There is a smoother transition for MRR1, but most of the surface is much higher. If we choose \( b \) near optimal, then it appears that the choice of λ is not crucial; a poor choice of \( b \) means that we must be very careful in choosing λ.

### 3.1.2 Summary of Example 2

Example 2 mentioned in section (2.4.1.2) was generated from the following model:

\[
y = 5\sin(2\pi X) + \epsilon
\]

at twenty equally spaced \( X \)-values from 0 to 1 by 0.05, where \( \epsilon \sim N(0, 1) \). A cubic model would generally be specified for the underlying model.

When using the MRR1 procedure to fit the data generated in Example 2, we should note the flat surface along the bandwidth axis with no signs of curvature. As identified in Figure 3.3, when the mixing parameter, λ, is near 0 for any choice of \( b \), this describes the behavior of an OLS fit to the data. In Figure 3.3b we see that for small values of \( b \) the AVEMSE
is considerably small. The slight curl for the very small bandwidth as $\lambda$ approaches 1 is a

penalty imposed by AVEMSE for too small values of $b$. From the viewpoint of Figure 3.3c we see that as $\lambda$ approaches 1 (i.e. model misspecification increases and we use more of the LLR fit) the AVEMSE increases quite rapidly, especially if the value of $b$ is too large.

When selecting the values for $b$ and $\lambda$ simultaneously, the AVEMSE is 0.269 with $b_0 = 0.084$ and $\lambda_0 = 0.478$. This choice of the bandwidth value confirms the golden section search range [0.05, 0.30] for the bandwidth as stated by Waterman (2002). When values for $\lambda$ are too small there is not enough nonparametric fit to the data. The optimal choice for both the value of $b_0$ and $\lambda_0$ are underneath the curl in Figure 3.3b, where the bandwidth is small and there is a symmetrical mix of OLS and LLR fits to the data.

MRR2 handles the data generated in example 2 quite differently. We immediately notice that most of the surface in Figure 3.4 is lower than that of the MRR1 surface in Figure 3.3. MRR1 and MRR2 as suspected depend upon different values of $b$, because MRR1 smooths
responses and MRR2 smooths residuals. Using the MRR1 procedure $b_0 = 0.084$, but with

due to the difference in procedures we see a visibly larger flat surface for MRR2 than that of MRR1. This noncurved region is consistent with the OLS fit to the data, which is the worst case scenario when using the MRR2 procedure. In Figure 3.4 viewpoint (a) we see that the addition of the residuals only has an effect on the fit if the bandwidth is small. If the value for $b$ is mischosen in the beginning, the LLR fit to the residuals has no way of correcting itself or possibly minimizing the MSE any better than OLS alone. The simultaneous selection process resulted in a AVEMSE$_0$ value of 0.273 with $\lambda_0 = 0.842$. In Figures 3.4 viewpoints (b) and (c) we notice a slight dip in the graph, this would signal us that we are in the area of the optimal bandwidth. Comparing MRR2 to MRR1, we can say that MRR1 is really bad for large values of $b$ if a large $\lambda$ is selected; this does not happen for MRR2 (large $\lambda$ and large $b$ still give basically an OLS fit).
3.1.3 Summary of Example 3

The last example used is based on the same model as example 1, given in Equation (2.1). In the case of Example 3, two observations are taken at each of the ten evenly spaced values over the interval \(X = [1, 10]\). Using MRR1, the larger sample results in a better fit than that of Example 1. In Example 1 the AVEMSE\(_0\) value was 8.15 with the values of \(b_0\) and \(\lambda_0\) as 0.087 and 0.445, respectfully. With the larger sample size, the AVEMSE\(_0\) is 4.95 and the values for \(b_0\) and \(\lambda_0\) are 0.08 and 0.613, respectively. Even with the larger sample size, we see in Figure 3.5 that the graph is very similar to that of the Example 1 plot (see Figure 3.1).

Using MRR2, the larger sample results yield plots that are noticeably different than that of Example 1 although the overall structure of the plots is similar. As with MRR1, we would suspect a better fit to the data when the number of observations are increased. In Figure 3.6 we see more dramatic curvature, but more chances for improvement (AVEMSE dips much
There is more importance on $\lambda$ to maintain these improvements. In example 1 the AVEMSE$_0$ value was 8.09 with the values of $b_0$ and $\lambda_0$ as 0.099 and 0.513, respectively. With the larger sample size the values for $b_0$ and $\lambda_0$ are 0.097 and 0.742 with a AVEMSE$_0$ of 4.84, resulting in a fit that is better, but more of the residual fit is required to achieve this. In Figure 3.6 we see that is essential to choose a value of $b$ close to optimal; if not, $\lambda$ has no way of improving the fit.

3.1.4 Summary of AVEMSE Examples

The 3-D plots have provided us the ability to identify several key points about the behaviors of MRR1 and MRR2 in the three examples. The intent of the previous examples was to demonstrate how the two techniques behave in search of the optimal values, $b_0$ and $\lambda_0$. There are definitely some observable differences in the overall surface structure (i.e. behavior) of the techniques. Several comparisons about the two techniques will be made in order to
clearly understand correct behavior as well as future expectations.

With the MRR1 technique we continuously notice a wave-like formation in the plotted surface. This formation explains how the range of AVEMSE values changes as $\lambda$ is increased for all possible values of $b$. This wave-like surface gives us a visual impression of how MRR1 converges from an OLS fit ($\lambda = 0$) to a LLR fit ($\lambda = 1$). When the simultaneous selection process does not find the optimal fit to be the OLS fit, this would indicate that an addition of the LLR fit is required to improve the fit of the model. Since the nonparametric fit to the data is needed, then the choice of $b$ becomes critical. Remember that when $b$ is close to 0 there is an increase in the variance term, whereas on the other extreme, when $b$ is close to 1 there is an inflation of the bias term. We clearly understand that the optimal value for $b$ must deliver a compromise between the two.

From observation of viewpoint (b) in each of the MRR1 examples, we see that there is a small range of acceptable values for $b$, once again reinforcing the fact that the selection of $b$ is vital in the nonparametric fitting process. The MRR1 techniques found the optimal values for $b$ in the range of $[0.08, 0.087]$. Once the appropriate bandwidth is selected, then the mixing parameter controls the amount of the LLR fit that is added to the OLS fit. Noting that if $b$ is chosen greater than the acceptable region, then the use of $\lambda$ will be ineffective in improving the fit. It is easy to see that if $b$ is large, as the nonparametric fit is added to the model, the range of AVEMSE values is seriously inflated beyond that of the OLS fit alone.

The MRR2 technique gives a very different picture in terms of how the use of $\lambda$ becomes important. The majority of the MRR2 surface resembles the OLS fit to the data. We
remember that the MRR2 technique simply handles model misspecification by using the mixing parameter to control the amount of a LLR fit to the residuals added back to the OLS fit to the data. When the addition of the residual fit is required, the optimal bandwidth fell in the interval of \([0.097, 0.126]\), well larger and wider than that of the MRR1 procedure. If the bandwidth is chosen greater than the optimal value, we see that \(\lambda\) will be unable to improve the model, but on the other hand it will not over-inflate the AVEMSE like that of MRR1. Notably, the range of AVEMSE values is much smaller when using the MRR2 technique than those for the MRR1 technique. The AVEMSE values are only over-exaggerated when \(b\) is too small (i.e. below optimal range) and the mixing parameter increases toward 1, implying that the variance term is escalating. Overall, we find in the examples that MRR2 outperforms MRR1 for a couple of reasons. First, we discover that there will generally be a much more acceptable range for \(b\) around \(b_0\) when using MRR2. Secondly, it is easy to see that the range of AVEMSE values for MRR2 will be smaller than those for MRR1, indicating that as long as \(b\) is in an acceptable range, \(\lambda\) has more of a chance of improving the model, unlike with MRR1. Lastly, it is interesting (and comforting) to note that the AVEMSE plots for MRR1 follow very similar patterns for each of the three examples; this is also the case for the three AVEMSE plots for MRR2.
3.2 Graphical Selection Criterion Study

