Research Report
Department of Statistics
Göteborg University
Sweden

Variance estimates based on knowledge of monotonicity and concavity properties

Ulla Dahlbom

Research Report 1998:7
ISSN 0349-8034

Mailing address:	Fax	Phone	Home Page:
Department of Statistics	Nat: 031-773 12 74	Nat: 031-773 10 00	http://www.stat.gu.se
Göteborg University	Int: +46 31 773 12 74	Int: +46 31 773 10 00	
Box 660			
SE 405 30 Göteborg			
Sweden			

VARIANCE ESTIMATES BASED ON KNOWLEDGE OF MONOTONICITY AND CONCAVITY PROPERTIES

ABSTRACT

In problems dealing with regression functions the choice of model and estimation method is due to a priori information about the regression function. In some situations it is motivated to consider regression functions with specific non-parametric characteristics, for instance monotonicity and/or concavity/convexity. In situations when we only have one y-observation for each x_i we propose two new variance approximation methods, one for curves that fulfil monotonicity restrictions and one for curves that fulfil concavity/ convexity restrictions.

***Key Words**: isoton; non-parametric regression.*

1 INTRODUCTION

In all regression problems the choice of estimation method is due to the a priori information about the regression function. In some applications the shape of the regression function is known and the estimation problem is reduced to an estimation of some parameters. But usually this is not the case. Instead the relationship between the dependent and independent variables is determined by some constraints on the variables. The simplest characteristic is monotonicity. In this case the only assumption about the regression function is, that it has a non-decreasing phase or/and a non-increasing phase. If the regression function consists of one non-decreasing (or non-increasing) phase only a suitable estimation method is isotonic regression. This is a non-parametric regression method that is very often referred to. The basic theory of isotonic regression is described in Barlow et al (1972) and Robertson et al (1989). If on the other hand the regression function consists of two phases - one non-decreasing and one non-increasing - with a known or unknown mode we can use unimodal regression. This estimation method is described in Frisén (1980). Further statistical properties are given in Dahlbom (1986).

Another simple shape characteristic is concavity (or convexity). In many applications it is motivated to consider regression functions with both monotonicity and concavity/convexity restrictions. The concave regression problem was first formulated by Hildreth (1954) for the estimation of marginal productivity curves. He adopted the LS estimation method and formulated it as a quadratic programming problem. Dent (1973) and Holloway (1979) continued the work and the concave regression function was obtained by a more general framework of quadratic programming with linear inequality constraints. The consistency of these LS concave regression estimators has been proved by Hanson and Pledger (1976). Wu (1982) proposed two algorithms for concave regression. One involves some quadratic programming. Unfortunately this algorithm giving the LS solution has no assured convergence. The other is just an approximation, but very easy to implement. An iterative method for regression with convexity and monotonicity restrictions was proposed by Holm and Frisén (1985). This method gives the LS estimate in a finite number of steps. Fraser and Massam (1987) formulated an algorithm using cylinder projections. Theoretically they obtained the LS solution but the estimation method involves a matrix inversion which in practice implies numerical uncertainty in the obtained results. An iterative estimation method which is closely related to the method proposed by Holm and Frisén (1985) was proposed by Dahlbom (1994). This method gives the LS estimate in a finite number of steps. It is also evaluated for more general situations which contain different combinations of monotonicity and concavity/convexity restrictions like increasing concave/convex, decreasing concave/convex, unimodal concave/convex and sigmoid regression functions.

In these situations we might have difficulties with the variance estimation depending on the situation. The estimation problem can be solved in different ways. In many cases we already have estimates of Var(Y) from another analogous survey. In other situations we can do some realistic theoretical assumptions. One assumption that is commonly used is that all y-observations, Y_i, i=1,..,N, have the same variance s^2. If we do not have any a priori information about the variance we can easily estimate if there is more than one y-observation for each x_i. In some situations, however, we have only got one y-observation for each x_i. If we have no a priori knowledge of the variances then we can have a more severe problem. This problem has not been solved before.

In this paper approximation methods have been developed in the situations when the regression function is monotonic and/or concave/convex and when we only have one y-observation for each x_i. Some simple properties of these approximations are also examined by simulation studies. Finally we illustrate the two variance estimation methods by using examples, one for each method.

2 APPROXIMATION OF THE VARIANCE FOR ISOTONIC REGRESSION FUNCTIONS

If the regression function consists of one non-decreasing (or non-increasing) phase only, a suitable estimation method is isotonic regression. An approximation method of the variance of the observations has been developed in this situation using the assumption that $\sigma^2(x_i) = \sigma^2$ for all i. Since some notion in the variance estimation method is used that will need knowledge of the estimation method we start by giving a short general description.

2.1 THE LSE-METHOD OF ISOTONIC REGRESSION

Isotonic regression is used for monotonic increasing or monotonic decreasing regression functions. A least squares estimate of the increasing regression function, $\mu(x_i)$, is obtained according to the following:

Suppose we have a probability distribution for each x_i, i=1,......,N. Also suppose that for each x_i a random sample $\{y_j(x_i), j=1,....,n(x_i)\}$ is available. The number of y-observations can be different for different x_i. The isotonic regression is identical to the regression function obtained from $\bar{y}(x_i)$ with weights

$$w(x_i) = \frac{n(x_i)}{\sum_{i=1}^{N} n(x_i)}.$$

Suppose that the x-observations are ordered, i.e. $x_1 < x_2 < \dots\dots\dots < x_N$. If $\bar{y}(x_i)$ is monotonicly increasing, then the averages themselves are the isotonic regression function. (The case when $\bar{y}(x_1) \geq \bar{y}(x_2) \geq \dots\dots \geq \bar{y}(x_N)$ is treated analogously). If $\bar{y}(x)$ do not fulfil the monotonicity restrictions we compare the averages successively until the restrictions are not fulfilled.

