Chapter 2 : Estimation of the Means Model

The regression model describes the relationship between a set of k regressor variables $(X_1, X_2, \cdots, X_k)$ and a response variable of interest (y). Written in its most general form, the model is given as

$$y_i = h (X_{1i}, X_{2i}, \cdots, X_{ki}) + \varepsilon_i, \quad i = 1, 2, \ldots, n \qquad (2.1)$$

where ε_i is a random error term whose expected value is taken to be zero. The primary goal of regression analysis is to provide some estimate $\hat{h}$ of h such that the value of the response can be predicted for any point of interest $\mathbf{x}_o' = (X_{1o} \; X_{2o} \; \cdots \; X_{k\,o})$. Since $E (y \mid X_1 , X_2 , \cdots , X_k) = h$, it follows that $\hat{h}$ is an estimate of the conditional mean of y at $\left(X_1 \; X_2 \; \cdots \; X_k\right)$. We therefore refer to (2.1) as the process means model. The technique used in the estimation of h depends on the form that the user specifies for h . In general, there are three forms of h that the user may specify : a parametric form, a nonparametric form or a semiparametric form. In this chapter we will discuss the estimation of h in each of the aforementioned cases and point out the strengths and weaknesses associated with them.

2.A Parametric Approach

The most common parametric form expresses h as a linear model involving parameters $\beta_o, \beta_1, \cdots, \beta_k$:

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

The linear model may be expressed in matrix notation as :

$$\mathbf{y} = h\left(\mathbf{X} ; \boldsymbol{\beta}\right) + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} . \qquad (2.A.2)$$

In this notation, $\mathbf{y}$ is an (n x 1) vector of responses, $\mathbf{X}$ is an (n x (k + 1)) matrix of the k regressors augmented by a column of ones, $\boldsymbol{\beta}$ is a ((k + 1) x 1) vector of unknown parameters, and $\boldsymbol{\varepsilon}$ is the (n x 1) vector of random errors. In developing the current research, we will keep our discussion limited to linear models of a single regressor variable where the model terms $(X_{1i}, X_{2i}, \cdots, X_{ki})$, will simply be polynomial expressions of the single regressor x_i (i.e. $X_{1i} = x_i$, $X_{2i} = x_i^2$, etc.). Thus we will write :

$$\begin{aligned} y_i &= h\left(\mathbf{x}_i ; \boldsymbol{\beta}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + \varepsilon_i, \text{ for } i = 1, \ldots, n \end{aligned}$$

$$= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$

where $\mathbf{x}_i' = \left(1\ x_i\ x_i^2 \cdots x_i^k\right)$ is the i^{th} row of the X matrix.

2.A.1 Ordinary Least Squares

The goal of parametric regression is to provide the best possible estimates of the unknown parameters in $\boldsymbol{\beta}$ which in turn, leads to the estimated mean responses, given as $\hat{\mathbf{y}} = \hat{\boldsymbol{h}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{h}}$ is a (n x 1) vector. Assuming that the random errors follow a Gaussian distribution with constant variance, σ^2, the uniform minimum variance unbiased estimates (UMVUE) of the parameters in $\boldsymbol{\beta}$ are obtained by maximizing the normal log-likelihood with respect to $\boldsymbol{\beta}$. Writing the normal log-likelihood as :

$$l(\boldsymbol{\beta}, \sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2, \qquad (2.A.1.1)$$

it is clear that $l(\boldsymbol{\beta}, \sigma)$ is maximized for the value of $\boldsymbol{\beta}$ which yields the 'least' sum of squared errors, given as : $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$. Thus, the estimation technique is often termed 'least squares'. If we let e_i denote the residual at the point x_i, we have that $e_i = y_i - \hat{y}_i$, where $\hat{y}_i = h\left(\mathbf{x}_i; \hat{\boldsymbol{\beta}}\right) = \hat{h}_i$. The estimated mean responses, obtained from ordinary least squares can be written as

$$\hat{\boldsymbol{h}} = \hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}^{(ols)}\mathbf{y}. \qquad (2.A.1.2)$$