In the previous section we observed how the fitting techniques, MRR1 and MRR2, behaved during the simultaneous selection of the optimal choices of \( b \) and \( \lambda \) when found by minimizing the AVEMSE. The purpose of this section is to provide 3-D plots that describe how the data-driven selectors, PRESS* or PRESS**, behave when the bandwidth and mixing parameters are chosen simultaneously. As mentioned in sections (1.2.2.5) and (1.2.2.6), both PRESS* and PRESS** were possible candidates for bandwidth and mixing parameter selection. Both of these data-driven selectors have possible areas of concern due to their imposing of penalties. In previous research by Mays et al. (2001), it was found that PRESS* tends to choose large values for \( b \) and small values for \( \lambda \), where PRESS** does the complete opposite, it chooses small values of \( b \) and large values for \( \lambda \).

To evaluate the performance of the selectors, the same five data sets that were considered in section (2.4.2) are used. Remember that data set 1 is the same as the data used in Example 1 in section (2.4.1.1). Data set 2 is actually data set 1 without the error term. Data set 3 is the same as Example 3 in section (2.4.1.3); recall that Example 3 is actually Example 1 with two observations taken at each \( X \)-value. Data set 4 is from Example 2 from section (2.4.1.2). The last data set for this study is generated from the model

\[
y = 2(10X - 5.5)^2 + 50X + 3.5 \sin(4\pi X) + \epsilon
\]

at twenty-one evenly spaced \( X \) values on the interval \([0, 1]\) and \( \epsilon \sim N(0, 1) \).
3.2.1 Graphical PRESS* Results

As mentioned in section (1.2.2.5), the PRESS* statistic is denoted as

\[
PRESS^* = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i,-i})^2}{n - tr(H^{(ker)})} = \frac{PRESS}{n - tr(H^{(ker)})},
\]

Since the denominator of PRESS* imposes a penalty for choices of small bandwidths, within the 3-D plots we would expect to see a slight difference in the values for \( b \) and \( \lambda \), where the results for these parameters may be on opposite ends of the interval from 0 to 1 as compared to the AVEMSE examples. Remember in the previous examples, the AVEMSE was minimized in order to obtain the values for \( b \) and \( \lambda \). In each of the previous sections; (3.1.1), (3.1.2), and (3.1.3), the optimal choices tend to have small bandwidths and conservative values for the mixing parameter. In the following examples we will examine how PRESS* behaves as a data-driven selector. The comparisons will be how PRESS* results for MRR1 and MRR2 compare to the optimal results of the AVEMSE.

3.2.1.1 Data Set 1 (Same as Example 1 in section 3.1.1)

In data set 1 we see that when using PRESS* for selecting \( b \) and \( \lambda \) we get a very similar type surface as generated for AVEMSE example. PRESS* also holds true to its claim for choosing large bandwidths coupled with small values for \( \lambda \). In Figure 3.7 the flat area of the surface along the axis for \( b \) when \( \lambda = 0 \) indicates an OLS fit to the data. The raised section of the surface when \( \lambda = 1 \) represents the nonparametric, LLR, fit to the data. In this area the curled portion represents the increase in the variance term due to small choice of \( b \) and
the large $\lambda$, this curl reaches a PRESS* of 152. Initially one may suggest that PRESS* is minimized where $b$ is small and a significant amount of the nonparametric fit is added back (i.e. large mixing parameter) due to the fact that the overall structure of the surfaces appear to be very similar. Evidence contrary to that belief is visible in Figure 3.7, viewpoint (c). In this figure we see that there may be an appropriate choice where $b$ is small and $\lambda$ is near 1. This curl in the surface is somewhat misleading, because the dip portion in the surface is not parallel to the $b-\lambda$ plane, and we remember that PRESS* penalizes for choices of bandwidth that are too small. In this figure we see that the value for $b$ associated with that dip in the surface is considered to be too small, thus a larger value must be considered. Reflecting on the use of the bandwidth, we are reminded that the bandwidth controls the amount of smoothness in the fit of the model, where small values (close to 0) for $b$ would be overfitting (increase in variance term) and large values of $b$ (close to 1) would be underfitting (increase in bias term). Since PRESS* eleviates the threat of overfitting by the penalization for small

Figure 3.7: PRESS* Plot for Data Set 1, using MRR1 as Fitting Technique
Our focus turns to the other end of the axis for choices of \( b \). PRESS* stays constant along the bandwidth axis, so the selection of \( \lambda \) becomes vital in this process.

PRESS* favors a parametric fit of OLS where incremental increases in \( \lambda \) would have devastating effects on the fit. Clearly we see that for large values of \( \lambda \), there is a possibility of PRESS* becoming seriously over-inflated. On the other hand, near the OLS fit, where the choices for \( \lambda \) are small, this does not seem to create any noticeable effects to the surface. Even if the wrong bandwidth is chosen here, a reasonable fit is dependent upon a near optimal choice of the mixing parameter. The value of \( b \) in the PRESS* results is 1 with \( \lambda = 0.032 \) as compared to the optimal results of 0.087 and 0.445, respectively. The AVEMSE\(_0\) is 8.147 where AVEMSE\(_{(PRESS^*)} = 9.957\). PRESS* seems to miss the mark in the case of fitting this model as compared to the optimal choices in example 1, where a poor choice for the bandwidth can only be corrected by selecting a near 0 value for \( \lambda \). These inappropriate values for \( b \) and \( \lambda \) may be accredited to the biased nature of the penalty imposed on PRESS (see section 1.2.2.5).

In Figure 3.8, when using PRESS* with MRR2 we see a much flatter surface except where \( b \) increases and \( \lambda \) remain small (i.e. close to 0). Not shown in the figure, is that there is a spike in the surface where \( b \) is between 0.05 and 0.09 and as \( \lambda \) is increased toward 1. This spike in the surface would represent an increase in the variance term of the model due to the fact that too much of the LLR fit of the residuals is added back to the parametric fit of the data. We notice in Figure 3.8 for situations where \( b \) is near 0.09 and the mixing parameter is large (i.e. close to 1) that the value for PRESS* is large as indicated by the spike.
As with MRR1, the choice of $b$ does not seem to be as important as for the selection of $\lambda$.

Figure 3.8: PRESS* Plot for Data Set 1, using MRR2 as Fitting Technique

As noted in the graph of MRR1 (see Figure 3.7), PRESS* prefers OLS as its choice of fits and uses very little of the LLR fit. Here the flat surface represents OLS with a small (or no) portion of the residual fit being added back in. If the bandwidth is chosen incorrectly, we see that there is a wide range of acceptable values for the mixing parameter that keep the value of PRESS* low. The results from using PRESS* indicate that a much more appropriate LLR fit is used on the residuals of the OLS fit to the data. The possible threat of damaging the fit only arises in the case where $\lambda$ is too large and $b$ is too small, creating an increase in the variance term. When using MRR2, this inflated variance raises PRESS* to a value of 72.38 compared to 152 with MRR1.

