The Estimation of Error Variance in Nonparametric Regression Models

Yousef F. Alharbi  
Department of Mathematics,  
Faculty of Science,  
Taibah University, Medinah, Saudi Arabia  

Prakash N. Patil  
Department of Mathematics and Statistics,  
Mississippi State University, USA

In this paper a new class of estimators for estimating the error variance is proposed in the setting of the homoscedastic nonparametric regression models. In order to obtain the error variance, we propose to average the product of the residual and response variable as opposed to averaging the squared residuals in the standard residual-based estimators. Definition of a new estimator is provided using kernel smoothing. Then, the asymptotic properties of the proposed methodology are studied in order to assess its goodness. Comparative discussion with that of the existing methods is also provided. The finite sample performance of the proposed estimator is exhibited through simulation studies.

Introduction

Our main aim is to focus on the constant error variance estimation. Suppose the following homoscedastic nonparametric regression model

\[ Y_i = m(x_i) + \varepsilon_i, \text{ for } i = 1,2,\cdots, n, \]

where \( m(x_i) \) is the mean function \( E(Y_i|x_i) \), \( Y_i \)s represent the response variable and \( x_i \)s denote the design points. The errors \( \varepsilon_i \)s are assumed to be independent, identically distributed random variables with zero mean and variance \( \sigma^2 \). Since the knowledge of \( \sigma^2 \) is essential for drawing inference about the mean function, estimation of the error variance is also very important. And therefore many different estimators of the variance in the model (1) or its variant are discussed in the literature.

All existing estimators for error variance can be classified into two different classes. The first class is called the residual-based estimators. This is because it depends on the sum of the squared residuals, which are obtained from a nonparametric regression fit for the mean function. In general, all residual-based estimators can be written in the form,
The Estimation of Error Variance in Nonparametric Regression Models

\[ \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \hat{m}(x_i))^2, \quad (2) \]

where \( \hat{m}(x_i) \) represents the estimated mean function. The estimation of the mean function in equation (2) is usually done by spline or kernel smoothing. For example, by taking \( \hat{m}(x_i) \) to be a kernel-based estimator, Hall and Marron (1990) have proposed the following estimator,

\[ \hat{\sigma}^2_{HM} = \frac{\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{n} w_{ij} Y_j)^2}{(n-2) \sum_{i=1}^{n} w_{ii} + \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}^2}, \quad (3) \]

where \( w_{ij} = K\left(\frac{x_i - x_j}{h}\right) / \sum_{k=1}^{n} K\left(\frac{x_i - x_k}{h}\right) \), for \( 1 \leq i, j \leq n; \) \( h \) is a suitable smoothing parameter and \( K(\cdot) \) is a kernel function. For error variance estimators using spline smoothing, see for instance Reinsch (1967), Carter and Eagleson (1992), Buckley et al. (1988) and Thompson et al. (1991).

The second class of the error variance estimators is the difference-based estimators. The idea of this class is to remove the trend using differencing as one does in the time series analysis. Thus, the estimation of the mean function is not required. The simplest possible estimator in this class, defined by Rice (1984), is

\[ \hat{\sigma}^2 = \frac{1}{2(n-1) \sum_{l=2}^{n}} (Y_i - Y_{i-1})^2. \quad (4) \]

Gasser et al. (1986) defined a second ordered difference-based estimator. Hall et al. (1990) generalize the concept of the difference based estimator to define \( r \)th ordered difference-based estimator where \( r \geq 1 \). The class of the error variance estimators based on the \( r \)th order differences is

\[ \hat{\sigma}^2_{HRT} = (n - r)^{-1} \sum_{j=1}^{n-r} \left( \sum_{k=0}^{r-d} d_k Y_{k+j} \right)^2, \]

where \( d \) is a differences sequence such that \( \sum_{k=0}^{r} d_k = 0, \sum_{k=0}^{r} d_k^2 = 1 \). For more details, see for example Seifert et al. (1993). Another variant of the difference-based estimator, which is not necessary fixed order difference-based estimator, is discussed in Müller et al. (2003) and it is defined as

\[ \hat{\sigma}^2_{MSW} = \frac{1}{n(n-1) \sum_{i \neq j}^{n}} \sum_{i \neq j}^{n} (Y_i - Y_j)^2 w_{ij}, \quad (5) \]

where \( w_{ij} \) is a weight function. A modified version of the estimator in (5) is developed by Tong et al. (2008). Tong and Wang (2005) used the expected value of the lag-k version of the Rice (1984) estimator to estimate the error variance as the intercept of a simple linear model. Park et al. (2009) modified the Tong and Wang estimator using the local quadratic approximation approach.