The matrix $\mathbf{H}^{(ols)}$ is known as the ordinary least squares "HAT" matrix. The term "HAT" comes from the fact that the HAT matrix produces the 'y-hat' values through a transformation of the observed y values. The HAT matrix plays a major role in much of regression analysis and many of its properties will be referred to in this research. Some of these properties are :

$\mathbf{H}^{(ols)}$ is symmetric and idempotent. (2.A.1.3)

$\mathrm{tr}(\mathbf{H}^{(ols)}) = k + 1$ where k is the number of regressors in the in the model. (2.A.1.4)

$\sum_{j=1}^{n} h_{ij}^{(ols)} = 1$ where $h_{ij}^{(ols)}$ is the i,j^{th} element of $\mathbf{H}^{(ols)}$ (2.A.1.5)

$$e_i = (1 - h_{ii}^{(ols)})\, y_i - \sum_{j \neq i}^{n} h_{ij}^{(ols)} y_j \qquad (2.A.1.6)$$

$$Var(e_i) = E(e_i^2) = (1 - h_{ii}^{(ols)})\, \sigma^2 \qquad (2.A.1.7)$$

$$Var(\hat{y}_i^{(ols)}) = \sigma^2 h_{ii}^{(ols)} \qquad (2.A.1.8)$$

For the development of these properties, see Myers (1990) and Hoaglin and Welsch (1978). Before moving on, there is a subtlety that is implied by (2.A.1.1) that should be discussed. Notice from (2.A.1.2) that the fitted value $\hat{y}_i^{(ols)}$ at x_i can be written as :

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}^{(ols)} y_j \qquad (2.A.1.9)$$

Thus, the fit $\hat{y}_i$ at x_i , is a weighted average of the observed y_j's where the weights are the elements of the i^{th} row of $\mathbf{H}^{(ols)}$. The values of these weights (the h_{ij}'s) depend on the model chosen by the researcher. For instance, in simple linear regression

$$h_{ij}^{(ols)} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \qquad (2.A.1.10)$$

Notice that those points which have the most influence on prediction at a point x_i are those points which are the farthest away from x_i. Conversely, the x_j's which are closest to x_i are given less consideration. Thus, in least squares, points which are observed at the extremes of our x-space have greater 'leverage' on the overall fit than interior points of the x-space. If the model assumed by the researcher is incorrect, it may be more beneficial to consider a different weighting scheme. The idea of an alternative weighting philosophy will be discussed later in this chapter when nonparametric regression techniques are described.

2.A.2 Weighted Least Squares

In the discussion of ordinary least squares, it was assumed that the random errors possessed constant variance, σ^2, across the domain of the data. This assumption however, is often invalid. For instance, it is not uncommon to have a process in which the variation in response depends on the magnitude of the regressors. An example of this is seen in Myers (1990) presentation of the Transfer Efficiency data. In this data set, it is believed that two regressors, air velocity (X_1) and voltage (X_2), influence the efficiency of a particular type of spray paint

equipment (the response variable, y). Not only do these two regressors influence y, but it is pointed out that as voltage increases, there is greater variation in the measurements of y. Therefore, in estimating the mean regression function, it appears intuitive to weight our observations in such a way as to give more influence to those observations which are known to have small variability and less influence to those with large variability.

Assuming that the variances of the errors at the *n* data points are given by σ_1^2 , σ_2^2 , $\cdots$, σ_n^2 , it is easy to show that the UMVUE estimates for $\boldsymbol{\beta}$ can be obtained through weighted least squares (WLS) where the weights are just the reciprocals of the variances at each of the data points. The WLS estimate of the underlying regression function is given by the expression

$$\begin{aligned} \hat{\boldsymbol{h}}\left(\mathbf{X};\boldsymbol{\beta}^{(wls)}\right) &= \hat{\mathbf{y}}^{(wls)} = \mathbf{X}\,\hat{\boldsymbol{\beta}}^{(wls)} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ &= \mathbf{H}^{(wls)}\mathbf{y}, \end{aligned} \tag{2.A.2.1}$$

where $\mathbf{V} = \text{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2)$ and $\mathbf{H}^{(wls)} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$. Like $\mathbf{H}^{(ols)}$, $\mathbf{H}^{(wls)}$ has many important properties which will be referred to in this research. Some of these are :