The values for $b$ and $\lambda$ are 0.191 and 0.384, respectfully, as compared to those of the optimal fit, where the chosen values for $b_0$ and $\lambda_0$ are 0.099 and 0.713, respectfully. The AVEMSE using PRESS* is 9.088 where the optimal value, AVEMSE$_0$, is 8.091. When using
PRESS* we see that a poor choice of $b$ does not seriously affect the fit of either MRR1 or MRR2, but the choice of $\lambda$ is much more important with MRR1 than with MRR2. MRR2 outperforms MRR1 for several reasons. First of all, MRR1 results in poor fits for a wide range of values for $\lambda$. Secondly, when using PRESS* with the MRR2 technique we do not get the over-inflated PRESS* for large $b$ and large $\lambda$ as we did with MRR1. Finally, even though MRR2 displays an inflated variance term, it is not as severe as the comparable spike when using MRR1, where this same inflated portion is nearly doubled in the case of MRR1 and the selected values for $b$ and $\lambda$ are much more appropriate.

3.2.1.2 Data Set 2

Data set 2 generated a surface structure similar to that represented by the optimal selection process used in Example 1 (see Figure 3.1) and that of data set 1 (see Figure 3.7). We observe the same situation with MRR1 as in data set 1 with a large bandwidth and a small value for $\lambda$. It is evident that small incremental increases in $\lambda$ have an adverse affect on the value for PRESS*, especially when $b$ is large. The absence of the error term seems to have minimal effect on how PRESS* treats the structure or behavior of the data. In Figure 3.19 PRESS* continuously places more dependence upon the value of the mixing parameter than that of the bandwidth. Evidence of this is present in Figure 3.9 by how the graph is flat along the bandwidth axis.

Once again PRESS* favors the parametric fit of OLS with a seemingly high bias term, where the value of $b$ in the PRESS* results is .74 with $\lambda = 0.004$ as compared to the
optimal results of 0.059 and 0.922, respectfully. The optimal AVEMSE is 0.905 where $AVEMSE_{(PRESS^*)}$ is 5.386. Even though the absence of the error term had no effect on the structure of the surface there was a definite effect on the outcome of the AVEMSE and PRESS*.

The graphs for data set 2 using the MRR2 technique look similar overall to that for

Figure 3.9: PRESS* Plot for Data Set 2, using MRR1 as Fitting Technique

Figure 3.10: PRESS* Plot for Data Set 2, using MRR2 as Fitting Technique
data set 1 using MRR2, where the error term is $\sim N(0, 16)$. However, in Figure 3.10 we do notice some sort of bow tie or twisting effect on the data in the area of high variability. On the other hand, the surface is flat for most of the graph indicating that OLS alone is not a bad fit. As more of the residual fit is added back to improve the fit, PRESS* requires that the curve is oversmoothed by increasing the bandwidth. Here we see that the choice of $b$ is significant as we add the nonparametric residual fit back. As mentioned in section 3.2.1.1 the combination of MRR1 when using PRESS* favors OLS as its choice of fits, but here when using MRR2 and PRESS* there is a greater contribution from the residual LLR fit. The values for $b$ and $\lambda$ are 0.74 and 0.8, respectfully, as compared to those of the optimal fit, where the chosen values are 0.072 and 1, respectfully. The AVEMSE when using PRESS* is 5.38 where AVEMSE$_0$ = 0.872. There is certainly room for improvement here because in the case of data set 2, PRESS* fails to perform well with either technique when compared to the optimal values.

3.2.1.3 Data Set 3 (Same as Example 3 in section 3.1.3)

We recall that data set 3 is the same as example 3, where example 3 is actually example 1 with two observations taken at each $X$-value, as in section (2.4.1.3). We notice that the PRESS* graph for data set 3 when using MRR1 (Figure 3.11) looks just like the AVEMSE graphs of examples 1 and 3 (See Figures 3.1 and 3.5). The very same structural traits demonstrated in the optimal selection process are visible here when using MRR1 and PRESS* with this data set. The values for $b$ and $\lambda$ when for data set 3 are 0.067 and 0.554, respectfully with
an AVEMSE of 4.983. In the simultaneous selection process used in example 3, \( b_0 = 0.08 \), \( \lambda_0 = 0.613 \), and AVEMSE$_0 = 4.945$.

The graph illustrates that PRESS* is much more effective for larger sample sizes, because PRESS* is known for choosing large values for \( b \) and small values for \( \lambda \), but here we find values near the optimal values found in example 3. In the case of data set 3, it seems that as \( n \) get larger, PRESS* exhibits similar or the same behavior as the simultaneous selection process using AVEMSE to find \( b_0 \) and \( \lambda_0 \).

The increase of observations at each data point has a similar effect on the MRR2 technique when using PRESS* as it did when \( b_0 \) and \( \lambda_0 \) were found by minimizing the AVEMSE in example 3. In Figure 3.12 we see a much flatter surface compared to Figure 3.6. This smooth surface is indicative of the OLS fit of the data. The curvature along the mixing parameter axis is not as drastic as it was in example 3 but it is exhibiting the same behavior. This would indicate that the need for the LLR residual fit is not as dire for minimizing PRESS*.

When using MRR2 the choice of \( b \) is still crucial because if \( b \) is chosen too large it is evident
that no value for $\lambda$ will improve the fit. If the correct bandwidth is selected, then changing $\lambda$ can have a significant impact on the fit of the model. From viewpoint (b) of Figure 3.12 we see that when $b$ is near optimal, $\lambda$ has a much better chance of correcting model misspecification. The AVEMSE$_0$ is 4.842 and the AVEMSE$_{(PRESS^*)}$ = 4.942 when using PRESS*. The values for $b$ is 0.072 and $\lambda = 0.587$ when using PRESS*, where $b_0 = 0.097$ and $\lambda_0 = 0.742$. PRESS* does an excellent job as a data-driven selector in both cases for MRR1 and MRR2 as long as $n$ is sufficiently large. Even though PRESS* delivers results near optimal when using MRR1, we see that the AVEMSE$_{(PRESS^*)}$ when using the MRR2 technique is lower than AVEMSE$_0$ (4.945) when using MRR1. MRR2 is shown to be superior over MRR1 for a couple of reasons. If we compare the overall surface of the two techniques it is clear to see that the range of PRESS* values is smaller for MRR2. Secondly, the threat that may arise from choosing the value for $b$ too large is not as devastating for MRR2 as it would be for MRR1.
3.2.1.4 Data Set 4 (Same as Example 2 in section 3.1.2)

We remember that data set 4 is from example 2 in section (3.1.2), which is generated from the following model:

\[ y = 5 \sin(2\pi X) + \epsilon \]

at twenty equally spaced \(X\)-values, where \(\epsilon \sim N(0, 1)\). In example 2 we concluded that the selection of the value \(b\) was essential for minimizing AVEMSE and the smoothing of the curve. If the appropriate bandwidth is chosen, then gradual incremental increases to the mixing parameter \(\lambda\) can possibly improve the fit. In Figure 3.13 we see that PRESS* exhibits the same characteristics as in Example 2 for the optimal fits (See Figure 3.3). It is easy to see that an OLS fit to the data may provide a fit that may be considered sufficient where \(\lambda\) is considered to be near 0.

The mixing parameter has a drastic effect on the fit of the data for larger values of \(b\)

![Figure 3.13: PRESS* Plot for Data Set 4, using MRR1 as Fitting Technique](image)

(Three rotations of the same graph)

as we notice the wave-like formation that increases the PRESS* values. We see that if \(b\) is
selected incorrectly, the choice of \( \lambda \) becomes very important. If \( \lambda \) is too large, we run the risk of damaging the model by adding too much of the LLR fit. In Figure 3.13, viewpoint (b), we see that small values of \( b \) are appropriate here. As long as \( b \) is near optimal, then \( \lambda \) will not have much impact on the performance of the model.

Using this data set, PRESS* is minimized where \( b = 0.121 \), \( \lambda = 0.578 \), and AVEMSE is 0.294. The resulting AVEMSE\(_{(PRESS^*)}\) is near optimal as compared to the results in example 2, where \( b_0 = 0.084 \), \( \lambda_0 = 0.478 \) and AVEMSE\(_0\) = 0.271.