In this paper a class of the error variance estimators is proposed where one averages \( e_i Y_i \)’s as opposed to averaging \( e_i^2 \)’s as one does in the residual-based estimators. Thus, our proposed estimator can be written as

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i Y_i = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}(x_i)) Y_i. \quad (6) \]

In the classes of estimators in (2) and (6), \( i \)th observation is used to estimate the mean function and again through residual in order to estimate the error variance. The proposed class of error variance estimators in (6) can be redefined as

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{m}(x_i)) Y_i, \quad (7) \]
where \( m_{-i}(x_i) \) denotes the estimate of the mean function without using the \( i \)th observation. For simplicity, our results will be given for an estimator based on kernel smoothing, but the idea can be generalized to any other smoothing techniques such as wavelet and spline approaches. In particular, the simplest estimator, based on kernel smoothing, is

\[
\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} Y_i^2_i - \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij} Y_i Y_j, \quad (8)
\]

where \( w_{ij} = K \left( \frac{x_i - x_j}{h} \right) / \sum_{i 
eq j} K \left( \frac{x_i - x_j}{h} \right), 1 \leq i, j \leq n; h \) is a suitable bandwidth, \( K(\cdot) \) is a suitable kernel function and \( \sum_{j \neq i}^n w_{ij} = 1 \) for each \( i \). Our main interest is to seek the answers to the questions of whether averaging of \( e_i Y_i \) has any advantage over averaging \( e_i^2 \), and whether using, or not using, observation \( Y_i \) in the estimation of \( m(x_i) \) has any advantage.

## The Main Results

### The Asymptotic Results

The following assumptions are made in order to find the asymptotic behaviour of the new estimator in (8)

\[ A1: \quad \int u^t K(u) \, du = \begin{cases} \frac{1}{p!}, & \text{for} \; t = 0; \\ \frac{1}{(p-1)!}, & \text{for} \; t = 1 \ldots p-1; \\ 0, & \text{for} \; t = p. \end{cases} \]

\[ A2: \quad \text{The kernel function } K(\cdot) \text{ is bounded, symmetric around 0, compactly supported on } (-1,1) \text{ and its first derivative exists}. \]

\[ A3: \quad x_i \text{s are chosen randomly from the } U[0,1] \text{ distribution}. \]

\[ A4: \quad \text{The mean function } m(x) \text{ is differentiable, bounded and has } p \text{-continuous derivatives and the fourth moment } \mu_4 = \mathbb{E}(m(x_i) - m(x))^4 \text{ exists}. \]

\[ A5: \quad n \to +\infty, \; \alpha > 0 \text{ such that } nh \to \infty. \]

The bias and variance formulas of \( \hat{\sigma}^2 \) proposed in (8) are given in the following theorem.

**Theorem 2.1** Suppose the assumptions A1, A2, A3, A4 and A5 are satisfied and \( h \sim n^{-\alpha} \), where \( \alpha \) is positive number such that \( 1/3 < \alpha < 1 \), then for \( \hat{\sigma}^2 \) in (8)

\[(i) \quad E(\hat{\sigma}^2) - \sigma^2 = h^p \cdot C_1 + o(h^p) + o(n^{-1}), \]

\[(ii) \quad \text{Var} (\hat{\sigma}^2) = n^{-1} C_2 + n^{-2} h^{-1} C_3 + o(n^2 h^{-1}), \]

where

\[ C_1 = \frac{(-1)^p}{p!} \int K(y)y^p \, dy \int m(t)m^{(p)}(t)\, dt, \]

\[ C_2 = \mu_4 - \sigma^4, \]

\[ C_3 = 2\sigma^4 \int K^2(y)\, dy + 4\sigma^2 \int K^2(y)\, dy \int m^2(x)\, dx \]

and \( m^{(p)}(x) \) is the \( p \) th derivative of \( m(x) \). The asymptotic optimal bandwidth of the estimator in (8) is \( n^{-2/3(p+1)} \).