$\mathbf{H}^{(wls)}$ is idempotent. (2.A.2.2)

$$e_i^{(wls)} = (1 - h_{ii}^{(wls)})\,y_i - \sum_{j \neq i}^{n} h_{ij}^{(wls)}\,y_j \tag{2.A.2.3}$$

where $h_{ij}^{(wls)}$ is the i,j^{th} element of $\mathbf{H}^{(wls)}$.

$$\text{Var}(e_i^{(wls)}) = E(e_i^{2\,(wls)}) = (1 - h_{ii}^{(wls)})^2\,\sigma_i^2 + \sum_{j \neq i}^{n} h_{ij}^{2\,(wls)}\,\sigma_j^2 \tag{2.A.2.4}$$

$$\text{Var}(\hat{y}_i^{(wls)}) = x_i'(X'V^{-1}X)^{-1}x_i \tag{2.A.2.5}$$

One could very well argue that weighted least squares is a procedure which is only practical in a theoretical sense since it assumes the variances at the *n* data points are known. This assumption, however, is rarely satisfied in practice. A possible solution is to estimate the variances of the *n* data points and perform an estimated weighted least squares (EWLS) analysis of the data. The EWLS estimate of the underlying regression function is given by

$$\begin{aligned} \hat{\boldsymbol{h}}\left(\mathbf{X};\boldsymbol{\beta}^{(ewls)}\right) &= \hat{\mathbf{y}}^{(ewls)} = \mathbf{X}\,\hat{\boldsymbol{\beta}}^{(ewls)} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \end{aligned} \tag{2.A.2.6}$$

$$= \mathbf{H}^{(\text{ewls})}\mathbf{y},$$

where $\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_n^2)$ and $\mathbf{H}^{(\text{ewls})} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}$. The concept of variance estimation and EWLS brings greater complexity to the analysis, many of which motivate this research and are discussed in detail in Chapter 3.

2.B Nonparametric Approach

Recall the general regression model in (2.1) and the fact that the underlying motive in regression is to provide the best possible estimate of the regression function, h. The linear, parametric approach to this problem assumes that h takes on the form given in (2.A.1) where h is described by known parameters. In the nonparametric regression setting however, the assumption regarding the form of h is less restrictive. The regression function h is only assumed to take on some arbitrary smooth form. Like parametric regression, nonparametric regression uses the data to estimate h and the estimate is a weighted sum of the response values (the y's). However, the 'weighting' philosophy in nonparametric regression is such that observations closest to the point of interest, x_o, have the most information about the mean response at x_o. Thus, those points in closest proximity to x_o are given more weight in obtaining $\hat{y}(x_o)$. In the next two sections two popular methods of nonparametric regression analysis are outlined.

2.B.1 Kernel Regression

As mentioned, the philosophy of nonparametric regression is to estimate the regression function h using a weighted average of the raw data where the weights are a function of distance in the x-space. In particular, the weights are a decreasing function of distance . A weighting scheme of this type is proposed by Nadaraya (1964) and Watson (1964) in which the weight associated with observation y_j, for prediction at x_i is given by :

$$h_{ij} = \frac{K\left(\frac{x_i - x_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right)}. \qquad (2.B.1.1)$$

The function K (u) is taken to be some appropriately chosen decreasing function in | u |. The parameter, b is known as the smoothing parameter or bandwidth. The choice of K and b are topics of discussion in sections 2.B.2 and 2.B.3. The kernel estimate of the regression function h at the point x_i is given by :

$$\hat{h}\,(\,x_i\,) \;=\; \hat{y}_i^{(\ker)} \;=\; \sum_{j\,=\,1}^{n} h_{ij}^{(\,\ker\,)}\, y_j \;=\; \mathbf{h}_i^{(\ker)'}\,\mathbf{y}\,. \qquad (2.B.1.2)$$

