Chapter 3 : Estimation of Variance

3.A Introduction

A foundational assumption of classical regression is that the model errors have constant variance. If this assumption is violated, the data are said to be heteroscedastic and the regression analysis must be adjusted accordingly. In many data sets, heteroscedasticity is present simply due to the underlying distribution of the random errors. For example, in dose / response data, the response is a Bernoulli random variable :

$$y_i = \begin{cases} 1 & \text{if subject } i \text{ responds} \\ 0 & \text{if subject } i \text{ does not respond} \end{cases} \qquad (3.A.1)$$

A typical model for the process mean in such settings would then be the logistic model :

$$E(y_i) = p_i = \frac{1}{1 + e^{-(\beta_o + \beta_1 \log(dose_i))}} \qquad (3.A.2)$$

where $dose_i$ is the dose given to the i^{th} subject. It follows then, from the definition of a Bernoulli random variable, that $Var(y_i) = \sigma_i^2 = p_i(1 - p_i)$. Thus, the variance of y_i changes with the regressors as a function of the mean. With the exception of normal random variables, this "natural heteroscedasticity" is present for all random variables with probability density functions that are members of the exponential family. Since the variance depends on the regressors solely through the mean, variance estimation occurs simultaneously with the estimate of the mean. For purposes of this research, we will not concern ourselves with this type of variability but will instead investigate variance which occurs above and beyond the natural variance. This type of variability is often termed overdispersion.

In this research we assume that the random errors follow a normal distribution. In the case of normal errors, the natural variance is taken to be one with constant overdispersion, σ^2. In this research, when we speak of heteroscedasticity, we are referring to variance which is present due to some phenomenon in our process and not variance which is naturally present due to the assumed error distribution.

Transformations

A popular method of accounting for non-constant variance is to transform the data in such a way as to stabilize the variances. The analysis is then conducted in the usual sense but on the set of transformed data. For details concerning the use of data transformations for variance stabilization, see Atkinson (1985) and Myers (1990). There are two general criticisms of data transformations. First, in many cases, no simple transformation exists which adequately stabilizes

the variances. Second, and perhaps more importantly, the use of data transformations sheds no light on the behavior of the variance as a function of relevant regressor variables. This has important implications in the area of quality control where the experimenter is interested in knowing where process variability is high and which factors influence variance. Thus, when heteroscedasticity is present, we not only want to account for its presence, but we would also like to gain some insight regarding its structure.

Replication

When heteroscedasticity is not "swept away" using a data transformation, it is typically accounted for in a "weighted" analysis where the "weights" are taken to be the reciprocals of the individual variances. In practice, however, the variances are rarely known and must be estimated. A common suggestion is to estimate the variances based on replication, where the experimenter takes *m* replicates at each data point and estimates the variances with the corresponding sample variances. Notice that in computing *n* different sample variances, the researcher has some feel for where the variance is high or low but still has no sense of structure for process variability. The use of replication for variance estimation is generally regarded as unsatisfactory for a variety of reasons. First, in many regression settings, the values of the regressor variables are often observed and cannot be replicated. Second, unless the number of replicates *m* is large $(m > 10)$, the sample variances are quite unstable and their use as estimates for the weights causes more problems than if heteroscedasticity were ignored all together (Williams,(1975)). Furthermore, work by Deaton, Reynolds and Myers (1983) suggests that, unless the heteroscedasticity is severe, $s_{max}/s_{min} > 3$ (where s_{min} and s_{max} are respectively the minimum and maximum sample standard deviations), an unweighted regression analysis is preferable to an estimated weighted analysis.

Variance Modeling

An appealing alternative to replication is to assume that the heteroscedasticity is present due to some underlying, smooth, variance function. Viewing the process variability in this manner leads naturally to the concept of using regression techniques to estimate this function. In constructing a model for the variance, the building blocks are taken to be the residuals from an unweighted fit to the means model. The intuition proceeds as follows. If we assume $y_i = h(\mathbf{x}_i) + \varepsilon_i$, then we know that $y_i - h(\mathbf{x}_i) = \varepsilon_i$ and $y_i - \hat{h}(\mathbf{x}_i) = \hat{\varepsilon}_i = e_i$. Thus, assuming the correct model ($h(\mathbf{x}_i)$) has been fit, any deviations from the fit ($\hat{y}_i$) and what is observed (y_i), are due to random error and not model misspecification. The random error is estimated by e_i and the variability in the e_i's represents the variation in the process. Thus, regression methods for variance estimation involve regressing some function of the residuals on an "appropriate" variance function. For example, one popular regression model involves regressing the squared residuals on an exponential variance function :

$$e_i^2 = \exp\left\{\mathbf{z}_i'\boldsymbol{\theta}\right\} + \eta_i \qquad (3.A.3)$$

where $\mathbf{z}_i$ is (q x 1) and η_i is the random error term. In this model, the vector $\mathbf{z}_i$ is composed of variables which influence the variance. These variables are referred to as "variance regressors" and may include all, some, or none of the variables which affect the process mean.