Suppose that the inequality $\bar{y}(x_k) \leq \bar{y}(x_{k+1})$ does not hold. We then estimate these two points according to

$$\bar{y}'(x_k) = \bar{y}'(x_{k+1}) = \frac{n(x_k)\bar{y}(x_k) + n(x_{k+1})\bar{y}(x_{k+1})}{n(x_k) + n(x_{k+1})}$$

Before the procedure is continued we have to check if the monotonicity restrictions between $\bar{y}(x_{k-1})$, $\bar{y}'(x_k)$ and $\bar{y}'(x_{k+1})$ are fulfilled, i.e.

$$\bar{y}(x_1) \leq \dots\dots\dots \leq \bar{y}(x_{k-1}) \leq \bar{y}'(x_k) = \bar{y}'(x_{k+1}).$$

If this is not the case we have to re-estimate the y-value for x_{k-1}, x_k and x_{k+1};

$$\bar{y}''(x_{k-1}) = \bar{y}''(x_k) = \bar{y}''(x_{k+1}) = \frac{n(x_{k-1})\bar{y}(x_{k-1}) + n(x_k)\bar{y}(x_k) + n(x_{k+1})\bar{y}(x_{k+1})}{n(x_{k-1}) + n(x_k) + n(x_{k+1})}$$

Now we have to check if

$$\bar{y}(x_1) \leq \dots\dots\dots \leq \bar{y}(x_{k-2}) \leq \bar{y}''(x_{k-1}) = \bar{y}''(x_k) = \bar{y}''(x_{k+1}).$$

If this is not the case we have to re-estimate the y-values of x_{k-2}, x_{k-1}, x_k and x_{k+1} using the weighted average of the means. This is repeated until the monotonicity restrictions are fulfilled for all the y-values corresponding to x_i, i=1,.....,k+1. We then continue to examine if the inequality $\bar{y}''(x_{k+1}) \leq \bar{y}(x_{k+2})$ holds and so on.

This procedure is repeated until estimates that fulfil the monotonicity restrictions have been obtained for all x_i. These estimates are the LS estimates, $\bar{y}^+(x)$, of $\mu(x)$. As mentioned before properties of isotonic regression are described in Barlow et al (1972) and Robertson et al (1989).

The estimate of a unimodal regression function is obtained in an analogous way. But in this situation we might also have the problem of estimating an unknown mode.

2.2 ESTIMATION OF VAR(Y) FOR ISOTONIC AND UNIMODAL REGRESSION FUNCTIONS

Let $\mu(x)$ be a isotonic (or unimodal) regression function and let Y_i be observations generated from this function. Suppose that $\sigma^2(Y_i) = \sigma^2 < \infty$ for all i. Then the variance can be estimated by

$$\hat{\sigma}_1^2(Y_i) = \frac{1}{N}\sum_{i=1}^{N}\left[y(x_i) - \mu(x_i)\right]^2 \quad i=1,\ldots,N.$$

When $\mu(x)$ is estimated by $\hat{\mu}(x) = \bar{y}^+(x)$ the variance can instead be estimated by

$$\hat{\sigma}_{21}^2(Y_i) = \frac{1}{N-p}\sum_{i=1}^{N}\left[y(x_i) - \hat{\mu}(x_i)\right]^2$$

where p is the number of plateaus in the estimate $\hat{\mu}(x) = \hat{y}^+(x)$, containing one observation or more. The number N-p acts as the degree of freedom because each plateau has a separate mean. However the regression function is not necessarily constant within the plateaus, which means that this is only an approximation. Its accuracy will be studied in the simulation study in section 2.3. For the plateaus with more than one observation the 'slope' of observations go in opposite directions (see lemma 2.3.1). This implies that $\hat{\sigma}_{21}^2(Y_i)$ will underestimate σ^2.

This might motivate the following ad hoc alternative for the estimation of σ^2.

Call the observations in the j:th block $\bar{y}^{+}(x_i)$, $i=m_{j-1}+1,\ldots,m_j$, where $m_0=0$ and $j=1,\ldots,K$. Then for block j, $j=2,\ldots,K-1$, $\hat{\mu}(x_i)$ is estimated by the straight line going through the points

$$\left[\frac{x_{m_{j-1}}+x_{m_{j-1}+1}}{2}, \frac{\bar{y}^{+}(x_{m_{j-1}})+\bar{y}^{+}(x_{m_{j-1}+1})}{2}\right] \quad \text{and} \quad \left[\frac{x_{m_j}+x_{m_j+1}}{2}, \frac{\bar{y}^{+}(x_{m_j})+\bar{y}^{+}(x_{m_j+1})}{2}\right]$$

i.e. $\hat{\mu}(x_i)=a_j+b_jx_i$ where

$$a_j = \frac{\bar{y}^{+}(x_{m_{j-1}})+\bar{y}^{+}(x_{m_{j-1}+1})}{2} - \frac{1}{2}b_j \cdot \frac{1}{m_j}$$

$$b_j = \left(\frac{\bar{y}^{+}(x_{m_j+1})-\bar{y}^{+}(x_{m_{j-1}})}{2}\right)\frac{1}{m_j}.$$

The estimate of $\mu(x)$ corresponding to block number 1 is obtained by letting the y-distance between the first and the second block be equal to the distance between $\bar{Y}_0^{+}$ and the first block. For block number K $\hat{\mu}(x)$ is obtained in an analogous way. The variance estimate obtained by using $\hat{\mu}(x_i)=a_j+b_jx_i$ will overestimate the variance. Denote the variance estimate using $\hat{\mu}(x)=\bar{y}^{+}(x)$ by σ_{21}^2 and using $\hat{\mu}(x_i)=a_j+b_jx_i$ by σ_{22}^2. The proposed estimate σ_g^2 of the variance is the geometric average of σ_{21}^2 and σ_{22}^2, i.e. $\sigma_g^2=\sqrt{\sigma_{21}^2\cdot\sigma_{22}^2}$. The motivation for using the geometric average is that for positive values it is at least as logical to use the geometric average compared to the arithmetic average. Since the logarithms are symmetric around $\sigma^2=1$ we get a better estimate, i.e. closer to 1.