Rewriting (2.B.1.2) in matrix notation we have

$$\hat{\boldsymbol{h}} \;=\; \hat{\mathbf{y}}^{(\ker)} \;=\; \mathbf{H}^{(\ker)}\,\mathbf{y} \qquad (2.B.1.3)$$

where

$$\mathbf{H}^{(\,\ker\,)} \;=\; \begin{bmatrix} \mathbf{h}_1^{(\ker)'} \\ \vdots \\ \mathbf{h}_n^{(\ker)'} \end{bmatrix}$$

and $\mathbf{h}_i^{(\ker)'} \;=\; \left(h_{i1}^{(\ker)},\,\cdots,\,h_{in}^{(\ker)}\right)$. The matrix $\mathbf{H}^{(\,\ker\,)}$ is referred to as the kernel HAT matrix, or kernel smoother matrix. Kernel predictions at an arbitrary point, x_o, may be obtained by using equation (2.B.1.2), replacing the "*i*" by "o". Then we can write $\hat{h}\left(x_o\right) \;=\; \mathbf{h}_o^{(\ker)'}\,\mathbf{y}$, where, $\mathbf{h}_o^{(\ker)'} \;=\; \left(h_{o1}^{(\ker)},\,\cdots,\,h_{on}^{(\ker)}\right)$. Notice that a disadvantage of kernel regression is that, unlike parametric regression, it offers no closed form estimate for *h* .

2.B.2 Choice of Kernel Function

The name 'kernel' regession comes from the fact that the estimated regression function at x_o is obtained by taking a weighted average of the y values where the weights are produced by the kernel function, K(u). H„rdle (1990) considers the issue of which kernel function is "optimal" and he illustrates through efficiency arguments that, for the general case of twice differentiable kernels, the choice of kernel function is not critical to the performance of the kernel regression estimator. The kernel function K(u) is typically chosen to be nonnegative, symmetric about zero, continuous and twice differentiable. Some popular functions are the Gaussian kernel, the uniform kernel, the Epanechnikov kernel (Epanechnikov (1969)), the quartic kernel and the cubic kernel. Since the choice of kernel function is not critical to the performance of the kernel regression estimator, for a matter of convenience, we will use the simplified Gaussian kernel written here as

$$K\,(\,u\,) \;=\; e^{-u^2} \qquad (2.B.2.1)$$

where $u = \left(\frac{x_i - x_j}{b} \right)$. Notice that this weighting scheme uses the simplified Gaussian probability density function to assign weights to the points around a given x_o of interest. In other words, if we picture a normal curve above the x axis, centered at x_o , points closest to x_o would receive the most weight for prediction at x_o whereas those points lying in the tails of the normal density (points far removed from x_o) receive minimal weight. Although the choice of kernel function is not a critical issue, the choice of the bandwidth is critical. The next section outlines various methods for bandwidth selection.

2.B.3 Choice of the Smoothing Parameter *b*

In the previous section it was mentioned that in the kernel regression weighting scheme, the magnitude of the weights decreases as an observation's distance from x_o increases. For Gaussian kernel functions, the speed at which the kernel weights decreases as a function of x depends on the spread of the normal density. Specifically, if we rewrite (2.B.2.1) substituting for u, we have

$$K\left(\frac{x_j - x_o}{b} \right) = e^{-\left(\frac{x_j - x_o}{b} \right)^2} . \qquad (2.B.3.1)$$

Notice that the spread of the normal density is determined by *b*. Thus, if *b* is large, then the normal curve will be "wide" and more observations will be utilized for prediction at a particular point x_o . However, if *b* is small, the density will be narrow and only a few observations will be given weight in determining the fit at x_o . In fact, if *b* is large enough, the normal density would appear uniform and the prediction at any x_o would just be the average of the observations (where all observations would have equal weight). This estimate of h could be likened to the situation in parametric regression of estimating h with the 'intercept only' model. Such estimates of h are described as "under- fit" models and are associated with large bias in estimation of the mean response. If *b* is small (i.e. ≈ 0), we go to the other extreme and "over-fit" the model. If $b \approx 0$, our window with which to assign weights has a width of approximately zero and the only point with any weight is the point of interest. Such an estimate of h would simply result in a "connect-the-dots" fit of the y values, producing a jagged and highly variable fit. Notice the trade-off which exists: a bandwidth which is too large results in bias problems whereas one that is too small produces a fit which is too variable. Thus, the selected bandwidth should be one which produces an estimate that possesses a suitable balance of bias and variance in the estimated fit.