When we view the process variance in the same light as the process mean: a phenomenon which changes smoothly and systematically as regressors are manipulated, the process is said to be described in terms of a "dual model". Dual modeling is appealing due to the flexibility which is afforded. Specifically :

1. Replication is not required to estimate the variances at the *n* data points.
2. The researcher can study variables which affect the process mean as well as variables which affect the process variance. This flexibility is extremely useful in quality control where it is important to have knowledge about variables which affect process variability. For details concerning applications to quality control see Box and Meyer (1986), Nair and Pregibon (1988) and Taguchi and Wu (1980).

Written in its most general form, the dual model is given as follows :

(1) Means Model $\quad y_i = h(\boldsymbol{x}_i) + g(\boldsymbol{z}_i)^{1/2}\,\varepsilon_i \qquad (3.A.4)$

where the ε_i are the random errors,
$h(\boldsymbol{x}_i)$ is the process mean function,
and $g(\boldsymbol{z}_i)$ is the process variance function.

(2) Variance Model $\quad \sigma_i^2 = g(\boldsymbol{z}_i) \qquad (3.A.5)$

In chapters 1 and 2 several approaches to regression function estimation were discussed within the context of a single model. That single model represented the process mean and was written as:

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

If *h* exists in a truly parametric form then parametric regression was described to be the method of choice. Conversely, if *h* has no (or very little) parametric structure, then a nonparametric procedure such as local linear regression is often best. This dilemma, regarding which approach to implement, is also an issue in dual modeling. The issue however, is perhaps more complex within the setting of a dual model due to the fact that often, estimation of the models is accomplished iteratively. That is, the mean estimate is used to estimate the variance and the variance is used to re-estimate the mean. As a consequence of this interdependency, a poor mean estimate will in turn cause a poor estimate of variance and vice versa. Thus, procedural selection for analysis of the dual model should not be taken lightly. The primary focus of this chapter is to provide an overview of the common approaches to dual model estimation .

3.B Dual Model Function Estimation

When describing the process of interest in terms of a dual model, there are four basic scenarios to consider. First, if the researcher is confident that the mean and variance functions can be expressed parametrically, the dual model is analyzed from a purely parametric point of view. Secondly, on the opposite side of the coin, situations arise in which neither the mean or variance functions take on a parametric form. In this case, the dual model is analyzed from a strictly nonparametric standpoint. For processes in which one of the two functions, the mean or variance, can be expressed parametrically but the other has no closed form, the dual model is approached using both parametric and nonparametric techniques. The remainder of this section will be devoted to discussing approaches to dual modeling within each of these four contexts.

3.B.1 Parametric Dual Modeling

When the user assumes that the mean and variance functions can be expressed adequately with parametric functions, the dual model can be written:

(1) Means Model $$y_i = h(\boldsymbol{x}_i;\boldsymbol{\beta}) + v(\boldsymbol{z}_i;\boldsymbol{\theta})^{1/2}\,\varepsilon_i \qquad (3.B.1.1)$$

(2) Variance Model $$T\left(e_i\right) = v(\boldsymbol{z}_i;\boldsymbol{\theta}) + \eta_i\,. \qquad (3.B.1.2)$$

Regarding the notation, $\boldsymbol{\beta}$ is a $(k+1)$ vector of means model parameters, $\boldsymbol{\theta}$ is a $(l+1)$ vector of variance model parameters, $T\left(e_i\right)$ is an appropriate function of residuals from the means model fit, η_i is the random error term in the variance model, and $h(\boldsymbol{x}_i;\boldsymbol{\beta})$ denotes the parametric means model function. It is assumed that $h(\boldsymbol{x}_i;\boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}$ throughout this research. The means model fit is obtained by estimated weighted least squares and can be written in matrix notation as

$$\begin{aligned}\hat{\mathbf{y}}^{(\text{ewls})} &= \mathbf{X}\hat{\boldsymbol{\beta}}^{(\text{ewls})} \qquad (3.B.1.3)\\ &= \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}\\ &= \mathbf{H}^{(\text{ewls})}\mathbf{y}\end{aligned}$$

where

$$\mathbf{H}^{(\text{ewls})} = \mathbf{X}\left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}. \qquad (3.B.1.4)$$