2.3 SOME PROPERTIES OF THE VARIANCE ESTIMATION METHOD FOR ISOTONIC REGRESSION

The estimated isotonic regression function has constant values in blocks of neighbouring points. This common value is equal to the weighted mean of the y-observations in the block. Let a weight function $w(x_i)$, i=1,...,N with positive values be defined. This means that the estimated function can be given as a sum of step functions. Let $A_0 = \{x_1, x_2, \ldots, x_N\}$, where $x_1 < x_2 < \ldots < x_N$, be the set of observed x-values and

$$f_k(x) = \begin{cases} 1 \text{ for } x \geq x_k \\ 0 \text{ for } x < x_k \end{cases} \qquad \text{for} \quad x \in A_0.$$

For each non-decreasing function f(x), $x \in A_0$, there exist constants $a_0, a_1, \ldots, a_N$, with $a_k \geq 0$ for $k \geq 1$, such that $f(x) = a_0 + a_1 f_1(x) + a_2 f_2(x) + \ldots + a_N f_N(x)$ for $x \in A_0$.

LEMMA 2.3.1: Let M be any block in the non-decreasing estimate of the regression function. Then for all k such that $x_k \in M$ we have

$$\sum_{x_i \in M} \left[y(x_i) - \bar{y}^+(x_i) \right] f_k(x_i) w(x_i) \leq 0$$

Proof: Suppose that

$$\sum_{x_i \in M} \left[y(x_i) - \bar{y}^+(x_i) \right] f_k(x_i) w(x_i) > 0$$

Then

$$\sum_{x_i \in M} \left[y(x_i) - \bar{y}^+(x_i) + \varepsilon\, f_k(x_i) \right]^2 w(x_i) =$$

$$\sum_{x_i \in M} \left[\left[y(x_i) - \bar{y}^+(x_i) \right]^2 w(x_i) + \varepsilon^2 f_k(x_i)^2 w(x_i) - 2\varepsilon \left[y(x_i) - \bar{y}^+(x_i) \right] f_k(x_i) w(x_i) \right]$$

For ε small enough and positive we can always get

$$2\varepsilon \sum_{x_i \in M} \left[y(x_i) - \bar{y}^+(x_i) \right] f_k(x_i) w(x_i) > \varepsilon^2 \sum_{x_i \in M} \left[f_k(x_i) \right]^2 w(x_i)$$

Choose an ε satisfying this and also being smaller than the increase in level mean from M to the neighbouring block to the right. This means that we can find another isotonic regression function that gives a smaller sum of squared deviation than $\bar{y}^+(x_i)$. But by definition $\overline{y}^{+}(x_i)$ is the isotonic regression function that gives the smallest sum of squared deviations. This means that $\sum_{x_i \varepsilon M}\left[y(x_i)-\overline{y}^{+}(x_i)\right] f_k(x_i)\,w(x_i)$ can not be positive.

Q.E.D.

The lemma implies that the linear regression function estimate with weights $w(x_i)$ for the x-observations in M always has a negative slope.

Simulations were made using 100 replications to examine the new variance estimation method for isotonic regression functions under the assumption that Y_i are identically normally distributed with $\mu(x)= -x^2$ and the common variance $\sigma^2=1$. In the simulations we used N fixed equidistant x-values. The number of x-observations varied from 50 to 500 within the x-interval (-5,5). The result of the simulations is shown in table 2.1.

TABLE 2.1

Number of x-observations, N, $-5 \le x \le 5$	$\hat{\sigma}^2_{21}$	$\hat{\sigma}^2_{22}$	$\hat{\sigma}^2_g = \sqrt{\hat{\sigma}^2_{21} \cdot \hat{\sigma}^2_{22}}$
50	0.66	1.97	1.14
75	0.69	1.60	1.05
100	0.75	1.44	1.04
200	0.82	1.24	1.01
300	0.86	1.18	1.01
400	0.89	1.15	1.01
500	0.91	1.13	1.01

Table 2.1. Random numbers were generated so that Y ~ N($-x^2$;1) for equidistant x-observations in the x-interval $-5 \le x \le 5$. The number of x-observations varied from 50 to 500. The variances were estimated according to the proposed estimation method.

We see from this that the geometric average $\hat{\sigma}_g^2$ of the two variance estimates $\hat{\sigma}_{21}^2$ and $\hat{\sigma}_{22}^2$ is a possible modification which is rather good but overestimate $Var(Y_i)$ a little. However when the numbers of x-observations increase $\hat{\sigma}_g^2$ decreases towards $\sigma^2=1$. Even for moderate values of N, $\hat{\sigma}_g^2$ is very close to σ^2. This study indicates that the estimation method is rather good. It also seems that N-p is a rather good approximation for the degrees of freedom.

2.4 AN EXAMPLE

Let y_i be the average corn yield and x_i be the amount of nitrogen fertilizer. In the examined x-interval we can estimate a non-decreasing regression function to the data. Since we only want to use the monotonicity restriction isotonic regression will be used. The data together with the estimates are

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130
y	23	49	42	45	65	66	59	73	82	85	81	79	82	80
$\bar{y}^+$	23	45.33	45.33	45.33	63.33	63.33	63.33	73.00	81.50	81.50	81.50	81.50	81.50	81.50

x	140	150	160	170	180
y	88	96	97	95	94
$\bar{y}^+$	88.00	95.50	95.50	95.50	95.50

The degrees of freedom is estimated by 19-7=12 where 19 is the number of observations and 7 is the number of plateaus.

Using these data we obtain $\hat{\sigma}_{21}^2$ = 6.653, which underestimate the variance. To be able to determine $\hat{\sigma}_{22}^2$ we first calculate the seven straight lines, one for each plateau:

No	the equation of the straight line	the estimated y-values
1	$\hat{y}_i = 0.67 + 22.33\, x_i$	$\hat{y}_1 = 23.000$
2	$\hat{y}_i = 30.804 + 6.72167\, x_i$	$\hat{y}_2 = 37.526$, $\hat{y}_3 = 44.2473$, $\hat{y}_4 = 50.969$
3	$\hat{y}_i = 52.024 + 4.612\, x_i$	$\hat{y}_5 = 56.636$, $\hat{y}_6 = 61.248$, $\hat{y}_7 = 65.86$
4	$\hat{y}_i = 63.6225 + 9.085\, x_i$	$\hat{y}_8 = 72.7075$
5	$\hat{y}_i = 76.625 + 1.25\, x_i$	$\hat{y}_9 = 77.875$, $\hat{y}_{10} = 79.125$, $\hat{y}_{11} = 80.375$, $\hat{y}_{12} = 81.625$, $\hat{y}_{13} = 82.875$, $\hat{y}_{14} = 84.125$
6	$\hat{y}_i = 81.25 + 7.00\, x_i$	$\hat{y}_{15} = 88.250$
7	$\hat{y}_i = 90.8125 + 1.875\, x_i$	$\hat{y}_{16} = 92.688$, $\hat{y}_{17} = 94.563$, $\hat{y}_{18} = 96.438$, $\hat{y}_{19} = 98.313$

From these y-estimates we obtain $\hat{\sigma}_{22}{}^2 = 16.332$. This implies that $\hat{\sigma}_g = 3.229$.