**Remark:**

According to the results of the above theorem, the mean squared error of \( \hat{\sigma}^2 \) is
The Estimation of Error Variance in Nonparametric Regression Models

\[
MSE(\hat{\sigma}^2) = (E(\hat{\sigma}^2) - \sigma^2)^2 + Var(\hat{\sigma}^2) = n^{-1} C_1 + n^{-2} h^{-1} C_2 + h^2 p C_3^2 + o(n^2 h) + o(h^2 p) .
\]  

(9)

Note that \( C_1 \) represents the contribution of the bias, whereas \( C_2 \) and \( C_3 \) denote the contributions of the variance. Clearly, if the bandwidth \( h \) is chosen as in the above theorem, we obtain

\[
E(\hat{\sigma}^2 - \sigma)^2 = n^{-1} \cdot var(\epsilon^2) + o(n^{-1}) .
\]

(10)

That is, in the first order, the mean squared error of \( \sigma^2 \) is the same as that of the estimators of Hall and Marron (1990), Müller et al. (2003) and Tong et al. (2008). The above optimal rate of MSE is not achieved by the fixed order difference-based estimators such as the estimators of Rice (1984) and the Hall et al. (1990).

The following theorem establishes the asymptotic normality of \( \sigma^2 \).

Theorem 2.2 Under the assumptions A1, A2, A3, A4 and A5, and \( h \sim n^{-\alpha} \), where \( \alpha \) is positive number such that \( \alpha < 1 \)

\[
\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \overset{d}{\to} N(0, \mu_4 - \sigma^4).
\]

It should be noted the asymptotic distributions of \( \hat{\sigma}^2 \) and the Hall and Marron estimator are exactly the same, because \( \mu_4 - \sigma^4 = Var(\epsilon^2) \).

Discussion

The second order kernel function is assumed in this discussion, but the results can be generalized to a kernel of any order. Using equation (9), it is easy to see that the asymptotic optimal bandwidth of the estimator in (8) is \( n^{-2/5} \), whereas the asymptotic optimal bandwidth for estimating the mean function is \( n^{-1/5} \). The asymptotic optimal bandwidth of the Hall and Marron estimator (1990) is \( n^{-2/9} \). That is the optimal bandwidth of the Hall and Marron estimator is close to the optimal bandwidth for estimating the mean function. In other words, the Hall and Marron estimator approximately estimates the mean function optimally. This means that the proposed estimator \( \hat{\sigma}^2 \) estimates the mean function with a smaller bias compared to that of the Hall and Marron estimator. According to Wang et al. (2008), this property is often desirable. Thus, the optimal bandwidth of the proposed estimator is roughly the square of that of the Hall and Marron estimator.

The proposed estimator and the Hall and Marron estimator approximately have the same mean squared error in the first order. However, the Hall and Marron estimator clearly has smaller second order term, in \( MSE(\hat{\sigma}^2_M) \sim n^{-1} [1 + n^{-7/9}] \). Whereas the proposed estimator \( MSE(\hat{\sigma}^2) \sim n^{-1} [1 + n^{-3/5}] \). That is the Hall and Marron estimator has smaller relative error. According to Dette et al. (1998) and Tong et al. (2008), the smallest relative error does not lead to a better finite sample behavior. To illustrate this further, by excluding the constants of the MSEs of the proposed, and the Hall and Marron estimators, we obtain

\[
\frac{MSE(\hat{\sigma}^2_M)}{MSE(\hat{\sigma}^2)} = \frac{n^{-1}[1 + (n^{-7/9})]}{n^{-1}[1 + (n^{-3/5})]} = \frac{1 + (n^{-7/9})}{1 + (n^{-3/5})} .
\]

The above ratio roughly equals to 0.97 for \( n = 100 \). This means that the second order of the mean squared errors of these estimators has a very small effect in their finite sample performances. Thus, the finite sample behaviors of the estimators are more likely to be determined and dominated by the constants \( C_i \)s of the MSEs.