This goal of selecting the bandwidth to strike the proper blend of bias and variance leads naturally to minimization of a mean squared error criterion for determining the appropriate bandwidth. The literature is rich regarding bandwidth selection and numerous procedures have been developed. The remainder of this section will serve to provide an overview of one of the more popular selection techniques known as cross-validation.

Cross - Validation (PRESS)

The most practical method of determining if a selected bandwidth is appropriate is to evaluate its performance based on some global error measure for the particular regression fit. One such global error measure is the average mean squared error (AVEMSE) for the regression curve. The AVEMSE for a given kernel estimate of *h* using bandwidth *b* can be expressed as

$$\text{AVEMSE}\left[\ \hat{h}_b(\mathrm{x})\ \right] = n^{-1} \sum_{j=1}^{n} \mathrm{E}\left[\hat{h}_b(\mathrm{x}_j) - h(\mathrm{x}_j)\right]^2 , \qquad (2.B.3.2)$$

where $\hat{h}_b(\mathrm{x})$ denotes the kernel estimate of the true regression function *h*. The only problem with the expression in (2.B.3.2) is that its use in determining *b* assumes that the researcher can explicitly state the true regression function *h*. As a solution to this dilemma, instead of choosing the value of *b* which minimizes the AVEMSE, perhaps we can choose the value of *b* which minimizes a unbiased estimate of the AVEMSE.

One such unbiased estimate of the AVEMSE is the cross-validation statistic of Stone (1974), given here as

$$\mathrm{CV}\ (\ b\ ) = \mathrm{n}^{-1} \sum_{i\ =\ 1}^{n} \left(\mathrm{y}_i - \hat{\mathrm{y}}_{i,-i}\right)^2 \mathrm{w}(\mathrm{x}_i)\ . \qquad (2.B.3.3)$$

The notation $\hat{\mathrm{y}}_{i,-i}$, implies that this is the estimated value of our regression function at x_i when observation $\left(\mathrm{y}_i, \mathrm{x}_i\right)$ is left out of the weighted average of the y values. For example, we write $\hat{\mathrm{y}}_{i,-i}^{(\mathrm{ker})} = \sum_{j\ \neq\ i}^{n} \mathrm{h}_{ij,-i}^{(\mathrm{ker})}\ \mathrm{y}_j$ where $\mathrm{h}_{ij,-i}^{(\mathrm{ker})}$ denotes the $i\ ,\ j^{\mathrm{th}}$ element in the matrix $\mathbf{H}_{i,-i}^{(\ \mathrm{ker}\ )}$.

It is easy to show that $\mathrm{h}_{ij,-i}^{(\mathrm{ker})} = \dfrac{\mathrm{h}_{ij}^{(\mathrm{ker})}}{1\ -\ \mathrm{h}_{ii}^{(\mathrm{ker})}}$. Ignoring the dependence on *n* and *w*, the expression in (2.B.3.3) is just the familiar PRESS statistic given by Allen (1974). Using this statistic, numerical methods are used to find the optimal *b* by finding the bandwidth which minimizes (2.B.3.3).

A problem which results from using the PRESS criterion is that it often selects bandwidths which are too small. This problem was noted by Einsporn (1987) and as a solution to this problem the following penalized PRESS (known as PRESS*) was proposed

$$\frac{\text{PRESS}}{\mathrm{n}\ -\ \mathrm{tr}(\ \mathbf{H}^{(\ \mathrm{ker}\ )})}\ . \qquad (2.B.3.4)$$

The "penalty" used in PRESS* is found in its denominator. If a small bandwidth is used, the diagonal elements of $\mathbf{H}^{(\ker)}$ get larger, resulting in a larger value for tr($\mathbf{H}^{(\ker)}$). As a result, the denominator gets smaller and PRESS is thus penalized for using a small bandwidth.