3 ESTIMATION OF VAR(Y) FOR CONCAVE/CONVEX REGRESSION FUNCTIONS

Suppose that we have the same estimation situation as in section 2.2, i.e. we have only got one y-observation for each x_i. If $\mu(x_i)$ is a concave (or convex) function and Y_j j=1,...,N, are observations with $Y_i \sim N(\mu(x_i); \sigma^2(x_i))$ where the variances $\sigma^2(x_i)$ are assumed to be the same for all x_i, then the variance might be estimated by

$$\hat{\sigma}_1^2 = \frac{1}{N}\sum_{i=1}^{N}\left[y(x_i) - \hat{\mu}(x_i)\right]^2 w_i$$

When we use concave regression to estimate $\mu(x)$, $(\hat{\sigma}_1)^2$ will underestimate the variance σ^2. Therefore another estimation method has been developed for this situation.

Suppose that Y_i, i=1,...,N, have the same variance σ^2. We start by looking at three observations, since this is the smallest possible number of observations to form a concave (or convex) function.

Suppose that we have three observations, μ_1, μ_2, μ_3, forming a concave regression function. Let the corresponding observations on the least squares estimate of the regression line for these three observations be denoted by μ_1^*, μ_2^* and μ_3^*. Thus $\mu_1^* \geq \mu_1$, $\mu_2^* \leq \mu_2$ and $\mu_3^* \geq \mu_3$. Denote the distance between x_1 and x_2 by Δ_{12} and the distance between x_2 and x_3 by Δ_{23}. Then the corresponding distance between μ_1^* and μ_2^* will be $d_1=\hat{\beta}\Delta_{12}$ and between μ_2^* and μ_3^* $d_2=\hat{\beta}\Delta_{23}$ where $\hat{\beta}$ is the slope of the regression line through μ_1^*, μ_2^* and μ_3^*.

<u>Case 1</u>: If we have the special case that the concave regression function itself is a straight line we get equality for μ_i^* and μ_i, i=1,2,3, i.e. $\mu_1^* = \mu_1$, $\mu_2^* = \mu_2$ and $\mu_3^* = \mu_3$. If the x-observations are equidistant, i.e. if $\mu_2^* = 0.5\ (\mu_1^* + \mu_3^*)$, then we can use the corresponding real observations Y_1, Y_2, Y_3 to calculate

$$s^2 = \frac{2}{3}\left[Y_2 - 0.5(Y_1 + Y_3)\right]$$

But usually this is not the case. Therefore we will use a more general formula.

The relation between μ_1^*, μ_2^* and μ_3^* can be written as $\mu_2^* = (d_1+d_2)^{-1}(d_2\mu_1^* + d_1\mu_3^*)$. Then we can do the corresponding calculation

$$s^2 = \left((d_1+d_2)^2 + d_1^2 + d_2^2\right)^{-1}(d_1+d_2)^2 * [Y_2 - (d_1+d_2)^{-1}(d_2Y_1 + d_1Y_3)]^2 \ldots\ldots\ldots\ldots\ldots\ldots (3.1)$$

Denote $U = Y_2 - (d_1+d_2)^{-1}(d_2Y_1 + d_1Y_3)$. Since $E[U] = E[Y_2 - (d_1+d_2)^{-1}(d_2Y_1 + d_1Y_3)] = \mu_2 - (d_1+d_2)^{-1}(d_2Y_1 + d_1Y_3) = 0$ thus

$$E(U^2) = (d_1+d_2)^{-2}\left((d_1+d_2)^2 + d_1^2 + d_2^2\right)\sigma^2.$$

From this calculation we obtain that $E(s^2) = \sigma^2$ and therefore (3.1) will serve as the estimated variance for these three observations. But usually a regression function consists of more than three observations. Then we can use

$$s^2 = (N-2)^{-1}\sum_{i=2}^{N-1}\frac{(d_{i-1}+d_i)^2}{(d_{i-1}+d_i)^2+d_{i-1}^2+d_i^2}\left[Y_i-\frac{1}{d_{i-1}+d_i}(d_iY_{i-1}+d_{i-1}Y_{i+1})\right]^2$$

as an estimate of the variance, σ^2, of the observations.

<u>Case 2</u>: Suppose that the concave regression function consists of three observations, μ_1, μ_2 and μ_3. Also suppose that we don't have the special case where the concave regression function itself is a straight line. Then we obtain the inequalities $\mu_1^* > \mu_1$, $\mu_2^* < \mu_2$ and $\mu_3^* > \mu_3$. These order conditions give us

$$E[Y_2-(d_1+d_2)^{-1}(d_2Y_1+d_1Y_3)] = \mu_2 - (d_1+d_2)^{-1}(d_2\mu_1+d_1\mu_3) > 0.$$

Thus

$$V = s/\sigma \sim N\left(\sqrt{\frac{(d_1+d_2)^2}{(d_1+d_2)^2+d_1^2+d_2^2}}\cdot\lambda\frac{1}{\sigma};1\right) \quad \text{where } \lambda = \mu_2 - (d_1+d_2)^{-1}(d_2\mu_1 + d_1\mu_3) \text{ and}$$

$$E(V^2) = Var(V) + [E(V)]^2 = 1 + \left((d_1+d_2)^2+d_1^2+d_2^2\right)^{-1}(d_1+d_2)^2(\lambda/\sigma)^2 > 1.$$