In the difference-based estimators, one may view that the mean function is estimated implicitly with small possible bias by taking \( h \sim n^{-1} \). In general, the bandwidth of the difference-based estimators is smaller than
that of the estimator in (8). Thus, in some sense, the asymptotic optimal bandwidth of the proposed method is in the middle between that of the residual- and differences-based methods.

**Finite Sample Behavior**

Consider the model in (1) where the mean functions are:

(i) \( m_1(x) = 4.7 + 2.4x + 5x^2 + 4.3x^3 \).

(ii) \( m_2(x) = \exp(-29 - 10x) \).

The noise levels are chosen to be \( \sigma^2 = 1, 4, 25 \) and 100 in order to assess their effect on the performance of \( \hat{\sigma}^2 \). In figures (1) and (2), the \( Y_i \)'s are generated by taking design points to be uniform distributed on \([0, 1]\) and the errors to be \( N(0, \sigma^2) \) random variables. Then, the error variance is obtained using the proposed estimator \( \hat{\sigma}^2 \) and the Hall & Marron estimator. These two steps are replicated for 1000 times. From the observed values of the proposed and Hall & Marron estimators, the estimated kernel density functions of these estimator are plotted in figures (1) and (2) corresponding to the models (i) \( Y_i = m_1(x) + \epsilon \) and (ii) \( Y_i = m_2(x) + \epsilon \), respectively. The sample size is selected to be \( n = 200 \) for the first model, while \( n = 100 \) for the second model. The bandwidth of the estimator in (8) is chosen to be square of that of the Hall and Marron estimator as discussed in subsection 3.3.

The numerical results provide a clear verification of the normality for the asymptotic distribution of the proposed estimator. The numerical results show that the means and variances of estimated density functions of the proposed and Hall & Marron estimators, and their asymptotic distribution are approximately the same when the mean function is exponent of a negative function. See for example figure (2). However, for the mean function \( m_1(x) \), the variances of the estimated distributions of the proposed and Hall & Marron estimators, and their asymptotic distribution are different, but the means are approximately the same. In general, we studied the behavior of the considered estimators in several different cases using similar procedures of the simulation studies described in previous paragraph. We can summarized our main findings as follows. When the mean function is a constant, simple linear function or exponent of a negative function, the means and variances of these distributions are roughly the same when \( \int_0^1 m^2(x) \, dx < 1 \) where \( m(x) \) is polynomial regression function of order \( \geq 3 \). However, when \( \int_0^1 m^2(x) \, dx > 1 \), the estimated distributions of the proposed and Hall & Marron estimators, and their asymptotic distribution have approximately the same means, but the variances are different. See for instance figure (1). For detailed discussion, we refer the reader to Alharbi (2013).

Using similar procedures of the simulation studies described in the first paragraph of the current section, the effect of the bandwidth on the finite sample performance of the proposed estimator is also studied. In these simulation studies, the mean function is fixed, whereas the bandwidth is allowed to vary. For more details, see Alharbi (2013). From these simulation studies, we can conclude that the chosen bandwidth has a clear effect on
the finite sample behavior of the proposed estimator. Large choices of the bandwidth increase the bias and decrease the variance, while the opposite occurs for small choices. The interval of the appropriate choices of the bandwidth is narrow for the small level of noises. This interval increases as the error variances rises. One can check this by plotting a chosen values of the bandwidth versus their mean squared error.