Regarding notation, $\hat{\mathbf{V}} = \text{diag}\left\langle \hat{\sigma}_1^2, \cdots, \hat{\sigma}_n^2 \right\rangle$ where $\hat{\sigma}_i^2$ is estimated through the analysis of the model given in (3.B.1.2). The i^{th} means model residual is given by

$$e_i^{(ewls)} = y_i - \hat{y}_i^{(ewls)} \qquad (3.B.1.5)$$

and it is these residuals that serve as the building blocks for the estimation of process variance. Before discussing the variance estimation procedure, it is important to consider the function $T\left(e_i^{(ewls)}\right)$, which represents the i^{th} response in the variance model regression.

The Choice of $T\left(e_i\right)$

If one expects to obtain an accurate estimate of the variance through modeling, it is pertinent that the right data be used to do the modeling. In normal theory regression, the function of the means model residuals that is generally agreed to yield an "appropriate" response for variance modeling is $T\left(e_i^{(ewls)}\right) = e_i^{2\,(ewls)}$. The rationale is simple and proceeds as follows. We have, assuming the correct model has been fit to the mean, that

$$e_i^{(ewls)} \sim N\left(0\,,\ (1 - h_{ii}^{(ewls)})^2\sigma_i^2 + \sum_{j \neq i}^{n} h_{ij}^{2\,(ewls)}\sigma_j^2\right) \qquad (3.B.1.6)$$

where $h_{ij}^{(ewls)}$ is the (i, j) element of $\mathbf{H}^{(ewls)}$. Assuming that the $h_{ij}^{(ewls)}$'s are small (see Cook and Weisberg (1983)), we have that

$$e_i^{2(ewls)} \Big/ \left(1 - h_{ii}^{(ewls)}\right)^2 \ \dot{\sim}\ \sigma_i^2\,\chi_{(1)}^2. \qquad (3.B.1.7)$$

Thus, apart from leverage, we have $E\left(e_i^{2(ewls)}\right) \doteq \sigma_i^2$ and thus $T\left(e_i^{(ewls)}\right) = e_i^{2\,(ewls)}$ becomes an obvious choice for the form of the function T. It should be noted, however, that the use of squared residuals as variance model responses sets the variance model up to be very susceptible to outliers. If an outlier exists in the data set, then this will show up as a large residual in the means model fit and a large residual becomes even more "outlying" when it is squared. This point was first noted by Cohen et al. (1984). Davidian and Carroll (1987) suggest the use of $T\left(e_i^{(ewls)}\right) = \left|e_i^{(ewls)}\right|$ as responses in the variance model when one suspects the presence of outliers in the data.

Estimation of θ

After choosing the appropriate data with which to estimate the variance, it becomes necessary to estimate the parameter vector θ in the variance model:

$$e_i^2 = v(\mathbf{z}_i;\theta) + \eta_i \ . \qquad (3.B.1.8)$$

However, before θ can be estimated, it is important to specify the form of the regression function, $v(\mathbf{z}_i;\theta)$. The function v can assume a wide variety of forms, the most popular of which are listed below in Table 3.B.1.1 :

Table 3.B.1.1 Common Variance Functions

Name	$v\ (\ h(\boldsymbol{x}_i),\ \boldsymbol{z}_i\ ,\ \theta)$
log - linear	$e_i^{2\,(ewls)} = \exp\left\{\boldsymbol{z}_i'\,\theta\right\}$
power-of-the-mean	$e_i^{2\,(ewls)} = \theta_o\ h(\boldsymbol{x}_i)^{\theta_1}$
inverse	$e_i^{2\,(ewls)} = 1 \Big/ \boldsymbol{z}_i'\,\theta$
expanded power-of-the-mean	$e_i^2 = \theta_o + \theta_1\left[h(\boldsymbol{x}_i)\right]^{\theta_2}$

Regarding notation, the variance function in the table heading is written as $v\ (\ h(\boldsymbol{x}_i),\ \boldsymbol{z}_i\ ,\ \theta\ )$ where $h(\boldsymbol{x}_i)$ is the function underlying the process mean. Also, in the last three models mentioned in Table 3.B.1.1, one could replace $e_i^{2\,(ewls)}$ with $\log(\ e_i^{2\,(ewls)}\ )$ to insure a positive valued variance estimate. The appropriate functional form of v is generally decided upon based on either prior experience or through construction of a scatter plot. The squared residuals are plotted against the values of the variance regressors to determine v as a function of $\mathbf{z}_i$ and the squared residuals are plotted against the predicted values (the $\hat{y}_i^{2\,(ewls)}$) to determine v as a function of the mean.