From this inequality follows that

$$E(s^2) = E(\sigma^2V^2) = \sigma^2 + \left((d_1+d_2)^2+d_1{}^2+d_2{}^2\right)^{-1}(d_1+d_2)^2\lambda^2 > \sigma^2.$$

Thus s^2 will overestimate the variance σ^2. If the regression function contains more that three observations the proposed estimate

$$s^2 = (N-2)^{-1}\sum_{i=2}^{N-1}\frac{(d_{i-1}+d_i)^2}{(d_{i-1}+d_i)^2+d_{i-1}^2+d_i^2}\left[Y_i-\frac{1}{d_{i-1}+d_i}(d_iY_{i-1}+d_{i-1}Y_{i+1})\right]^2$$

can be used as a rough estimate of the variance σ^2. To get a better estimate we can use

$$\sigma^2 = E(s^2) - \left((d_1+d_2)^2 + d_1^2 + d_2^2\right)^{-1}(d_1+d_2)^2\ \lambda^2$$

where λ is the average curvature of the regression function. This can be estimated according to the following:

Suppose we have least squares estimates of straight lines, $f(x_i) = \mu_i + \beta_i x_i$, i=1,...,N-1, through μ_i and μ_{i+1}. Thus the estimates consist of N-1 regression lines where the slopes $\beta_1 > \beta_2 > \ldots > \beta_{N-1}$ form a concave function. We can give an approximate model for the slope β_k for $\beta_1 > \beta_k > \beta_{N-1}$;

$$\beta_k = \beta_1 + (k-1)(N-1)^{-1}(\beta_{N-1}-\beta_1)$$

From this model we obtain

$$\beta_k - \beta_{k-1} = (N-1)^{-1}(\beta_{N-1}-\beta_1)$$

For the first two x-intervals we get

$$\lambda = \mu_2 - (d_1+d_2)^{-1}(d_2\mu_1 + d_1\mu_3) = (d_1+d_2)^{-1}(d_2\beta_1 - d_1\beta_2) =$$

$$(d_1+d_2)^{-1}\left[d_1(N-1)^{-1}(\beta_1-\beta_{N-1}) + (d_2-d_1)\beta_1\right]$$

This gives us the desired relation for the two first x-intervals

$$\frac{(d_1+d_2)^2}{(d_1+d_2)^2 + d_1^2 + d_2^2}\ \lambda^2 = \left((d_1+d_2)^2 + d_1^2 + d_2^2\right)^{-1}\left(\left[d_1(N-1)^{-1}(\beta_1-\beta_{N-1}) + (d_2-d_1)\beta_1\right]\right)^2$$

For any two x-interval we get the corresponding expression

$$\left((d_{i-1}+d_i)^2 + d_{i-1}^2 + d_i^2\right)^{-1}(d_{i-1}+d_i)^2\ \lambda^2 =$$

$$\left((d_{i-1}+d_i)^2 + d_{i-1}^2 + d_i^2\right)^{-1}\left(\left[d_{i-1}(N-1)^{-1}(\beta_1-\beta_{N-1}) + (d_i-d_{i-1})\beta_1\right]\right)^2$$

Since in this expression we have β_1 and β_{N-1} we must estimate them. This can be done according to the following:

Suppose we have a least squares estimate of a concave regression function with M bending points $x_0^{(1)} < x_0^{(2)} < \ldots < x_0^{(M)}$. Thus the estimate consists of M regression lines where the slopes $\hat{\beta}_1 > \hat{\beta}_2 \; ..... > \hat{\beta}_M$ form a concave function. The estimates of β_1 and β_{N-1} are $\hat{\beta}_1$ and $\hat{\beta}_M$, where $\hat{\beta}_1$ and $\hat{\beta}_M$ are the end slopes in the concave estimate of the whole regression function $\mu(x_i)$, i=1,...,N. The proposed variance estimate will be

$$s^2=(N-2)^{-1}\sum_{i=2}^{N-1}\Big\{ \left(\left(d_{i-1}+d_i\right)^2 d_{i-1}{}^2+d_i^2\right)^{-1}\left(d_{i-1}+d_i\right)^2\left(Y_i-\left(d_{i-1}+d_i\right)\left(d_iY_{i-1}+d_{i-1}Y_{i+1}\right)\right)^2 -$$
$$\left(\left(d_{i-1}+d_i\right)^{-1}\left[d_{i-1}(N-1)^{-1}\left(\hat{\beta}_1-\hat{\beta}_M\right)+\left(d_i-d_{i-1}\right)\hat{\beta}_1\right]\right)^2 \Big\}$$

If the x-observations are equidistant this formula will be simplified to

$$\hat{\sigma}^2 = \tfrac{2}{3}(N-2)^{-1}\sum_{i=2}^{N-1}\left(Y_i-\tfrac{1}{2}\left(Y_{i-1}+Y_{i+1}\right)\right)^2-\tfrac{1}{6}\left((N-1)^{-1}\left(\hat{\beta}_1-\hat{\beta}_M\right)\right)^2$$

From the proposed formula we can also see that if we have the described situation in case 1, that the regression function already is a straight line, then we obtain $\beta_1 = \beta_{N-1}$ and the correction term vanishes. This gives us the variance estimate that was proposed in case 1.

3.1 IMPROVEMENT OF THE PROPOSED VARIANCE ESTIMATION METHOD FOR REGRESSION FUNCTIONS WITH BIG CURVATURE

The estimation method proposed in the preceding section is based on short linear pieces, which are weighted together. If the regression function has a big curvature we can expect that the variance will be overestimated rather much. In this situation it might be appropriate to consider second-degree functions instead of linear pieces. If we use the assumption that all Y_i have the same variance then the corresponding variance estimation method can be evaluated according to the following discussion:

Since we will use second-degree functions three parameters are to be estimated. Therefore we start by looking at five observations forming a concave (or convex) function. This gives us two degrees of freedom for the variance estimate.