In work by Mays (1996), it was observed that the penalty in PRESS* for small bandwidths is too severe, resulting in choices of b that are too large. Mays proposes the following solution (which he termed PRESS**):

$$\text{PRESS}^{**} = \frac{\text{PRESS}}{n - \text{tr}(\mathbf{H}^{(\ker)}) + (n - 1)\dfrac{\text{SSE}_{\text{tot}} - \text{SSE}_b}{\text{SSE}_{\text{tot}}}}, \qquad (2.B.3.5)$$

where SSE_{tot} is the total sum of squares for the y's and SSE_b is the sum of squared errors resulting from any candidate bandwidth b. Notice that the extra term in the denominator, $\dfrac{\text{SSE}_{\text{tot}} - \text{SSE}_b}{\text{SSE}_{\text{tot}}}$, takes on values between 0 and 1. This expression goes to 0 for $b \to 1$ and approaches 1 for $b \to 0$. Thus, this penalty structure is what is desired for improving the problems encountered with PRESS*. The PRESS** statistic has been shown to work well in a variety of applications (Mays, 1996) .

2.B.4 Local Polynomial Regression

Kernel regression has received a great deal of attention over the years due to its intuitive approach and ability to be extended to higher dimensional problems. In our presentation of kernel regression, we stated that the approach used in estimating the regression function at x_o is simply a weighted average of the sample y values. Unfortunately, this simple approach to estimation has several flaws. First of all, when we assume a symmetric kernel function such as the Gaussian probability density function, the kernel estimate experiences bias problems at the boundaries of the x-space. The problem of bias can also exist within the interior of the data if the x_{i}'s are non-uniform or if there is substantial curvature in the regression function. These problems become magnified when the regressors are multidimensional. In an attempt to address the problems of kernel regression, Cleveland (1979) introduced the technique known as local polynomial regression. Local polynomial regression can be thought of as an expansion of kernel regression. Consider the kernel estimate of $h(x_o)$ given in expression (2.B.1.2) written in expanded form as :

$$\hat{h}(x_o) = h_{01} y_1 + h_{02} y_2 + \cdots + h_{0n} y_n \qquad (2.B.4.1)$$

Equivalently, in matrix-vector notation we have

$$\hat{h}(x_o) = 1 (\mathbf{1}' \mathbf{W}(x_o) \mathbf{1})^{-1} \mathbf{1}' \mathbf{W}(x_o) \mathbf{y} \qquad (2.B.4.2)$$

where $\mathbf{W}(x_o)$ is a diagonal matrix of the kernel weights associated with x_o, $\mathbf{1}$ is a n dimensional vector of unity elements and 1 is a scalar. Now recall from the discussion in Section 2.A.2 on weighted least squares (a parametric approach), that the prediction at an arbitrary point $\mathbf{x}_o'$ was given by

$$\hat{h}(\mathbf{x}_o') = \mathbf{x}_o'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad (2.B.4.3)$$

where $\mathbf{x}_o' = (1\ x_o\ x_o^2 \cdots x_o^k)$, $\mathbf{X} = \left[\mathbf{x}_1'\ \mathbf{x}_2' \cdots \mathbf{x}_n'\right]$ and $\mathbf{V} = \text{diag}\left(\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2\right)$. Notice the similarities in the fits given in (2.B.4.2) and (2.B.4.3). Ignoring the difference in weight matrices used, expression (2.B.4.3) can be viewed as a higher dimensional fit at x_o than (2.B.4.2). Kernel regression can be thought of as a 'kernel weighted' least squares where an intercept only model is being fit.

Cleveland uses this observation to propose a more sophisticated nonparametric regression fit. Instead of the intercept only approach used in kernel regression, Cleveland suggests fitting a k^{th} degree polynomial ($k > 0$) at each x_o. Cleveland's approach (termed "local polynomial regression) can be viewed as WLS where $\mathbf{V}^{-1}$ is replaced the matrix of kernel weights. It should be noted that polynomials of degree $k = 1$ or $k = 2$ effectively address the problems associated with kernel regression. For purposes of this research we will consider polynomials of degree $k = 1$, commonly referred to as local linear regression. The local linear estimate of the underlying regression function is given here in matrix notation as

