ISSN 0974-3235

ProbStat Forum, Volume 01, July 2008, Pages 15-20

Variance of estimates of variance of deviations from regression function

Tsitsiashvili G. Sh.
690041, Vladivostok, Radio str. 7
Institute for Applied Mathematics
FAR Eastern Branch of RAS
e-mail: *guram@iam.dvo.ru*

Abstract. An algorithm is given to calculate the variance of an estimate for the variance of deviations between observations at integer-valued points and a polynomial regression function. This is realized without estimating the regression coefficients.

Mathematics Subject Classification. 60K25, 60K30

Keywords. polynomial regression, algorithm, time series, variance estimation.

1 Introduction

In classical statistics the problem of variance estimation is solved by considering emperical variance. But even for this widely used statistics it is complicated to calculate its own variance. Now to calculate the variance of an estimate for the variance of deviations from a polynomial regression function is much more difficult. Nevertheless it is possible to solve this problem if observations are made at integer-valued points. The problem of estimation of deviation between observations and a regression function has its origin from an analysis of time series of overground air dynamics in connection with the global climate warming phenomenon.

Here this problem is discussed when the regression function is a polynomial of integer-valued argument and is solved by a special algorithm which is realized without estimating the regression coefficients. In the first step an analogue of emperical variance is considered so that it is possible to calculate its own variance. In the second step, for the regression function, represented by a polynomial of integer-valued argument, a special recurrence procedure is constructed which transforms random observations on the polynomial into a sequence of i.i.d r.vs. Then it is possible to use the results in the first step. The algorithm can be generalized into a multidimensional setup easily in which case estimation of variance is replaced by estimation of covariance matrix. This generalization has applications in mathematical geodesy.

In the next two sections we consider a modified estimate of variance and apply it to the context of deviations from a polynomial regression function. In section 4 the multidimensional case is discussed. In all considered cases accuracy formulas for variances of suggested estimations are obtained.

2 Modified estimation of variance

Suppose that $x_1, x_2, \dots$ is the sequence of independent and identically distributed (i.i.d) random variables (r.v) with the common distribution function $F(t)$. Denote

$$Ex_1 = \int_{-\infty}^{\infty} tdF(t) = a,$$

$$Var\ x_1 = \int_{-\infty}^{\infty} t^2 dF(t) - a^2 = E(x_1 - Ex_1)^2 = b,$$

and assume an absolute convergence of integrals in the mathematical expectation a and the variance b definitions. Usual estimates of a, b are the empirical expectation and the empirical variance (see Rozanov, 1971, p.318):

$$\hat{a}_n = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \hat{b}_n = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{a}_n)^2,$$

which are unbiased. The empirical expectation $\hat{a}_n$ has the variance $Var\ \hat{a}_n = b/n$. But a calculation of the variance of $\hat{b}_n$ is sufficiently complicated procedure. So in this section we consider the following unbiased estimate of b for which a calculation of its variance is apparent.

Introduce i.i.d r.v's $z_1 = x_2 - x_1,\ z_2 = x_4 - x_3, \dots$, satisfying the equalities

$$Ez_1 = 0,\ Var\ z_1 = 2Var\ x_1 = 2b. \tag{2.1}$$

Using r.v's $x_1, \dots, x_{2n}$ define the estimate b'_n of b by the formulas:

$$b'_n = \frac{1}{2n}\sum_{i=1}^{n} z_i^2.$$

For an arbitrary r.v u define $\overline{u} = u - Eu$, then $b = E\overline{x}_1^2$, denote $d = E\overline{\overline{x}_1^2}^2$. From the equalities (2.1) obtain:

$$Eb'_n = b,\ N = 2n,\ Var\ b'_n = \frac{d + 2b^2}{N}. \tag{2.2}$$

Here N is the total volume of a sample from the random sequence $x_1, x_2, \dots$, which is necessary to obtain the estimate b'_n.