Case 1: Suppose we have equidistant x-observations with the distance Δ between two successive observations. Let the points be centered at x=0.0. Thus the five observations will be Y_{-2}, Y_{-1}, Y_0, Y_1 and Y_2 and the corresponding x-observations -2Δ, $-\Delta$, 0.0, Δ and 2Δ. The parameters β_i, i=0,1,2, to be estimated are contained in the function $y= \beta_0 + \beta_1 x +\beta_2(x_2- \gamma)$. We can simplify the evaluation if we calculate an appropriate value of γ to obtain orthogonality:

$$((-2\Delta)^2-\gamma)1+((-\Delta)^2-\gamma)1 -\gamma 1+(\Delta^2-\gamma)1+((2\Delta)^2-\gamma)1 = 0 \quad \Leftrightarrow \quad \gamma = 2\Delta^2.$$

Then we obtain the estimates

$$\hat{\beta}_0 = \frac{1}{5}\sum_{i=-2}^{2} Y_i \ , \quad \hat{\beta}_1 = \frac{\sum_{i=-2}^{2} i \cdot Y_i}{10\Delta} \text{ and } \hat{\beta}_2 = \frac{2Y_{-2} - Y_{-1} - 2Y_0 - Y_1 + 2Y_2}{14\Delta^2} .$$

From the five points we obtain the following contribution to the variance estimate

$$\frac{1}{2}\sum_{i=-2}^{2}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2(x_i^2 - 2\Delta^2)\right)\right]^2$$

This expression can be evaluated to hold for several second-degree pieces. Suppose that we obtain the sets of five observations successively. Denote the estimated parameters from the k:th set by $\hat{\beta}_i^{(k)}$, i=0,1,2. Thus we can use

$$s^2 = (N-4)^{-1}\sum_{k=3}^{N-2}\frac{1}{2}\sum_{i=k-2}^{k+2}\left[(Y_i - Y_k) - \left(\hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)}(x_i - x_k) + \hat{\beta}_2^{(k)}((x_i - x_k)^2 - 2\Delta^2)\right)\right]^2$$

as an estimate of the variance, σ^2, of the observations on the whole curve.

Case 2: If the x-observations are not equidistant then it is no simplification to central the x-observations or determine a value of γ to obtain orthogonality. Instead we will use the ordinary normal equations to estimate β_i, i=0,1,2. Thus the corresponding estimate of the variance, σ^2, for the observations on the whole curve is reduced to

$$s^2 = (N-4)^{-1}\sum_{k=3}^{N-2}\frac{1}{2}\sum_{i=k-2}^{k+2}\left[Y_i - (\hat{\beta}_0^{(k)} + \hat{\beta}_1^{(k)} x_i + \hat{\beta}_2^{(k)} x_i^2)\right]^2$$

If the curvature of the regression function is big this alternative will not overestimate the variance as much as the one proposed for linear pieces. However, for functions with small curvature we will obtain almost the same size of the estimate of the variance. In practice this is probably a minor problem. Therefore the simulation studies in the following section are made for the estimation method proposed for linear pieces.

3.2 SOME PROPERTIES OF THE VARIANCE ESTIMATION METHOD FOR CONCAVE REGRESSION OBTAINED BY SIMULATIONS

A simulation study was performed to examine the properties of the proposed variance estimation method for a regression function with constant normalized curvature (NC). We used 25 000 replications for the regression function $Y_i = \mu(x_i) + \varepsilon_i$, where $\mu(x) = -cx^2$, $\varepsilon_i \sim N(0;1)$ for $-10 \le x \le 10$ and $NC = \frac{c\Delta^2}{\sigma/\sqrt{n}}$.

TABLE 3.1.

NC	$\hat{E}(s^2)$	correction term	$\hat{\sigma}^2$	$\hat{\sigma}$
0.0125	1.0010	0.0009	1.0001	1.000
0.05	1.0068	0.0036	1.0032	1.000
0.20	1.0258	0.0184	1.0074	1.004
0.50	1.1657	0.1024	1.0633	1.031
0.75	1.3740	0.2356	1.1377	1.067
1.00	1.6655	0.4421	1.2222	1.106
1.50	2.4987	1.0994	1.3977	1.182
2.00	3.6568	2.1300	1.5335	1.238

Table 3.1. Random numbers were generated so that $Y \sim N(-cx^2;1)$ for equidistant x-observations in the x-interval $-10 \le x \le 10$. The variances were estimated according to the proposed estimation method for $0.0125 \le NC \le 2.00$.

The x-observations were chosen equidistantly. The position of the nearest x-value compared to x_{max} was uniformly distributed in the interval $(-0.5\Delta, 0.5\Delta)$. In the simulations both the estimated expected value of s^2, the estimated correction term, the estimated σ^2 and the corresponding standard deviations of the estimates were determined. The results were rescaled and shown in table 3.1.

From this table we can see that both the estimated expected value of s^2, the estimated correction term and the estimated σ^2 increase when NC increases. But still when NC=2.0, the estimated σ^2 will be rather moderate. Bigger values of NC are useless to examine since then the observations almost themselves will form a concave regression function. From the table we can see that the estimated σ^2 seems to converge to σ^2 when the normalized curvature decreases to zero.

The simulation results also indicate that the variance of the estimated σ^2 increases when NC increases.

TABLE 3.2.

NC	μ_3 /(stand dev(σ^3))	μ_4 /(stand dev(σ^4))
0.05	1.008	4.495
0.20	1.008	4.498
0.50	0.981	4.424
0.75	0.900	4.245
1.00	0.801	3.982
1.50	0.707	3.217
2.00	1.197	3.973

Table 3.2. Random numbers were generated so that Y ~ N($-cx^2$;1) for equidistant x-observations in the x-interval $-10 \leq x \leq 10$. The standardized skewness and the standardized kurtosis were obtained for the variance estimates for $0.05 \leq NC \leq 2.00$.

To examine the variation in the results the 3:rd, μ_3, and the 4:th, μ_4, moments were determined and the estimates of μ_3/(stand dev(σ)^3) and μ_4/(stand dev(σ)^4) were calculated for different values of the normalized curvature, NC, together with the estimated distributions. The results were rescaled and shown in table 3.2.