In Figure 3.14 we see that when PRESS* is used with MRR2 for this data set the resulting graph is nearly the same as the MRR2 technique used in Example 2 for the optimal fits (See Figure 3.4). When performing the simultaneous process using PRESS* to find the values of \( b \) and \( \lambda \), we get the values 0.137 and 1, respectively. The optimal values are \( b_0 = 0.126 \), \( \lambda_0 = 0.842 \). The comparable AVEMSE value is 0.276, where AVEMSE\(_0\) = 0.277. PRESS* continues to stake claim to its reputation of choosing small values for \( b \) and large \( \lambda \) values.

We see in Figure 3.14 that most of the surface explains the OLS fit to the data, but there

\[ \text{Figure 3.14: PRESS* Plot for Data Set 4, using MRR2 as Fitting Technique} \]
are some noticeable reactions where $b$ is small. PRESS* is significantly improved when a portion of the residual fit is added back into the model. The bandwidth is desired to be small so that the residual fit is not oversmoothed and inflating the bias.

In Figure 3.14, viewpoints (b) and (c), we see a rise in the surface of the graph when the values of $b$ is too small. The rise in the graph is considered to be an increase in the variance term of the model. We can derive from this image that if $b$ is too small then there is no chance that adding any of the residual fit back will improve the model; in fact this choice for $b$ will render $\lambda$ useless in fitting the model. Once again it is evident that the choice of the bandwidth is critical if we hope to improve the fit of a misspecified model, especially when using MRR2. If the correct value for $b$ is chosen, then increases in the mixing parameter will gradually improve the fit, thus minimizing PRESS*. The values for the bandwidth and the mixing parameter when using the MRR2 techniques are more appropriate than those when using MRR1. Even though both regression procedures performed well under PRESS*, MRR2 provides more protection from damaging the fit of the model and offers a smaller range for the PRESS* values. This smaller range for PRESS* indicates that an increase in the variance term for MRR2 is not as crucial or damaging as it would be if the MRR1 technique was used.

3.2.1.5 Data Set 5

Data set 5 is generated from the model

$$y = 2(10X - 5.5)^2 + 50X + 3.5sin(4\pi X) + \epsilon$$
at twenty-one evenly spaced $X$ values on the interval $[0, 1]$ and $\epsilon \sim N(0, 1)$. In Figure 3.15

![Graph showing PRESS* plot for Data Set 5, using MRR1 as Fitting Technique](image)

*Figure 3.15: PRESS* Plot for Data Set 5, using MRR1 as Fitting Technique*

for MRR1, a very familiar surface structure is identified. The flat smooth surface along the bandwidth axis when $\lambda = 0$ represents an OLS fit to the data and the wave-like formation explains how the fit of the model is improved or destroyed as the nonparametric fit is added to the OLS fit. The effect of the wave is dependent upon the choice for $b$. It is clear to see that as $\lambda$ increases, it is highly possible that the value for PRESS* will also increase.

Remember that PRESS* is favorable toward small values for $b$ and large values for $\lambda$, and we see in Figure 3.15 that the choice of $b$ is quite critical. If $b$ is higher than optimal, then we must be careful with our choice of $\lambda$; on the other hand, if the correct (i.e. optimal) bandwidth is chosen, then $\lambda$ has very little negative impact on the fit.

Using PRESS*, the chosen values were $b = 0.051, \lambda = 1$ and AVEMSE is 0.565, where the optimal simultaneous selection process yields $b_0 = 0.05, \lambda_0 = 1$ and AVEMSE$_0 = 0.565$. The PRESS* procedure selected the same values as the optimal selection process, and this could be accredited to the sample size. We remember that if $n$ is sufficiently large, PRESS* selects
When PRESS* is used in combination with MRR2, we see in Figure 3.16 that the OLS fit is represented by the flat surface where PRESS* is the highest. The dipped or sloping portion in the surface shows us how the combinatory choices of small \( b \) values and \( \lambda \) are improving the fit of the model. It is quite evident that the correct or near optimal choice of \( b \) is critical. If \( b \) is chosen poorly (too large), then the additional LLR fit to the residuals could not improve the situation. In the case of choosing \( b \) too large, MRR2 does not allow the addition of the LLR residual fit to make matters any worse than they currently are by increasing PRESS*. The correct bandwidth is vitally important so that as gradual increases in \( \lambda \) can significantly improve the fit. As mentioned previously, PRESS* chooses small values for \( b \) and large values for \( \lambda \) where the selected value for \( b \) is 0.052 and the choice for \( \lambda \) is 1, with AVEMSE = 0.51. The optimal values for \( b_0 \) and \( \lambda_0 \) are 0.057 and 1, respectfully, and the AVEMSE_0 = 0.503. Just as with MRR1, PRESS* derived values close to optimal for MRR2 for this large sample size.
3.2.1.6 Graphical PRESS* Results Summary

We find that PRESS* performs well with data sets 3, 4, and 5, but rather poorly for the first two. In data sets 1 and 2 where the sample size is small ($n = 10$), PRESS* seems to be less effective as a data-driven selector, especially in the case of data set 2. Even though PRESS* performed below par for the two data sets, the graphical analysis turns out to be quite informative. The graphs allow us to answer some important questions about how MRR1 and MRR2 behave when using PRESS* as a selector for $b$ and $\lambda$.

For instance, data sets 3, 4, and 5 give us a better image of what sort of effect occurs as a result of changing the bandwidth or $\lambda$ (i.e., what causes the most change in the surface of the plot). These three data sets resemble those described in the examples used to derive the optimal values, $b_0$ and $\lambda_0$. When MRR1 is use with PRESS*, it still delivers a wave-like shape where the range of “acceptable” bandwidth values seems to be pretty narrow. The wave is created as the amount of the nonparametric fit added to the OLS fit is increased, where the height of the wave provides a look at the wide range of possible AVEMSE values associated with the range of PRESS*. If the value for $b$ is chosen too large, then $\lambda$ cannot add a “reasonable” amount of the nonparametric fit to the parametric fit and the regression is a disaster. Careful consideration must be given to the choice of $\lambda$ if $b$ is mischosen.

When using PRESS*, MRR2 displays the same sort of surface structure as seen in section (3.1) for the optimal fits. The range of the MRR2 PRESS* values is much smaller than those exhibited by the MRR1 technique. The values for the bandwidth and the mixing parameter when using MRR2 are closer to the optimal values. At first glance data set 5 looks like
a slightly different surface compared to the other four, but looking closer reveals that it actually does have a similar underlying structure. We see that the choice of bandwidth and mixing parameter can have a large impact the plot surfaces. Here the majority of the surface rests where the PRESS* values are at the maximum, but the use of a bandwidth near optimal and the complete addition of the residual fit back into the model gives significant improvements. The worst case scenario in this case of selecting a value for $b$ that is too large for MRR2 does not have the same adverse effect as selecting $b$ too large in MRR1. Thus, MRR2 surpasses MRR1 due to the fact that if the bandwidth is selected to be too large, the model is not negatively effected. MRR2 does not impose a change on the OLS fit when $b$ is too large, whereas a large choice of $b$ will affect the surface of MRR1 as $\lambda$ increases.