Once the form of the regression function has been decided upon, attention is focused on the estimation of the parameters. There are a number of methods that can be used to estimate θ. One method is the "maximum likelihood type" of an approach that utilizes the fact that $e_i^{2\,(ewls)} \Big/ \sigma_i^2 \ \sim\ \chi^2_{(1)}$. The terminology "maximum likelihood type" is used because the

distribution of $e_i^{2\,(ewls)} \big/ \sigma_i^2$ is only approximately Chi-Square. Using a variable transformation, it is easy to show that the approximate distribution of the i^{th} squared means model residual is given by

$$f\left(e_i^{2\,(ewls)}\right) \approx \exp\left\{-ln\left(e_i^{(ewls)}\right) - ln\,\sigma_i - e_i^{2\,(ewls)} \big/ 2\sigma_i^2\right\} \qquad (3.B.1.9)$$

and the log-likelihood of the squared means model residuals is approximately

$$lnL\left(e_1^{2\,(ewls)}, \cdots, e_n^{2\,(ewls)}\right) \approx -\sum_{i=1}^{n} ln\left(e_i^{(ewls)}\right) - \sum_{i=1}^{n} ln\,\sigma_i - \sum_{i=1}^{n} e_i^{2\,(ewls)} \big/ 2\sigma_i^2 . \quad (3.B.1.10)$$

Assuming that $\sigma_i^2 = v(\mathbf{z}_i;\boldsymbol{\theta})$, and differentiating the likelihood expression given in (3.B.1.10) with respect to $\boldsymbol{\theta}$, the following score equations result:

$$\sum_{i=1}^{n}\left[e_i^{2\,(ewls)} - v(\mathbf{z}_i;\boldsymbol{\theta})\right] v^{-2}(\mathbf{z}_i;\boldsymbol{\theta})\, \partial v(\mathbf{z}_i;\boldsymbol{\theta}) \big/ \partial\boldsymbol{\theta} = \mathbf{0}, \qquad (3.B.1.11)$$

where $\mathbf{0}$ is a vector of zeros. It is interesting to note that these are the same score equations which result from approaching the estimation of $\boldsymbol{\theta}$ through weighted least squares :

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \left(e_i^{2\,(ewls)} - v(\mathbf{z}_i;\boldsymbol{\theta})\right)^2 \big/ v^2(\mathbf{z}_i;\boldsymbol{\theta}) . \qquad (3.B.1.12)$$

Often, the score equations given in (3.B.1.11) are nonlinear in $\boldsymbol{\theta}$, requiring a nonlinear algorithm such as the Gauss-Newton procedure to determine a solution.

Another popular approach to variance estimation is the pseudo-likelihood procedure. The pseudo-likelihood estimates of $\boldsymbol{\theta}$ are obtained by maximizing the normal log-likelihood, $l(\boldsymbol{\beta},\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, where :

$$l(\boldsymbol{\beta},\boldsymbol{\theta}) = -\frac{n}{2}\ln(2\boldsymbol{\pi}) - \sum_{i=1}^{n} \ln\left(v(\mathbf{z}_i;\boldsymbol{\theta})\right) - \frac{1}{2}\sum_{i=1}^{n} (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2\, v^{-1}(\mathbf{z}_i;\boldsymbol{\theta}) \qquad (3.B.1.13)$$

and $\boldsymbol{\beta}$ is assumed to be known. It is easy to show that when v does not depend on the mean, the pseudo-likelihood estimate of $\boldsymbol{\theta}$ is asymptotically equivalent to WLS and maximum likelihood estimates (based on squared residuals) of $\boldsymbol{\theta}$ (Davidian and Carroll (1987)). Pseudo-likelihood

is appealing because it provides a method of estimation in the dual model which is based on a single objective function, $l(\boldsymbol{\beta},\boldsymbol{\theta})$. This approach was studied by Aitkin (1987), who showed that, when the variance does not depend on the mean, the values of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ which simultaneously maximize (3.B.1.13) are asymptotically equivalent to individual WLS estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. However, when the variance does depend on the mean, Jobson and Fuller (1980) suggested using individual WLS techniques for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$.