From this table we can see that the standardized skewness was around 1.0 for all values of NC and the standardized kurtosis near 4.0. To see if the estimate could be approximated by a χ^2-distribution, we also more carefully examined the estimated

distribution of the estimated σ^2. This knowledge can be very useful if we try to determine approximately 95% limits of the estimate. The estimated cumulative distribution for NC=0.5 obtained from the simulations is shown in the appendix.
From the same simulation study we also obtained the 0.025-limits and the 0.975-limits of the estimated distributions for 0.20≤ NC ≤2.00. The results are shown in table 3.3. This table shows that both the value of the lower and the upper limit increase when NC increases.

TABLE 3.3.

NC	**0.20**	**0.50**	**0.75**	**1.00**	**1.50**	**2.00**
lower limit	0.350	0.400	0.425	0.450	0.505	0.600
upper limit	2.075	2.150	2.225	2.350	2.765	3.450

Table 3.3. Random numbers were generated so that Y ~ N($-cx^2$;1) for equidistant x-observations in the x-interval $-10 \le x \le 10$. The 0.025 and 0.975 limits of the estimated cumulative distribution of the estimated σ^2, using the proposed estimation method, were obtained for 0.20 ≤ NC ≤ 2.00.

It is also obvious that the 95% confidence interval becomes wider when NC increases. Since the χ^2-distribution in the corresponding situation has 20 degrees of freedom we can compare the obtained results, rescaled as $(N-1)\hat{\sigma}^2 / \sigma^2$, with the 95% confidence limits of the χ^2-distribution, which are 9.59 and 34.17.

TABLE 3.4.

NC	**0.20**	**0.50**	**0.75**	**1.00**	**1.50**	**2.00**	**χ^2(20 df)**
lower limit	7.00	8.00	8.50	9.00	10.1	12.0	9.59
upper limit	41.5	43.0	44.5	47.0	55.3	69.0	34.17

Table 3.4. Random numbers were generated so that Y ~ N($-cx^2$;1) for equidistant x-observations in the x-interval $-10 \le x \le 10$. The 0.025 and 0.975 limits of the estimated distribution of the estimated σ^2, using the proposed estimation method, were rescaled for 0.20≤N≤2.00 and compared to the χ^2-distribution with 20 df.

The result of this comparison is shown in table 3.4. The result in this table shows that the lower level corresponds rather well to the χ^2-distribution but the upper limit is much higher in the estimated distributions. It is also obvious that the obtained limits are some functions of NC.

To get an expression for the connection between NC and the estimated variance we use some results of Mammen (1988). He gives an expression, containing $\mu(x)$ and $\hat{\mu}(x)$, that converges to a limit distribution for concave regression. Let

$\mu''(x_0)$ = the curvature in x_0,
$\sigma(x_0)^2 = Var(Y|X=x_0)$,
$N\,F'(x_0)$ = the density of observations in x_0,
G = the universal limit distribution.

Using these notations we have according to Mammen

$$N^{2/5} * \frac{F'(x_0)^{2/5}}{\sigma(x_0)^{4/5}\mu''(x_0)^{1/5}}\left(\hat{\mu}(x_0)-\mu(x_0)\right) \quad \rightarrow \quad G$$

From this expression we see that the distribution depends on the curvature of $\mu(x)$ and the variance of the observations. The product $N^{2/5}\,F'(x_0)^{2/5}$ is used as a scaling factor. The limit distribution G has a very complicated form and Mammen gives no theoretical expression for it. Probably the mathematical expression for how the curvature affects the estimation of the variance is not easy to find.

3.3 AN EXAMPLE

In this situation we use the same example as for the estimation method using isotonic regression, i.e. let y_i be the average corn yield and x_i be the amount of nitrogen fertilizer. Since we use concavity restriction concave regression will be used. The data are

x	0	10	20	30	40	50	60	70	80	90	100	110	120	130
y	23	49	42	45	65	66	59	73	82	85	81	79	82	80

x	140	150	160	170	180
y	88	96	97	95	94

When we use concave regression we obtain the following estimated regression function together with the corresponding bending points:

$\hat{\beta}_i$	1.9560	0.4953	0.2028	0.1546	-0.1000
bending points, $x_{(i)}$	1	2	5	9	17

Since the x-observations are equidistant we can use the simplified formula

$$\hat{\sigma}^2 = \frac{2}{3}(N-2)^{-1}\sum_{i=2}^{N-1}\left(Y_i - \frac{1}{2}(Y_{i-1}+Y_{i+1})\right)^2 - \frac{1}{6}\left((N-1)^{-1}\left(\hat{\beta}_1 - \hat{\beta}_M\right)\right)^2.$$

Then we obtain the following results:

$$\sum_{i=2}^{N-1}\left(Y_i - \frac{1}{2}(Y_{i-1}+Y_{i+1})\right)^2 = (49\text{-}0.5(23+42))^2 + \ldots + (95\text{-}0.5(97+94))^2 = 666.75$$

$$\left((N-1)^{-1}\left(\hat{\beta}_1 - \hat{\beta}_M\right)\right)^2 = 18^{-1}\,(1.9560+0.1000)^2 = 0.013047$$

This gives us the variance estimate $\hat{\sigma}^2 = 26.14488$ and the corresponding estimated standard deviation $\hat{\sigma} = 5.113$. If we use the proposed method that improves the estimated for regression functions with big curvature we obtain the estimated standard deviation $\hat{\sigma} = 5.334$, which indicates that we do not have a big curvature in this problem.

ACKNOWLEDGMENTS

I would like to thank professor Kasra Afsarinejad for interesting discussions and good advice. I am also very grateful to professor Marianne Frisén for her good advice.