### 3.2.2 Graphical PRESS** Results

As mentioned in section (1.2.2.6), the PRESS** statistic is denoted as

$$PRESS^{**} = \frac{PRESS}{n - tr(H^{(ker)}) + (n - 1)\frac{SSE_{mean} - SSE_b}{SSE_{mean}}}$$

The denominator of PRESS** imposes a double penalty; one for choosing bandwidths too small and the other for choosing bandwidths that are too large. With this added penalty, PRESS** should choose smaller values for $b$ and larger values for $\lambda$ than those derived with PRESS*. The problem of choosing $b = 1$ is resolved as a result of the additional penalty. When comparing the 3-D plots generated from using PRESS** to those created using PRESS*, we may expect to see different surface plots. In the following examples we
will examine how PRESS** behaves as a data-driven selector. An analysis will be made on how the PRESS** results (for MRR1 and MRR2) compare to the optimal (AVEMSE) results for $b_0$ and $\lambda_0$.

The axis for the PRESS** values ($z$-axis) for the data sets when using MRR1 as the fitting technique had to be suppressed. This limiting of the maximum PRESS** values was only done for the reason of being able to analyze the surface in the area of importance. The double penalty of PRESS** created spikes in the surface that demonstrated how the bias term decreases as more LLR is used and the variance increases as the fit of the data is transitioned from an OLS fit to a LLR fit. The anomalies in the graphs occurred along a ridge of the $b - \lambda$ values in a semi-circular pattern. Due to the possible extreme values for PRESS** in these cases, the surfaces of the graphs were difficult to interpret. This problem did not exist when using PRESS** in combination with MRR2. Since this problem was not exhibited in the data sets when using MRR2, the PRESS** value axis for the MRR1 surface is limited to a comparable range to that of the surfaces for MRR2.

3.2.2.1 Data Set 1 (Same as Example 1 in section 3.1.1)

The underlying model for data set 1 is

$$y = 2(X - 5.5)^2 + 5X + 3.5\sin\left(\frac{\pi(X - 1)}{2.25}\right) + N(0, 16).$$

Recall that this is the same model used in Example 1 (see section 3.1.1). Even though the same underlying model is used for the AVEMSE example, just as it was also used for data set 1 when using PRESS* (see section 3.2.1.1), we get a very different surface structure than
observed in both cases. In Figure 3.17 we notice a flat surface where $\text{PRESS}^{**} = 20$. This flat area indicates the value that $\text{PRESS}^{**}$’s output was limited to due to an increase in the amount of bias.

In viewpoint (a) we get a good perspective of how the OLS fit to the data is transitioned into a LLR fit to the data by the use of the mixing parameter $\lambda$. The ridged area along the bandwidth axis where $\lambda$ is near 0 represents the pure OLS fit. This portion of the surface is elevated above the remainder of the observed area, indicating that the addition of a local linear fit to the data will definitely improve the situation. It is quite evident that the proper choice of $b$ is essential for any possibility of improving the fit and consequently lowering $\text{PRESS}^{**}$.

We see that for small values of $b$, $\lambda$ will have a large impact on the fit and deliver a lower $\text{PRESS}^{**}$ statistic. There must be special considerations given for the choice of $b$ here, because if $b$ is too large, $\text{PRESS}^{**}$ runs the risk of becoming overly inflated more than if
the bandwidth is chosen too small, thus not giving $\lambda$ an opportunity to help improve the fit to the data. The mixing parameter lowers the surface (i.e. improves the fit) as long as $b$ is considered to be near the optimal value. In Figure 3.17, viewpoint (b), we clearly see how the denominator of PRESS** imposes a penalty for large bandwidths as well as a penalty for bandwidth values too small. Evidently a choice for $b$ too small (i.e. near 0) is not as severe as a choice of $b$ above the “acceptable” region. The “acceptable” region may be defined by the portion of the surface that is sloping as $\lambda$ is increased toward 1.

The optimal values for this data set are $b_0 = 0.087$, $\lambda_0 = 0.445$, and AVEMSE$_0 = 8.147$. PRESS** favors a nonparametric (LLR) fit over the parametric fit of OLS (with an increased value of $\lambda$). The comparable values when using the data-driven selector are $b = 0.1$, $\lambda = 0.885$, and AVEMSE = 9.341. There is a noticeable difference in the chosen value for $\lambda$, where this selection process requires double of the LLR fit than that of the optimal fit. PRESS* demonstrated the complete opposite (See Section 3.2.1.1) where $b = 1$, $\lambda = 0.032$, and AVEMSE = 9.957. It is easy to see from Figure 3.17 that if the the wrong bandwidth is chosen, especially too large, then the only way $\lambda$ can help deliver a “reasonable” fit is to take on a value near 0, giving close to the OLS fit.

When using MRR2 in combination with PRESS** we get a very similar surface to that of MRR1 in combination with PRESS**. Remember in Figure 3.17 that the $z$-axis of PRESS** values was limited to large values. In Figure 3.18 the true maximum PRESS** is given as 19.36. This alone indicates that the given combination has less threat of destroying the fit as happened with MRR1 using PRESS**.
Just as with the previous figure we see that there is a definite need for larger values for $\lambda$. As we see in Figure 3.18 the mixing parameter is having a noticeable effect on the surface of the plot. For most choices of $b$ the surface is flat, indicating the OLS fit to the data. The mixing parameter is able to reduce PRESS** for values of $b$ less than 0.367. There are greater improvements possible as the value of $b$ gets smaller. As seen from viewpoint (b) in Figure 3.18, smaller values for the bandwidth will significantly assist in getting a “reasonable” fit, but if $b$ is too small, $\lambda$ is limited in the amount of help it will provide.

We find that there is great importance placed on the use of the mixing parameter, where

![Three rotations of the same graph](image)

Figure 3.18: PRESS** Plot for Data Set 1, using MRR2 as Fitting Technique

the selected value for $\lambda$ is 0.972 and $b = 0.091$. The optimal selection process chose $\lambda_0 = 0.513$ with $b_0 = 0.099$. Notice that the bandwidth value is nearly the same but the value for $\lambda$ is nearly twice that of the optimal selection process. Unfortunately, the $\text{AVEMSE}_{(\text{PRESS**})}$ for MRR2 is 10.042 where the $\text{AVEMSE}_0 = 8.091$. In section 3.2.1.1 when using PRESS*, we recall that the selected values for $b$ and $\lambda$ are 0.191 and 0.384, respectfully, while the
AVEMSE\(_{(PRESS^*)}\) = 9.088. The edge of the surface has a vertical rise when \(\lambda = 1\), and approaches a PRESS\(^**\) value just above 19.36, which is the worst that can happen if we add the entire residual fit to the OLS fit to the data. There is still a level of comfort by using the MRR2 technique over the MRR1 technique due to the fact that there is less variability with the PRESS\(^**\) values. This consistency when using PRESS\(^**\) is illustrated in the remaining examples as well.

3.2.2.2 Data Set 2

As previously mentioned, data set 2 is actually data set 1 without the error term, but \(\epsilon \sim N(0, 1)\) for calculations. In Figure 3.19 it is clear to see that there is a definite problem with the bias as created by the penalty functions in PRESS\(^**\). The combination of MRR1 and PRESS\(^**\) is only effective for the small values of \(b\) near 0.05 and \(\lambda\) has a noticeable impact once its value is greater than 0.33 (see Figure 3.19, viewpoint (a)). One can conclude that any choice of \(b\) within the “acceptable” range can assist in lowering PRESS\(^**\) (see Figure 3.19, viewpoint (c)) as long as the appropriate value of \(\lambda\) is chosen, giving obvious improvements over the OLS fit where \(\lambda = 0\). The values derived from using PRESS\(^**\) are \(b = 0.08\) and \(\lambda = 0.826\), with AVEMSE = 1.244, whereas the AVEMSE when using PRESS\(^*\) is 5.386. The optimal values for this data set are \(b_0 = 0.059\) and \(\lambda_0 = 0.922\), with AVEMSE\(_0\) = 0.905.