The Parametric Generalized Least Squares (GLS) Algorithm

Estimation of parameters in the parametric dual model takes place within the context of a generalized least squares (GLS) algorithm. The GLS algorithm is given as follows :

1. Let $\hat{\mathrm{V}} = \mathrm{diag}\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_n^2\right)$ where $\hat{\sigma}_i^2 = 1$.
2. Using estimated weighted least squares, obtain estimates of the parameters in the means model : $\hat{\boldsymbol{\beta}}_{\mathrm{o}}^{(\mathrm{ewls})} = \left(\mathrm{X}' \hat{\mathrm{V}}^{-1} \mathrm{X}\right)^{-1} \mathrm{X}' \hat{\mathrm{V}}^{-1} \mathbf{y}$
3. From the estimates found in step 2, form the squared residuals, given as : $\mathrm{e}_i^{2\,(\mathrm{ewls})} = \left(\mathrm{y}_i - \hat{\mathrm{y}}_i^{(\mathrm{ewls})}\right)^2$ where $\hat{\mathrm{y}}_i^{(\mathrm{ewls})} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{\mathrm{o}}^{(\mathrm{ewls})}$.
4. Take the squared residuals from step 3 and use them as responses in a given variance model : $\mathrm{e}_i^{2\,(\mathrm{ewls})} = \nu\left(\mathbf{z}_i ; \boldsymbol{\theta}\right) + \eta_i$
5. From the regression model in step 4, obtain the estimates of the variance model parameters, $\boldsymbol{\theta}$, using an appropriate estimation technique.
6. Return to step 2 with $\hat{\sigma}_i^2$.
7. Cycle through steps 2-6 until convergence of the means model parameters, at which time $\hat{\boldsymbol{\beta}}^{(\mathrm{ewls})} = \hat{\boldsymbol{\beta}}^{(\mathrm{wls})}$.

This algorithm is appealing for a couple of reasons. First, it allows for simultaneous estimation of the parameters in both the mean and variance models. Second, the estimates obtained for the parameters in $\boldsymbol{\beta}$ are asymptotically equivalent to the estimates of $\boldsymbol{\beta}$ if the true variances were known (i.e. at convergence, $\hat{\boldsymbol{\beta}}^{(\mathrm{wls})} = \hat{\boldsymbol{\beta}}^{(\mathrm{gls})}$). It is important to note that the parametric dual modeling procedure's success hinges largely on the ability of the user to specify the correct mean and variance functions. Since the mean and variance are estimated simultaneously through the generalized least squares algorithm, if either the mean or variance models are misspecified over part of the data, parameter estimates and inferences on those estimates could be effected drastically.

3.B.2 Parametric Mean / Nonparametric Variance

In many situations, the researcher may be confident that the underlying structure of the mean can be described parametrically but the researcher may not know what kind of structure exists for the variance function. In this case the dual model is analyzed using both parametric and nonparametric procedures. This scenario was first considered by Carroll (1982) who assumed a one-regressor, linear model for the mean. The mean function is estimated parametrically through estimated weighted least squares but the variance is estimated nonparametrically. Since no parametric form is specified for the variance, Carroll suggests fitting the squared residuals from the means fit using kernel regression. The dual model is then estimated using the following GLS algorithm :

1. Let $\hat{\mathrm{V}} = \mathrm{diag}\left(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \cdots, \hat{\sigma}_n^2\right)$ where $\hat{\sigma}_i^2 = 1$.
2. Using estimated weighted least squares, obtain estimates of the parameters in the means model : $\hat{\boldsymbol{\beta}}^{(\mathrm{ewls})} = \left(\mathrm{X}' \hat{\mathrm{V}}^{-1} \mathrm{X}\right)^{-1} \mathrm{X}' \hat{\mathrm{V}}^{-1} \mathbf{y}$
3. From the estimates found in step 2, form the squared residuals, given as: $\mathrm{e}_i^{2\,(\mathrm{ewls})} = \left(\mathrm{y}_i - \hat{\mathrm{y}}_i^{(\mathrm{ewls})}\right)^2$ where $\hat{\mathrm{y}}_i^{(\mathrm{ewls})} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}^{(\mathrm{ewls})}$.
4. Estimate the underlying variance function by finding a kernel fit to the squared residuals, obtaining $\hat{\sigma}_i^{2\,(\mathrm{ker})} = \hat{g}\left(\mathbf{z}_i\right)$, where $\hat{g}\left(\mathbf{z}_i\right)$ is the kernel fit at the i^{th} value of the variance regressor.
5. Repeat Step 2, using the estimates from Step 4.