APPENDIX

TABLE

$\hat{\sigma}^2$	$F(\hat{\sigma}^2)$	$\hat{\sigma}^2$	$F(\hat{\sigma}^2)$	$\hat{\sigma}^2$	$F(\hat{\sigma}^2)$	$\hat{\sigma}^2$	$F(\hat{\sigma}^2)$
0.250	0.0001	1.150	0.6299	2.050	0.9665	2.950	0.9980
0.275	0.0016	1.175	0.6501	2.075	0.9691	2.975	0.9982
0.300	0.0042	1.200	0.6696	2.100	0.9715	3.000	0.9984
0.325	0.0079	1.225	0.6874	2.125	0.9736	3.025	0.9986
0.350	0.0124	1.250	0.7037	2.150	0.9756	3.050	0.9988
0.375	0.0178	1.275	0.7218	2.175	0.9774	3.075	0.9989
0.400	0.0247	1.300	0.7365	2.200	0.9790	3.100	0.9990
0.425	0.0319	1.325	0.7516	2.225	0.9805	3.125	0.9991
0.450	0.0406	1.350	0.7665	2.250	0.9822	3.150	0.9992
0.475	0.0497	1.375	0.7813	2.275	0.9836	3.175	0.9992
0.500	0.0605	1.400	0.7946	2.300	0.9849	3.200	0.9993
0.525	0.0730	1.425	0.8076	2.325	0.9864	3.225	0.9993
0.550	0.0868	1.450	0.8189	2.350	0.9876	3.250	0.9994
0.575	0.1041	1.475	0.8302	2.375	0.9889	3.275	0.9994
0.600	0.1208	1.500	0.8396	2.400	0.9899	3.300	0.9995
0.625	0.1408	1.525	0.8486	2.425	0.9905	3.325	0.9996
0.650	0.1607	1.550	0.8580	2.450	0.9914	3.350	0.9996
0.675	0.1816	1.575	0.8681	2.475	0.9919	3.375	0.9997
0.700	0.2024	1.600	0.8768	2.500	0.9924	3.400	0.9997
0.725	0.2264	1.625	0.8840	2.525	0.9927	3.425	0.9998
0.750	0.2523	1.650	0.8684	2.550	0.9932	3.450	0.9998
0.775	0.2767	1.675	0.8985	2.575	0.9938	3.475	0.9998
0.800	0.3001	1.700	0.9051	2.600	0.9941	3.500	0.9999
0.825	0.3254	1.725	0.9112	2.625	0.9948	3.525	0.9999
0.850	0.3504	1.750	0.9170	2.650	0.9951	3.550	0.9999
0.875	0.3751	1.775	0.9221	2.675	0.9955	3.575	0.9999
0.900	0.3997	1.800	0.9279	2.700	0.9960	3.600	0.9999
0.925	0.4248	1.825	0.9332	2.725	0.9963	3.625	0.9999
0.950	0.4485	1.850	0.9386	2.750	0.9968	3.650	0.9999
0.975	0.4727	1.875	0.9434	2.775	0.9970	3.675	0.9999
1.000	0.4956	1.900	0.9473	2.800	0.9972	3.700	0.9999
1.025	0.5174	1.925	0.9515	2.825	0.9974	3.725	0.9999
1.050	0.5410	1.950	0.9544	2.850	0.9976	3.750	0.9999
1.075	0.5631	1.975	0.9574	2.875	0.9978	3.775	0.9999
1.100	0.5866	2.000	0.9604	2.900	0.9978	3.800	0.9999
1.125	0.6086	2.025	0.9628	2.925	0.9980	3.825	1.0000

Table. Random numbers were generated so that Y ~ N($-cx^2$;1) for equidistant x-observations in the x-interval $-10 \leq x \leq 10$. The cumulative distribution of the estimated σ^2, using the proposed estimation method, was estimated for NC = 0.50.

BIBLIOGRAPHY

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D.; (1972): Statistical Inference Under Order Restrictions, Wiley, New York.

Dahlbom, U.; (1986): Some Properties of Estimates of Unimodal Regression, Research Report **1986:6**, University of Göteborg, Department of Statistics.

Dahlbom, U.; (1994): Estimation of Regression Functions with certain Monotonicity and Concavity/Convexity Restrictions, Almqvist & Wiksell International, Stockholm, ISBN-91-22-01641-4.

Dent, W.; (1973): A Note on Least Squares Fitting of Functions Constrained to be Either Nonnegative, Nondecreasing or Convex, Management Science **20**, pg 130-132.

Fraser, D. A. S. and Massam, H.; (1989): A Mixed Primal-Dual Bases Algorithm for Regression Under Inequality Constraints. Application to Concave Regression, Scandinavian Journal of Statisticts: pg 65-74.

Frisén, M.; (1980): U-Shaped Regression, COMPSTAT Proc Comput Statist **4**: pg 304-307.

Hanson, D.L. and Pledger, G.; (1976): Consistency in Concave Regression, Annals of Statistics **4**, pg 1038-1050.

Hildreth, C.; (1954): Point Estimates of Ordinates of Concave Functions, Journal of American Statistical Association **49**, pg 598-619.

Holloway, C.A.; (1979): On the Estimation of Convex Functions, Operations Research **27**, pg 401-407.

Holm, S. and Frisén, M.; (1985): Nonparametric Regression with simple curve characteristics, Research report **1985:4**, University of Göteborg, Dept. of Statistics.

Mammen, E.; (1988): Nonparametric Regression under Qualitative Smoothness Assumptions, Unpublished Manuscript, Universität Heidelberg, Institut für Angewandte Mathematik.

Robertson, T., Wright, F.T. and Dykstra, R.L.; (1989): Order Restricted Statistical Inference, Wiley, New York.

Wu, C-F.; (1982): Some Algorithms for Concave and Isotonic Regression, Studies in the Management Sciences **19**: pg 105-116.

Research Report

1997:1	Ekman, A.	Sequential probability ratio tests when using randomized play-the-winner allocation.
1997:2	Pettersson, M.	Monitoring a Freshwater Fishpopulation: Statistical Surveillance of Biodiversity.
1997:3	Jonsson, R.	Screening-related prevalence and incidence for non-recurrent diseases.
1997:4	Johnsson, T.	A stepwise regression procedure applied to an international production study - a multiple inference solution.
1998:1	Särkkä, A. & Högmander, H.:	Multiple spatial point patterns with hierarchical interactions.
1998:2	Frisén, M. & Wessman, P.:	Evaluations of likelihood ratio methods for surveillance. Differences and robustness.
1998:3	Järpe, E.:	Surveillance of spatial patterns. Change of the interaction in the Ising model.
1998:4	Pettersson, M.:	Evaluation of some methods for statistical surveillance of an autoregressive process.
1998:5	Pettersson, M.:	On monitoring of environmental and other autoregressive processes.
1998:6	Dahlbom, U.:	Least squares estimates of regression functions with certain monotonicity and concavity/convexity restrictions.