**Conclusion**

The proposed estimator with respect to the mean squared error is achieved the optimal rate in the first order, which was described in equation (10). We found that the proposed estimator estimates the mean function with a small bias compared with that of the Hall and Marron estimator. In other words, to obtain the best mean squared error of the estimator in (8), one is not required to estimate the mean function optimally. This property is often desirable according to Wang et al. (2008). We also found that the effect of not using i-th observation in the mean function estimation is negligible, because it does not improve the mean squared error or the finite sample performance of the proposed estimator in (8) significantly. We illustrate that the smaller relative error does not play an important role in the finite sample behavior. According to the results of the simulation studies, when the bandwidth of the proposed and the Hall & Marron estimators are selected appropriately, none of them is preferable across all mean functions and different noise levels.

**Acknowledgment**

The first author wishes to thank Taibah University in Saudi Arabia for financial support.

**References**


Appendix

An Outline of Proof of Theorem 2.1

First, to calculate the bias, we can show that
\[
E(\hat{\sigma}^2) = \sigma^2 + \frac{1}{n} \sum_{i=1}^{n} m^2(x_i) - \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right) m(x_i)m(x_j).
\] (11)

By approximate the summations in the above equation by integration, and then using Taylor series expansion, we obtain
\[
E(\hat{\sigma}^2) = \sigma^2 + \frac{h^p}{p!} \left( \int K(y) y^p \, dy \right) \left( \int m(t) m^{(p)}(t) \, dt \right) + o(h^p) + O(n^{-1})
\]
\[
= \sigma^2 + h^p C_1 + o(h^p) + O(n^{-1}),
\]
where \( C_1 = \frac{(p-1)!}{p!} \left( \int K(y) y^p \, dy \right) \left( \int m(t) m^{(p)}(t) \, dt \right) \). This completes the proof of part (i) of Theorem 2.1. Thus, the squared bias can be written as
\[
(E(\hat{\sigma}^2) - \sigma^2)^2 = h^2 C_1^2 + o(h^2) + o(n^{-2} h^{-1}).
\] (12)

By computing \( E(\hat{\sigma}^2)^2 \) and \( E(\hat{\sigma}^2) \), we can show that
\[
\text{Var}(\hat{\sigma}) = E(\hat{\sigma}^2)^2 - (E(\hat{\sigma}^2))^2
\]
\[-\frac{2}{n^2} \sum_{i=1}^{n} \left[ \mu_4 - \sigma^4 + 4\mu_3 m(x_i) + 4\sigma^2 m^2(x_i) \right]
\]
\[+ \frac{2}{n^2(n-1)^2 h^2} \sum_{i \neq j} K^2\left(\frac{x_i - x_j}{h}\right) \left[ \sigma^2 + \sigma^2 m^2(x_i) + \sigma^2 m^2(x_j) \right]
\]
\[+ \frac{\sigma^2}{n^2(n-1)^2 h^2} \sum_{i \neq j} \sum_{k \neq i,j} K\left(\frac{x_i - x_j}{h}\right) K\left(\frac{x_i - x_k}{h}\right) m(x_j)m(x_k)
\]
\[+ \frac{\sigma^2}{n^2(n-1)^2 h^2} \sum_{i \neq j} \sum_{k \neq i,j} K\left(\frac{x_i - x_j}{h}\right) K\left(\frac{x_k - x_i}{h}\right) m(x_j)m(x_k)
\]
\[+ \frac{\sigma^2}{n^2(n-1)^2 h^2} \sum_{i \neq j} \sum_{k \neq i,j} K\left(\frac{x_i - x_j}{h}\right) K\left(\frac{x_k - x_j}{h}\right) m(x_i)m(x_k)
\]
\[+ \frac{\sigma^2}{n^2(n-1)^2 h^2} \sum_{i \neq j} \sum_{k \neq i,j} K\left(\frac{x_i - x_j}{h}\right) K\left(\frac{x_k - x_i}{h}\right) m(x_i)m(x_k)
\]
\[= \frac{1}{n} \mu_4 - \sigma^4 \]
\[+ \frac{2}{n^3 h} \left[ \sigma^4 \int K^2(y) \, dy + 2\sigma^2 \int K^2(y)dy \int m^2(x) \, dx \right] + o(n^2 h^{-1})
\]
\[= n^{-1} C_2 + n^{-2} h^{-1} C_3 + o(n^{-2} h^{-1})
\]
where \( C_2 = \mu_4 - \sigma^4 \) and \( C_3 = 2\sigma^4 \int K^2(y) \, dy + 4\sigma^2 \int K^2(y)dy \int m^2(x) \, dx \). For the asymptotic optimal bandwidth, it is easy to show that \( h \sim n^{-2/2p+1} \) by solving the equation \( \frac{\partial \text{AMSE}(\hat{\sigma}^2)}{\partial h} = 0 \) where \( \text{AMSE}(\hat{\sigma}^2) = n^{-1} C_2 + n^{-2} h^{-1} C_3 + h^{2p} C_1^2 \). Thus, the proof of Theorem 2.1 is completed. For more details, see Alharbi (2011).