Carroll and Ruppert (1988) suggested that the algorithm be iterated until convergence is met for the parameters in the means model.

The GLS algorithms presented in the previous two sections rely heavily on the assumption that we have fit the correct model for the process mean. If the means model specified is insufficient, the residuals no longer contain information solely pertaining to process variance. These residuals are now "contaminated with lack of fit" (an expression to be explained fully in Chapter 4) and their use in variance modeling creates substantial bias problems when estimating the variance function. This problem was studied by Gasser, Sroka and Jennen-Steimetz (1987) who considered variance estimation in modeling human height growth from birth to adulthood. It is well known that parametric models are insufficient in this problem due to the presence of a growth spurt around seven years of age. This "spurt" causes a "wiggle" in the data that a parametric models cannot capture. The bias in the means fit produces biased residuals and consequently, a biased estimate of variance. Thus, Gasser, Sroka and Jennen-Steinmetz motivated the need for an estimate of variance which does not require an estimate of the process mean. This idea leads us to the next two dual model settings.

3.B.3 Nonparametric Dual Modeling

When the researcher is unable to express the mean or variance functions in a closed form, the two functions must be estimated nonparametrically. The nonparametric dual model is expressed as

$$\begin{aligned} &\text{Means Model :} \quad y_i = h(\mathbf{x}_i) + g^{1/2}(\mathbf{z}_i)\,\varepsilon_i \\ &\text{Variance Model :} \quad \sigma_i^2 = g(\mathbf{z}_i). \end{aligned} \tag{3.B.3.1}$$

The means model is generally estimated using one of the nonparametric techniques discussed in Chapter 2. Nonparametric variance estimation can be approached in a variety of ways. As mentioned in the last section, Gasser, Sroka and Jennen-Steinmetz proposed a method for nonparametric variance estimation that is independent of the fit to the means model.

The method of Gasser, Sroka and Jennen-Steinmetz is based on the concept of "pseudo-residuals", denoted as $\tilde{e}_i$. Pseudo-residuals are obtained by taking continuous triples of design points, x_{i-1}, x_i, and x_{i+1}, joining the outer two observations by a straight line and then computing the difference between the straight line and the middle observation, y_i. For example, the pseudo-residual at x_i is given as :

$$\begin{aligned} \tilde{\varepsilon}_i &= \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} Y_{i-1} + \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} Y_{i+1} - Y_i \\ &= a_i Y_{i-1} + b_i Y_{i+1} - Y_i. \end{aligned} \tag{3.B.3.2}$$

Variance estimation based on these $\tilde{\varepsilon}_i$ is often referred to as a "difference-based" scheme since the pseudo-residual is computed by taking the difference between the point y_i and the straight line connected by the two adjacent observations, y_{i-1} and y_{i+1}. They state that $c_i^2\,\tilde{\varepsilon}_i^2$, where $c_i^2 = a_i^2 + b_i^2 + 1$, is an asymptotically unbiased point estimate of σ_i^2. However, the variability of this point estimate is described as "untolerably" high and in order to reduce the variability, they suggest smoothing the individual point estimates.

Müller and Stadtmüller (1987) extend the ideas of Gasser, Sroka and Jennen-Steinmetz and estimate the local variances by considering continuous neighborhoods of m points $(m \geq 3)$. Their difference-based scheme involves defining a set of weights $\left(\omega_{-m_1}, \cdots, \omega_{-1}, \omega_0, \omega_1, \cdots, \omega_{m_2}\right)$, where $m_1 + m_2 + 1 = m$, for a given set of non-negative integers m_1 and m_2 satisfying : $\sum_{j=-m_1}^{m_2} \omega_j = 0$ and $\sum_{j=-m_1}^{m_2} \omega_j^2 = 1$. Initial local variance estimates, which are described as highly variable, are given as :

$\hat{\sigma}_i^2 = \left(\sum_{j=-m_1}^{m_2} \varpi_j Y_{i+j} \right)^2$. To reduce the variability, these initial estimates are smoothed using kernel regression.

Perhaps the most significant contribution of Müller and Stadtmüller is their suggestion of how to incorporate the variance estimates into the dual model. We have already discussed in Chapter 2 that when the process mean cannot be written in closed form, the typical estimation procedure is local linear regression. In our discussion, the kernel weights used in local linear regression were based on a global bandwidth (i.e. the same bandwidth is used for all data points). Müller and Stadtmüller suggest that the local linear procedure can be improved by using a local smoothing parameter which is constructed in such a way as to take into account the heterogeneity of variance. The idea is to smooth more (larger bandwidth) for regions of high variance and to smooth less (smaller bandwidth) where the data is less variable.