Based on the comparison of the AVEMSE for the three runs of this data set we see that the added penalty in PRESS\(^**\) offers some solution to the shortcomings of PRESS\(^*\), by
narrowing the interval for the choice of $b$. When using this data-driven selector a poor choice of $b$ is not as devastating as a bad choice of $\lambda$. PRESS** still outperforms PRESS* based on a mere comparison to the optimal values.

When we use MRR2, the key element to improving the fit is placed on the use of $\lambda$; this is detailed in Figure 3.20. The mixing parameter is having a significant effect on the surface of the plot for $b < 0.367$. One can conclude that for any choice of $b$ within an “acceptable”
range less than 0.367, PRESS** can be greatly lowered as long as a large value for \( \lambda \) is chosen. If the choice of \( b \) is greater than some “acceptable” interval, then a poor choice of \( b \) is not as devastating because that would revert the fit back to an OLS fit and \( \lambda \) has little impact at that point. We also notice how the values for PRESS** decrease rapidly once \( \lambda = 1 \) and \( b \) is small. There is also a noticeable difference between the range of PRESS** values for MRR1 and MRR2; when MRR1 is used the maximum value is inflated by the bias and when using MRR2 the maximum value is 19.05.

The optimal values for this data set are \( b_0 = 0.072 \) and \( \lambda_0 = 1 \), whereas the chosen values when using PRESS** are \( b = 0.03 \) and \( \lambda = 1 \). Using the data-driven selector PRESS** with MRR2 provides a much better fit than when using PRESS*, where the AVEMSE values are as follows: \( \text{AVEMSE}(\text{PRESS}^*) = 5.38 \), \( \text{AVEMSE}(\text{PRESS}^{**}) = 0.999 \) and \( \text{AVEMSE}_0 = 0.872 \).

3.2.2.3 Data Set 3 (Same as Example 3 in section 3.1.3)

We recall that data set 3 is the same as example 3, where example 3 is actually example 1 with two observations taken at each \( X \)-value (see Section (2.4.1.3)). We notice that the graph for MRR1 in Figure 3.21 has similar properties to that of data set 1 when using PRESS** (see Figure 3.17), except for the less dramatic spike in the PRESS** value for very small values of \( b \) when \( \lambda \) is near 1. The preferred bandwidth selection interval is identified in Figure 3.21, viewpoint (b). A value selected in this range will provide the chance for the addition of the nonparametric fit to lower the value of the PRESS** statistic as long as the value for \( \lambda \) is chosen appropriately.
We see that if we change the value for $\lambda$ there will be a more noticeable effect on the surface of the plot. When we look at Figure 3.21, viewpoint (a), it is evident that if $\lambda$ is too large for a small selection of $b$, such as $b = 0.05$, there will be a slight increase in the value for PRESS** (i.e. more variability). This rise in variation can be decreased by gradually increasing the value for $b$, thus still improving the fit of the misspecified model.

When using PRESS** and MRR1 in combination with this data set we did not achieve a fit any better than when using PRESS*. The given values here for $b$ and $\lambda$ are respectfully 0.058 and 0.765 with a AVEMSE = 5.69. The AVEMSE$_{(PRESS*)}$ is 4.984 which is slightly more comparable to the optimal values. The optimal values are $b_0 = 0.08$, $\lambda_0 = 0.613$, and AVEMSE$_0 = 4.945$.

Figure 3.22 for MRR2 clearly resembles Figure 3.6, where the optimal selection process is used to select the values for $b$ and $\lambda$. It is evident that $\lambda$ is having a large impact on the structure of the surface plot as long as the bandwidth is chosen in an appropriate range of (small) values. The flat area of the surface indicates that for any value of $b$ within this
range, all values of $\lambda$ will give a fit similar to the OLS fit to the data. It is critical here that $b$ is chosen correctly, and then $\lambda$ will dictate the amount of improvement to the model. Even though the overall shape of the surface is similar to that for AVEMSE, PRESS** was unable to select the desired values any better than PRESS*.

The maximum PRESS** value in Figure 3.22 for MRR2 is 15.58 and in Figure 3.21 for MRR1 the PRESS** value has been limited to a maximum of 16. We understand that if $b$ is mischosen ($b$ is chosen too large) when using MRR2 that the PRESS** value of 15.58 is the worst case scenario. This cannot be said for MRR1, where if $b$ is mischosen ($b$ is chosen too large) and an attempt is made to use $\lambda$ as a means to add in the nonparametric fit, we will certainly be in an area defined by a high bias term (not shown). Neither showed any significant advantage in terms of selecting $b$ or $\lambda$ if $b$ is selected too small, but as for the comfortability of having a less threatening outcome if $b$ is chosen too large, MRR2 wins. The values for $b$ and $\lambda$ are 0.061 and 0.784, respectively, where $b_0 = 0.097$ and $\lambda_0 = 0.742$. The comparable AVEMSE values are $\text{AVEMSE}(\text{PRESS*}) = 4.942$, $\text{AVEMSE}(\text{PRESS**}) = 5.684$, ...
and AVEMSE₀ = 4.842.

### 3.2.2.4 Data Set 4 (Same as Example 2 in section 3.1.2)

Data set 4 was generated from the following model:

\[ y = 5\sin(2\pi X) + \epsilon \]

as was Example 2 in section (3.1.2). The data set consists of twenty equally spaced \(X\)-values from 0 to 1 by 0.05, where \(\epsilon \sim N(0, 1)\). A cubic model would generally be specified for this data set if the true underlying model was unknown. As previously mentioned, (see Section 2.4.2.3) the selected values using PRESS** with MRR1 are noted as follows: \(b = 0.042\), \(\lambda = 0.999\), and AVEMSE = 0.551. In the current work, the observed structure of the surface plot was found to be near identical to those in previously analyzed surface plots. From an analysis of the plot, we see that when using PRESS** in combination with MRR1 that the values for PRESS** start seemingly high for small \(\lambda\) values (0.350), but one edge of the

![Graphical representation](attachment:image_url)
surface declines as the addition of the LLR fit begins improving the model and the PRESS** values are lowered (to 0.031). This is very different for how the optimal selection process develops when choosing the values for $b$ and $\lambda$, as seen in Figure 3.3. PRESS** continuously demonstrates how the double penalty separates the areas identified by the bias term and the variance term, as well as how the mixing parameter influences the total fit.

If the wrong value for $b$ is chosen, it is still possible to obtain a reasonable fit if the amount of the LLR fit is monitored with discretion. We are sure to improve the misspecified model if the initial choice of $b$ is chosen within an “acceptable” range. The range of acceptability is considered to be wider than that of $b_{(PRESS^*)}$. Because of the additional penalty imposed by PRESS**, it performed better than PRESS* when compared to the optimal values of $b_0 = 0.084$, $\lambda_0 = 0.478$, and $AVEMSE_0 = 0.271$ given in section 2.4.1.2. As previously stated, $b = 0.092$, $\lambda = 0.77$, and $AVEMSE = 0.3$ when using MRR1 with PRESS** as the data-driven selector.

The analysis of data set 4 (see Figure 3.24) when using PRESS** with MRR2 found the

![Figure 3.24: PRESS** Plot for Data Set 4, using MRR2 as Fitting Technique](image)

(Three rotations of the same graph)
values to be just as those stated in the previous research (see section 2.4.2.3). The chosen values are noted as \( b = 0.048, \lambda = 1, \) and \( \text{AVEMSE} = 0.482. \) It is quite obvious that the flat surface is characterizing an Ordinary Least Squares fit to the data, where any value for \( b \) for the residual fit in this range has no possibility in improving the fit if added back. If \( b \) is chosen within the sloping section of the plot, then \( \lambda \) can offer an improvement as compared to the original OLS fit.