$$\begin{aligned}\hat{h}_i\left(\mathbf{X};\boldsymbol{\beta}_i^{(\text{llr})}\right) &= \hat{y}_i^{(\text{llr})} = \boldsymbol{x}_i'\,\hat{\boldsymbol{\beta}}_i^{(\text{llr})} \qquad (2.B.4.4)\\ &= \boldsymbol{x}_i'(\mathbf{X}'\mathbf{W}^{(\text{llr})}(\boldsymbol{x}_i)\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(\text{llr})}(\boldsymbol{x}_i)\mathbf{y}\\ &= \mathbf{h}_i'^{(\text{llr})}\mathbf{y},\end{aligned}$$

where $\mathbf{W}^{(\text{llr})}(\boldsymbol{x}_i)$ is a diagonal matrix consisting of the kernel weights associated with x_i and $\mathbf{h}_i'^{(\text{llr})} = \boldsymbol{x}_i'(\mathbf{X}'\mathbf{W}^{(\text{llr})}(\boldsymbol{x}_i)\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(\text{llr})}(\boldsymbol{x}_i)$. In matrix notation, the n local linear fitted values can be expressed as $\hat{\mathbf{y}}^{(\text{llr})} = \mathbf{H}^{(\text{llr})}\mathbf{y}$, where

$$\mathbf{H}^{(\text{llr})} = \begin{bmatrix}\mathbf{h}_1'^{(\text{llr})}\\ \mathbf{h}_2'^{(\text{llr})}\\ \vdots\\ \mathbf{h}_n'^{(\text{llr})}\end{bmatrix}. \qquad (2.B.4.5)$$

2.C Parametric or Nonparametric

In the previous sections, several approaches to estimating the regression function have been presented and it is natural to ask if there is a globally optimal approach. The answer to this question invariably depends on the from of the underlying regression function. If the underlying regression function can be adequately expressed parametrically and the user specifies the correct parametric model, then obviously parametric regression should be used. However, if the researcher has no idea concerning the true form of the underlying regression function, then a nonparametric approach such as local linear regression should be utilized. The problem arises when we consider applications which fall between these two scenarios: the researcher has some feel for the underlying structure of h but there may be places where the data deviates from this structure. A purely parametric model would be inappropriate as the structure would be too rigid to capture specific deviations in the data's trend and the result would be an estimate that contains too much bias or "lack of fit". On the other hand, a purely nonparametric model is insufficient as the fit would likely be too variable and not use the researcher's knowledge of the underlying structure. Mays and Birch (1996) propose a procedure termed Model Robust Regression which essentially mixes together a parametric fit and a nonparametric fit to obtain a superior estimate of h. The details of this procedure are outlined in the next section.

2.D Model Robust Regression

The basic idea of model robust regression (MRR), is to improve the estimate of the regression function by combining parametric and nonparametric estimates via a mixing parameter, λ. The idea of using a mixing parameter to combine both parametric and nonparametric estimates was first proposed by Einsporn (1987) and Einsporn and Birch (1993). The proposed regression function estimate of Einsporn and Birch, known as the HATLINK estimate, is written as

$$\hat{\boldsymbol{h}} = \hat{\mathbf{y}}^{(\text{hatlink})} = \lambda\, \hat{\mathbf{y}}^{(\text{ker})} + (1 - \lambda)\, \hat{\mathbf{y}}^{(\text{ols})}. \qquad (2.D.1)$$

The name "hatlink" comes from the fact that when (2.D.1) is written in matrix notation we have

$$\begin{aligned} \hat{\mathbf{y}}^{(\text{hatlink})} &= \lambda\, \mathbf{H}^{(\text{ker})}\mathbf{y} + (1 - \lambda)\mathbf{H}^{(\text{ols})}\mathbf{y} \qquad (2.D.2) \\ &= \mathbf{H}^{(\text{hatlink})}\mathbf{y} \end{aligned}$$

where $\lambda \in [0,1]$ and $\mathbf{H}^{(\text{hatlink})} = \lambda\, \mathbf{H}^{(\text{llr})} + (1 - \lambda)\mathbf{H}^{(\text{ols})}$ with $\mathbf{H}^{(\text{ols})}$ and $\mathbf{H}^{(\text{ker})}$ being the ordinary least squares and kernel HAT matrices, respectively.