Estimation of the nonparametric dual model can occur in the following two steps :

1. Estimate σ_i^2 using a difference based method as proposed by Gasser, Sroka and Jennen-Steinmetz or Müller and Stadtmüller .
2. Using $\hat{\sigma}_i^2$ from Step 1, obtain the local linear fit for the mean, given as:

$$\hat{y}_i^{(\text{llr})} = \mathbf{x}_i' \left(\mathbf{X}' \mathbf{W}^{(\text{llr})}(\mathbf{x}_i) \mathbf{X} \right)^{-1} \mathbf{X}' \mathbf{W}^{(\text{llr})}(\mathbf{x}_i) \mathbf{y}_i$$

where $\mathbf{W}^{(\text{llr})}(\mathbf{x}_i) = \begin{bmatrix} h_{i1}^{(\text{ker})} & & & 0 \\ & h_{i2}^{(\text{ker})} & & \\ & & \ddots & \\ 0 & & & h_{in}^{(\text{ker})} \end{bmatrix}$,

and $h_{ij}^{(\text{ker})} = \dfrac{K\left(\dfrac{x_i - x_j}{b_i} \right)}{\sum_{j=1}^{n} K\left(\dfrac{x_i - x_j}{b_i} \right)}$ where b_i is the bandwidth used for data point $\mathbf{x}_i$ and is a function of $\hat{\sigma}_i^2$.

It should be noted that the **X** matrix used in writing the local linear hat matrix is not the same **X** matrix used in the parametric model since the i^{th} row of **X** in the parametric model is given by $\mathbf{x}_i' = \left(1 \; x_i \; x_i^2 \cdots x_i^d\right)$, whereas the i^{th} row of **X** in the nonparametric model is given by $\mathbf{x}_i' = \left(1 \; x_i\right)$.

The use of local bandwidths as a means for accounting for variance heterogeneity in nonparametric regression is a fairly new concept. Although the use of local bandwidths in the presence of heteroscedasticity has been shown to provide asymptotically superior results to local linear estimates using a global bandwidth, small sample comparisons is an area that needs more consideration. For more discussion on the use of local bandwidths, see Fan and Gijbels (1992 and 1995) and Ruppert (1996).

A note worth mentioning is that the difference based variance estimates are also useful in estimated weighted least squares when a parametric means model is proposed. However, if the parametric means model is valid, variance estimation done by regressing the squared residuals on a set of variance regressors is the best route for estimation. The reason is that purely nonparametric estimates of variance, which ignore information about the mean (i.e. difference based methods), are more biased and have larger variance than their parametric counterparts. However, when the means model is misspecified, the difference-based methods of variance estimation are more appealing.

One other approach to the nonparametric dual model was considered by Hall and Carroll (1988). Hall and Carroll fall back on the traditional means of variance estimation in which estimation occurs once the location effect is removed. They propose to estimate the variance function by smoothing the squared residuals from the nonparametric fit. Thus, their approach to dual model estimation can be summarized in the following three steps :

1. Using local linear regression, estimate the process mean function as:

$$\hat{y}_i^{(\mathrm{llr})} = \mathbf{x}_i'\left(\mathbf{X}'\mathbf{W}^{(\mathrm{llr})}(\mathbf{x}_i)\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^{(\mathrm{llr})}(\mathbf{x}_i)\mathbf{y}_i \text{ , where}$$

$$\mathbf{W}^{(\mathrm{llr})}(\mathbf{x}_i) = \begin{bmatrix} h_{i1}^{(\mathrm{ker})} & & & 0 \\ & h_{i2}^{(\mathrm{ker})} & & \\ & & \ddots & \\ 0 & & & h_{in}^{(\mathrm{ker})} \end{bmatrix}, \quad \text{and}$$

$$h_{ij}^{(\mathrm{ker})} = \frac{K\left(\frac{x_i - x_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{b}\right)}.$$