We see in Example 2 (Figure 3.4) from section (3.1.2) that the optimal selection shows how the use of \( b \) and \( \lambda \) combined together lead to improvements of the misspecified model, but when using PRESS** we see that the surface dips deeper for a minimum value (i.e. lower PRESS**). The optimal values for this data set are \( b_0 = 0.126, \lambda_0 = 0.842, \) and \( \text{AVEMSE}_0 = 0.277, \) where as the values when using PRESS* are \( b = 0.061, \lambda = 1, \) and \( \text{AVEMSE} = 0.395. \) PRESS** combined with MRR2 shows considerable improvements over MRR1 because the “acceptable” range for the value of \( b \) as seen in Figure 3.24, viewpoint(b), is wider than that in Figure 3.23. For this data set, PRESS** outperforms PRESS*, and MRR2 performs better than when using MRR1. Also note that the 3-D plots for MRR2 using PRESS** are again similar in form to those for the AVEMSE example (in Figure 3.4).

3.2.2.5 Data Set 5

Remember that the final data set was generated from the model:

\[
y = 2(10X - 5.5)^2 + 50X + 3.5\sin(4\pi X) + \epsilon
\]

at twenty-one evenly spaced \( X \) values on the interval \([0, 1]\), and \( \epsilon \sim N(0, 1). \) When using
the MRR1 technique, data set 5 gives a plotted surface nearly identical to the previous four data sets. In Figure 3.25 the “acceptable” bandwidth range is identified by the slanted area parallel to the \( \lambda \) axis. At the top of the sloping portion we observe that the surface starts high with a PRESS** value just just above 6.84. In the mentioned area the surface descends rapidly to the point where PRESS** is at its minimal value.

A serious bias inflation issue is always a possibility when using MRR1, especially if \( b \) is chosen too large. For that reason, the PRESS** in Figure 3.25 has been limited to a maximum value of 10 for the purpose of analyzing the portion where PRESS** is minimized. It is noteworthy that in Figure 3.25, viewpoint (b), that the choice of \( \lambda \) strongly impacts the slope of the edge of the surface. This reaction in the surface is only where the bandwidth values are small. Clearly, if the wrong bandwidth is chosen, we may find ourselves in a case where \( \lambda \) can only do so much; for example, if the value for \( b \) is selected too far above 0.03, the mixing parameter may not be as helpful. The smaller bandwidth and mixing parameter combination values do not “blow up” (i.e. increased variance term), except when \( b = 0.03, \)
as we see that there is a slight jump in PRESS**, but as \( \lambda \) approaches 1 this spiking effect is reversed, thus lowering PRESS**.

The results for this data set are near optimal, where the optimal values are \( b_0 = 0.05 \), \( \lambda_0 = 1 \), and AVEMSE\(_0\) = 0.503. The resulting values here are \( b_{(PRESS**)} = 0.042 \) and \( \lambda_{(PRESS**)} = 1 \) with AVEMSE\(_{(PRESS**)}\) = 0.600. PRESS** performed well as a data-driven selector for \( b \) and \( \lambda \) with this data set. Even though near optimal values were selected, there is still some concern about the devastating effects that can occur from possibly selecting a value for \( b \) that is too large.

MRR2 also performed well when using PRESS** with results near optimal, yielding \( b = 0.042 \), \( \lambda = 1 \), and AVEMSE = 0.577. We see in Figure 3.26 that the surface is flat for large values of \( b \) indicating that for any \( b \) within this range no damage is done by the addition of the LLR fit to the residuals. This portion of the surface for MRR2 is a representation of an OLS fit to the data. The PRESS** value associated with the height of the flat surface

![Three rotations of the same graph](image)

Figure 3.26: PRESS** Plot for Data Set 5, using MRR2 as Fitting Technique
is 7.75, where the maximum PRESS** value when using MRR1 will be considerably higher than that due the bias issues. If the correct bandwidth is chosen, MRR2 stands a better chance at choosing near optimal values, thus an AVEMSE just as good as optimal. In order for the regression to achieve its potential improvement, as seen in Figure 3.26, viewpoint (b), the bandwidth must be near 0.04. Once $b$ is selected close to this value, the mixing parameter will become important.

MRR2 still offers the protection of not seriously destroying the model as demonstrated by MRR1 in previous PRESS** data sets. In the case of data set 5, PRESS** selected the same exact parameters for both MRR1 and MRR2. The selected choices when using MRR2 are $b = 0.042$ and $\lambda = 1$, but the AVEMSE is slightly lower at 0.577 than the AVEMSE when using MRR1. The structure of the 3-D plots for MRR2 using PRESS** is very similar to the other examples studied throughout this chapter.

3.2.2.6 Graphical PRESS** Results Summary

The graphical representations of the two techniques, MRR1 and MRR2, when using PRESS** give a better perspective into the process of selecting the values for $b$ and $\lambda$. Previous research by Mays et al. (2000) indicates that MRR2 performed the best and that PRESS** is the data-driven technique that yields the most appropriate choices for $b$ and $\lambda$. Graphically, it was shown how the data-driven selector, PRESS**, continuously yields smaller values for AVEMSE. As well as PRESS** outperforming PRESS*, the MRR2 technique demonstrated why it is a superior method over MRR1 when choosing the bandwidth and mixing parameter.
simultaneously.

The combination of MRR2 and PRESS** exhibits similar behavior to that of the optimal selection process. The MRR1 technique did not display the same sort of behavior when used to correct the misspecified model. In the AVEMSE examples in section 3.1, the surfaces of the plots all displayed a wave-like structure as the overall fit transitioned from OLS to LLR. In the AVEMSE examples in section 3.1 the optimal bandwidth was found in the range of $[0.08, 0.087]$. The data-driven selector PRESS** selected those values in the range of $[0.042, 0.1]$ for data sets 1, 3, and 4. Remembering that the AVEMSE values for the data sets when using PRESS** were considerably higher than those of the optimal examples may be explained here because the range of values for $b_{(PRESS**)}$ is much wider than that of $b_0$.

In each of the data sets when using MRR1 we encounter issues of bias inflation. This problem may be accredited to the additional penalty of $(n - 1)\frac{SSE_{mean} - SSE_b}{SSE_{mean}}$, which penalizes PRESS for the choice of a large bandwidth. In other words, when looking at the grid of all possible choices of $b$, this penalty term tends to inflate the value of PRESS** for large values of $b$. All of the plots in this section show just how differently PRESS** performs as compared to the optimal selection process when using the MRR1 technique. They all clearly show that if $b$ is considered to be too large that there will be a devastating effect to the criterion. Small bandwidth values find comfort between the two penalizing functions of PRESS**. In all of the cases, $\lambda$ is definitely important as long as $b$ was found to be within the interval of $[0.042, 0.1]$. The mixing parameter values are in the interval of $[0.765, 1]$, which is slightly wider when using PRESS** than the range of the optimal examples, which
is $[0.445, 0.613]$. The use of $\lambda$ is significant when using this data-driven selector as it favors the nonparametric portion of MRR1. The plots show that $\lambda$ only creates concerns when $b$ is too small. When $b$ is too small and $\lambda$ is near 1 we see the surface turning upward, indicating an increase in the variance term. This slight inflation of the variance is not as severe as what happens at the other extreme where $b$ is large and $\lambda$ is small, where the fit is equivalent to an OLS fit to the data. Overall, PRESS** does a good job at selecting the values for $b$ and $\lambda$, but MRR2 may offer a few more benefits that MRR1 can’t seem to consistently offer.