A Sketch of Proof of Theorem 2.2

The estimator in (8) can be expressed as
The Estimation of Error Variance in Nonparametric Regression Models

\[ \hat{\sigma}^2 = \sum_{k=1}^{3} P_k(h) \]

where \( P_1(h) = \frac{1}{n} \sum_{i=1}^{n} \left[ Y_i^2 - b_i \left( Y_i - m(x_i) \right) \right] \), \( b_i = \frac{2}{(n-1)h} \sum_{j \neq i} K \left( \frac{x_i - x_j}{h} \right) \) \( m(x_j) \).

\[ P_2(h) = -\frac{1}{n(n-1)h} \sum_{i \neq j} K \left( \frac{x_i - x_j}{h} \right) U_i U_j, \quad U_i = Y_i - m(x_i) \]

and

\[ P_3(h) = -\frac{1}{n(n-1)h} \sum_{i \neq j} K \left( \frac{x_i - x_j}{h} \right) \left[ m(x_i) m(x_j) \right]. \]

Let \( Z_i = Y_i^2 - b_i \left[ Y_i - m(x_i) \right] \). Thus, the asymptotic normality of the term \( P_1(h) \) can be established using the Lindeberg-Feller central limit theorem, since \( Z_i \) are independent random variables. Therefore, we can verify that

\[ \sqrt{n} \left( P_1(h) - C_4 \right) \overset{d}{\to} N(0, C_2). \]

where \( C_4 = \sigma^2 + \int m^2(u) \, du \) and \( C_2 \) is defined in Theorem 2.1. Thus, the term \( P_1(h) \) can be represented as

\[ P_1(h) = \frac{1}{\sqrt{n}} (\mu_4 - \sigma^4)^{1/2} N_1 + \sigma^2 + \int m^2(x) \, dx \quad (13) \]

and the expected value of the term \( P_2(h) \) is zero and its variance is \( O(n^{-2}) \) = \( o(n^{-1}) \). Thus, we obtain

\[ \sqrt{n} \left( P_2(h) \right) \overset{P}{\to} 0. \quad (14) \]

Therefore, the effect of the term \( P_2(h) \) in the main distribution is negligible compared with that of the term \( P_1(h) \). The term \( P_3(h) \) is a deterministic term and can be approximated as

\[ \frac{1}{n(n-1)h} \sum_{i=1}^{n} \sum_{j \neq i} K \left( \frac{x_i - x_j}{h} \right) \left[ m(x_i) m(x_j) \right] = \int m^2(x) \, dx + O(h^r) + O(n^{-1}). \quad (15) \]

Also note that

\[ \text{Cov} \left[ P_1(h), P_2(h) \right] = 0. \quad (16) \]

So, \( P_k(h) \)s are uncorrelated. Using equations (13), (14) and (15), \( \hat{\sigma}^2 \) in (8) may be expressed as

\[ \hat{\sigma}^2 = \frac{1}{\sqrt{n}} (\mu_4 - \sigma^4)^{1/2} N_1 + \sigma^2 + \int m^2(x) \, dx - \int \int m^2(x) \, dx + O(n^{-1}) + O(h^r). \]

Therefore,

\[ \sqrt{n} \left( \hat{\sigma}^2 - \sigma^2 \right) \overset{d}{\to} N(0, \mu_4 - \sigma^4). \quad (17) \]