2. Form $e_i^{2\,(\mathrm{llr})} = \left(y_i - \hat{y}_i^{(\mathrm{llr})}\right)^2$.

3. Obtain $\hat{\sigma}_i^2$ by performing a local linear fit to the squared residuals from Step 2 : $\hat{\sigma}_i^2 =$

$$\mathbf{z}_i'\left(\mathbf{Z}'\mathbf{W}^{(\mathrm{llr_e})}(\mathbf{z}_i)\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{W}^{(\mathrm{llr_e})}(\mathbf{z}_i)\,\mathbf{e}^{2(\mathrm{llr})},$$ where

$$\mathbf{W}^{(\mathrm{llr_e})}(\mathbf{z}_i) = \begin{bmatrix} h_{i1}^{(\mathrm{ker_e})} & & & 0 \\ & h_{i2}^{(\mathrm{ker_e})} & & \\ & & \ddots & \\ 0 & & & h_{in}^{(\mathrm{ker_e})} \end{bmatrix}$$ is the i^{th} row of the Kernel Hat matrix used for fitting the n squared residuals, given in vector notation as $\mathbf{e}^{2\,(\mathrm{llr})} = \begin{bmatrix} e_1^{2\,(\mathrm{llr})} \\ e_2^{2\,(\mathrm{llr})} \\ \vdots \\ e_n^{2\,(\mathrm{llr})} \end{bmatrix}$, and

$$h_{ij}^{(\mathrm{ker_e})} = \frac{K\left(\frac{z_i - z_j}{b_e}\right)}{\sum_{j=1}^{n} K\left(\frac{z_i - z_j}{b_e}\right)}.$$

It is interesting to note that Hall and Carroll, unlike Müller and Stadtmüller , do not suggest using the variance estimates to obtain local bandwidths for re-estimation of the process mean.

3.B.4 Nonparametric Mean / Parametric Variance

The only dual model scenario that has not been discussed is the situation in which the researcher has no inclination about the structure of the mean but is confident in a parametric structure for the variance. Little, if any, literature discusses this type of situation. Traditionally, variance estimation occurs after the location effect has been removed and thus, if the researcher is not confident in a parametric model for the mean, it is doubtful that confidence will be expressed in a parametric form for the variance. If this scenario were to arise in practice, the logical method for estimation would involve a nonparametric estimate of the mean, such as local linear regression, and then the squared local linear residuals would be regressed in the specified variance model. Conceivably, the variance estimates could then be taken back to re-estimate the process mean, using local bandwidths as suggested by Müller and Stadtmüller (1987).

3.C Which Dual Modeling Method Is Best ?

In this chapter, we have presented four general approaches to estimating the mean and variance functions of a dual model. When the user knows the parametric forms of the underlying mean function, h , and variance function, g , a purely parametric procedure should be used. If the

user is only confident in the parametric form of one of the dual model functions, the dual model should be approached as discussed in Sections 3.C.2 and 3.C.4. When the researcher has no inclination regarding the underlying structures of either the mean or variance functions, a strictly nonparametric approach, such as the ones discussed in Section 3.C.3, should be employed. Selecting a procedure for function estimation in the dual model only becomes a dilemma when the researcher has some ideal about the parametric form(s) of h and/or g , but the form(s) are not adequate across the entire span of the data.

This research will consider the case in which the user specifies parametric forms for both the mean and variance but these forms are only appropriate for part of the data. If the specified parametric forms for h and g are appropriate across the entire data set, the GLS algorithm provided in Section 3.C.1 yields optimal (UMVU) parameter estimates. However, this algorithm's success hinges on the assumption that h and g are valid across the entire range of the data. For instance, if the means model is insufficient for portions of the data (i.e. bias problems), the squared residuals used for variance estimation no longer contain information solely pertaining to the process variance. The squared residuals are now contaminated with "lack of fit" and the bias in the means fit directly translates to a biased variance estimate. If the variance model is inappropriate, the wrong weights will be used in the weighted least squares analysis of the means model. These two scenarios lead to the consideration of nonparametric procedures. Specifically, if the means model is inappropriate, a nonparametric means fit would be able to capture the deviations in the data. However, the nonparametric fit to the mean is often too variable and fits the data to closely, consequently the means model residuals are often too meager to contain enough information for an accurate estimate of variance. A nonparametric approach to variance estimation in which the location effect is totally disregarded (i.e. difference based estimates of variance), generally results in a fit which is more biased and more variable than what is necessary.

The proposed research confronts this dilemma by offering a "middle-ground" alternative. A model robust GLS algorithm is developed which combines parametric and nonparametric procedures for estimation of the mean and variance. These semi-parametric methods of mean and variance estimation are designed to utilize any parametric information the user provides about the process models and to detect deviations from these models. The dual model robust procedure is described in detail in the next two chapters.