The mixing parameter, λ, ranges from 0 to 1 depending on the amount of misspecification in the users parametric model. For situations where the researcher's specified parametric model is correct, the optimal value of λ is 0, whereas if the user's parametric model is far from correct, λ

= 1 is optimal. Notice that the "hatlink" estimate, via the mixing parameter λ, attempts to provide a fit which possesses the proper mix between a parametric fit and a separate nonparametric fit to the raw data. However, if there are locations in the data in which both the parametric and nonparametric fits are positively biased or both negatively biased, the procedure has no means of resolving the problem. This is the motivation for the development of MRR.

Mays (1996), like Einsporn and Birch utilizes a mixing parameter but only one fit to the raw data is required. The raw data is fit by ordinary least squares using a researcher supplied parametric model. Any trend in the data not captured parametrically is presumed to be contained in the set of *n* residuals from the parametric fit, denoted here $\mathbf{e}^{(ols)} = \mathbf{y} - \hat{\mathbf{y}}^{(ols)}$. Local linear regression is then used to detect any structure in the set of parametric residuals and a portion of this structure, determined by the mixing parameter λ, is then added back to the parametric fit. The MRR estimate, originally referred to as the MRR2 estimate by Mays (1996), of the regression function is given by

$$\hat{\boldsymbol{h}} = \hat{\mathbf{y}}^{(mrr)} = \hat{\mathbf{y}}^{(ols)} + \lambda\,\hat{\mathbf{r}} \qquad (2.D.3)$$

where $\hat{\mathbf{r}} = \mathbf{H}^{(llr)}\mathbf{e}^{(ols)}$ and $\mathbf{H}^{(llr)}$ denotes the local linear hat matrix used to fit the set of *n* OLS residuals. Thus, we can write:

$$\begin{aligned}\hat{\mathbf{y}}^{(mrr)} &= \mathbf{H}^{(ols)}\mathbf{y} + \lambda\mathbf{H}^{(llr)}\mathbf{e} \qquad (2.D.4)\\ &= \mathbf{H}^{(mrr)}\mathbf{y}\end{aligned}$$

where $\mathbf{H}^{(mrr)} = \mathbf{H}^{(ols)} + \lambda\mathbf{H}^{(llr)}\left(\mathbf{I} - \mathbf{H}^{(ols)}\right)$ and $\lambda \in [0,1]$. Regarding the notation, **I** is an (*n* x *n*) identity matrix and $\mathbf{H}^{(mrr)}$ is known as the MRR "hat matrix". The next section outlines the data driven procedure used for determining the appropriate value of λ.

2.D.1 Choosing λ

In discussing the procedure of determining the optimal value of λ, it is important to recall the selection of the optimal bandwidth in nonparametric regression. The idea was that an appropriate bandwidth is one which generates a nonparametric fit which has the proper mix of bias and variance. These same issues regarding bias and variance are also present when choosing λ. Mays and Birch (1996) utilize the cross-validation technique of PRESS**, discussed in Section 2.B, for selecting the appropriate mixing parameter. The PRESS** statistic used for selection of the mixing parameter λ is written here as

$$\text{PRESS}^{**} = \frac{\text{PRESS}}{n - \text{tr}\left(\mathbf{H}^{(\text{mrr})}\right) + (n-1)\dfrac{\text{SSE}_{\text{tot}} - \text{SSE}_{\lambda}}{\text{SSE}_{\text{tot}}}} \qquad (2.D.1.1)$$

where $\mathbf{H}^{(\text{mrr})}$ is the MRR HAT matrix, SSE_{tot} is the total sum of squares of the data, and SSE_{λ} is the sum of squares error resulting from an MRR fit with a given value of λ.

2.D.2 Summary

The Model Robust Regression technique is very effective in offering a regression estimate that is robust to misspecification of the user's model. As mentioned in Chapter 1, this research focuses on processes in which it is of interest to estimate both the process mean and the process variance simultaneously. When no replication is present, the residuals from the means fit are often used as building blocks for a variance regression model. Thus, it is vitally important to estimate the means model with as little lack of fit as possible. The MRR procedure will be our method of choice for estimation in the means model setting so that we can offer model-misspecification-free residuals to the variance model. In the next chapter, our focus turns to the subject of estimating process variability.