MRR2 displays the same sort of behavior with all data sets. The graphs when using PRESS** for the data sets resemble those found in the earlier sections for the AVEMSE examples and those when using PRESS* (see Sections 3.1 and 3.2.1). Even though these surfaces all contain a flat area representing the OLS fit to the data, each required some assistance from the nonparametric fit of the residuals. A wide range of values for $b$ are visible that contribute to $\lambda$ gradually improving the fit to the data. The range of $b$ and $\lambda$ when using MRR2 is $[0.048, 0.091]$ and $[0.784, 1]$, respectively, as compared to the range of the optimal values for $b$ and $\lambda$, $[0.097, 0.126]$ and $[0.513, 0.842]$, respectively. The range for the bandwidth used in the MRR2 technique is nearly twice the width for that of the optimal process. However, the range of $b$ values for MRR2 is considerably smaller than that required when using the MRR1 technique. This smaller range of values indicates that MRR2 is much more consistent in how it uses the addition of the LLR to the residuals to handle misspecified models. The range of the mixing parameter reinforces the idea that the fit can be improved by the addition of the nonparametric fit, thus the choice of $\lambda$ is essential.
We see that PRESS** combined with the MMR2 technique follows a trend similar to that of the optimal selection process where minimizing AVEMSE to select $b$ and $\lambda$ is concerned. This technique also offers much more confidence in terms of not being able to destroy a model any worse than an OLS fit, as long as $b$ is in the “acceptable” range. When using PRESS**, MRR2 does not inflate the values (i.e. PRESS** or AVEMSE) like MRR1, where there is considerably more instability. If model misspecification is a concern, MRR2 performs well with all of the procedures used in the current work.

The intent of the current work was to analyze the behavior of the semiparametric techniques when using data-driven selectors and compare them to the structural behavior of the optimal selection process. In the analysis of PRESS** we found that the overall pattern for MRR2 across all data sets was similar if not identical to the optimal (AVEMSE) plots. MRR2 followed the same pattern of designating a flat region that represents the OLS fit to the data. The MRR2 technique always used the entire OLS fit, indicating that this flat area is the worst case scenario for fits. PRESS** values were lowered for small $b$ values within an “acceptable” range and a portion of the LLR fit to the residuals was added. The same traits were seen in the optimal process where a lower AVEMSE was desired.

The range of “acceptable” $b$ values showed more consistency when using PRESS** than when using PRESS*. As previously mentioned, the range of $b$ was in the interval of $[0.05, 1]$ when using MRR1 and PRESS* together, where when we use MRR1 and PRESS** combined the range of $b$ was from 0.042 to 0.1. Clearly, the range when using PRESS** is smaller than that of PRESS*. This tighter range may reinforce the consistency that we seek.
when selecting an appropriate value for $b$. We find the same to be true when we compare the combinations of PRESS* and MRR2 with that of PRESS** and MRR2. The range of $b$ when using PRESS* with MRR1 is $[0.061, 0.191]$, which is considerably wider than $[0.048, 0.091]$ when combining PRESS** with MRR2. Regardless of which technique used, PRESS** outperforms PRESS* by way of demonstrating consistent behavior in selecting $b$ within a narrow range.
Chapter 4

Conclusion and Future Research

Previous research by Mays et al. (2000) found that of the two semiparametric techniques discussed, MRR1 and MRR2, that MRR2 was the semiparametric technique that delivered the more appropriate values for $b$ and $\lambda$ as well as smaller AVEMSE values. In the current work there were confirming results that support these earlier results. The graphical representations of the two techniques, MRR1 and MRR2, when using the data-driven selection procedures, PRESS* and PRESS**, gave interesting results. The results in the current work delivered insight into the properties of the two aforementioned semiparametric techniques as well as the two data-driven selectors. For the majority of the five data sets, both techniques displayed nearly the same structural behavior as the optimal selection process. A significant difference was apparent for small sample sizes where $n = 10$ in section 3.2.1 for the two techniques when using PRESS*. As seen in Figures 3.7 and 3.9, MRR1 demonstrates a spike in the surface when $b$ is small (near 0) and $\lambda$ is large (near 1). The spike seen in these figures
indicates a variance term that is inflated. MRR2 also spiked for small values of \( b \) and large values of \( \lambda \). In Figures 3.8 and 3.10, the surface area when using PRESS** definitely did develop as those in the optimal (AVEMSE) plots.

Due to the fact that PRESS* tends to chose large bandwidth values, MRR1 showed vulnerabilities where this is concerned because the larger the \( b \) the more bias the fit to the data will have. Clearly, careful consideration must be made when selecting the value of \( \lambda \). As the sample size increased, the combination of PRESS* and MRR1 improved. The plots of MRR1 when using the PRESS** procedure did not follow the same structural behavior as the optimal process or that of the PRESS* procedure. When using PRESS**, MRR1 showed issues with bias when the bandwidth was large. In particular, when using MRR1 there were several possible ways that the model could be destroyed if the values of \( b \) and \( \lambda \) are mischosen.

The structure of MRR2 consistently behaved nearly the same under all conditions. It was explicitly shown that the MRR2 technique was superior to that of MRR1 simply because if \( b \) is too large the worst that can be done to the model is an Ordinary Least Squares fit to the data. Even when the sample was an issue with the PRESS* - MRR1 combination, MRR2 still displayed a consistency of a wide range of values that would represent OLS. The range of selected values for MRR2 remained smaller and closer to those of the optimal process. MRR2 also maintained the same behavior when using either procedure, PRESS* or PRESS**. MRR2 was not effected by the additional penalty of PRESS** like MRR1. Overall, the MRR2 technique and the PRESS** selection procedure were the superior choices.
Future research may be considered on the development of the joint distribution for \( b \) and \( \lambda: f_{B,\Lambda}(b, \lambda) \), in addition to the individual distribution when analyzing them separately using simulations. Also, the graphical approach in the current work was found to be quite insightful as a tool for analyzing the behavior of the selection criteria for the semiparametric techniques. Thus this approach may be beneficial for selecting \( b \) and \( \lambda \) for MRR2 when using the asymptotically optimal \( \lambda \) of Mays et al. (2001). A graphical analysis of the data-driven procedures, PRESS* and PRESS**, can be done when using semiparametric techniques by the use of \( b \) and the asymptotically optimal \( \lambda \) as well. Since MSE is a natural criterion used for selection of \( b \) and \( \lambda \), it may be just as useful to create surface plots for the bias and variance associated with the \( b - \lambda \) combinations for the semiparametric techniques mentioned in the current work. Finally, development of a simplified calculation for deriving the optimal (or optimal range) of \( b \) and \( \lambda \) when using AVEMSE and PRESS** for the MRR2 technique may be considered at another time as well.

This current work only analyzed a limited number of data sets (to match previous research studies), and it would be worthwhile to study other data sets as well. In particular, studies could be completed that look at the performance of the selectors across different sample sizes and different misspecification levels.
Bibliography


versity Press.

discussion), *Statistical Science*, 8, pp. 120-143.

27, pp. 415-438.

and Semiparametric Methods, *Ph.D. Dissertation*, Virginia Polytechnic Institute and
State University, Blacksburg.

Nonparametric, and Semiparametric Methods, *Technical Report Number 96-7*, Dept. of
Statistics, Virginia Polytechnic Institute and State University, Blacksburg.


ing Parametric, Nonparametric, and Semiparametric Methods, *Journal of Nonparametric
Statistics*, 13, pp. 245-277.


Vita

Derick Lorenzo Rivers was born on June 12, 1970 in Richmond, Virginia. After graduating from George Wythe High School in 1989, he received his Bachelor of Science degree in Mathematics in 1995 from Morris College, Sumter, South Carolina. He received a teaching certificate and taught for four years in Richmond Public Schools before returning to graduate school. He received a Master of Education in Mathematics Education in 2004 from Cambridge College, Cambridge, Massachusetts. He has been a Mathematics Instructor at ECPI College of Technology since May 2004.