Assume now that i.i.d r.v's $x_1, x_2, \ldots$ have normal distribution then b'_n is the likelihood estimate for the sequence $z_1, \ldots, z_n$. In an accordance with well known formulas (see Orlov, 2004, p.110) for moments of normal distribution and from (2.2) we obtain

$$Var\ b'_n = \frac{4b^2}{N}. \tag{2.3}$$

3 Variance of deviations from polynomial regression function with integer-valued argument

Suppose that the random sequence $\{y_{1,0}, y_{2,0}, \ldots\}$ satisfies the equalities

$$y_{i,0} = P_m(i) + \varepsilon_{i,0},\ i = 1, \ldots,$$

where

$$P_m(i) = \sum_{j=0}^{m} i^j p_{m,j}$$

is the polynomial of the degree m, $\varepsilon_{1,0}, \varepsilon_{2,0}, \ldots$ are i.i.d.r.v's satisfying the equalities

$$E\varepsilon_{1,0} = 0,\ Var\ \varepsilon_{1,0} = E\varepsilon_{1,0}^2 = \beta_0,\ \overline{E\varepsilon_{1,0}^2}^2 = f_0.$$

Define recurrently the random sequences

$$\{y_{1,1}, y_{2,1}, \ldots\},\ \{y_{1,2}, y_{2,2}, \ldots\}, \ldots, \{y_{1,m}, y_{2,m}, \ldots\}$$

as follows

$$y_{i,k+1} = y_{2i,k} - y_{2i-1,k},\ 1 \le i,\ k = 0, \ldots, m-1.$$

By the definition

$$y_{i,k+1} = P_{m-k-1}(i) + \varepsilon_{i,k+1},\ \varepsilon_{i,k+1} = \varepsilon_{2i,k} - \varepsilon_{2i-1,k},$$

$$P_{m-k-1}(i) = P_{m-k}(2i) - P_{m-k}(2i-1),\ 1 \le i,\ k = 0, \ldots, m-1.$$

Using the induction by k prove that $P_{m-k-1}(i)$ is the polynomial of the degree $(m-k-1)$ of i. Indeed, using the binomial theorem and the equality $a^r - b^r = (a-b)(a^{r-1} + a^{r-2}b + \ldots + b^{r-1})$, which is true for arbitrary natural number r, obtain

$$P_{m-k-1}(i) = P_{m-k}(2i) - P_{m-k}(2i-1) = \sum_{j=0}^{m-k} (2i)^j p_{m-k,j} - \sum_{j=0}^{m-k} (2i-1)^j p_{m-k,j} =$$

$$= \sum_{j=1}^{m-k} p_{m-k,j}((2i)^j - (2i-1)^j) = \sum_{j=1}^{m-k} p_{m-k,j} \sum_{t=0}^{j-1} (2i)^t (2i-1)^{j-1-t}.$$

Then from the binomial theorem it is easy to see that the function $P_{m-k-1}(i)$ may be represented as the polynomial of the degree $(m-k-1)$ of integer argument i :

$$P_{m-k-1}(i) = \sum_{j=0}^{m-k-1} p_{m-k-1,j} i^j.$$

So $E\varepsilon_{i,k+1} = 0,\ Var\ \varepsilon_{1,k+1} = \beta_{k+1} = 2\beta_k,$

$$E\overline{\varepsilon^2_{1,k+1}}^2 = f_{k+1} = 2f_k + 4\beta_k^2 = 2^{k+1}f_0 + 4\sum_{i=0}^{k} 2^{k-i}\beta_i^2 = 2^{k+1}f_0 + 4\beta_0^2\sum_{i=0}^{k} 2^{k+i} =$$

$$= 2^{k+1}f_0 + 4\beta_0^2 2^k(2^{k+1}-1) = 2^{k+1}(f_0 + 2\beta_0^2(2^{k+1}-1))$$

and consequently

$$\beta_m = 2^m\beta_0,\ f_m = 2^m(f_0 + 2\beta_0^2(2^m-1)),\ m \geq 1.$$

Define now $x_i = y_{m,i},\ i \geq 1,\ b = \beta_m,\ d = f_m$ and construct by the sample $x_1, \ldots, x_{2n}$ the estimate $b'_{n,m} = b'_n/2^m$, of $\beta_0 : Eb'_{n,m} = \beta_0,$

$$Var\ b'_{n,m} = \frac{d+2b^2}{2^{2m+1}n} = \frac{f_m + 2\beta_m^2}{2^{2m+1}n} = \frac{2^m(f_0 + 2\beta_0^2(2^m-1)) + 2^{2m+1}\beta_0^2}{2^{2m+1}n}$$

$$= \frac{f_0 + \beta_0^2(2^{m+2}-2)}{2^{m+1}n}.$$

If i.i.d r.v's $\varepsilon_{i,0},\ i = 1, \ldots,$ have normal d.f. then $f_0 = 2\beta_0^2$ and so

$$Var\ b'_{n,m} = \frac{2\beta_0^2}{n}.$$

To construct the estimate $b'_{n,m}$ it is necessary to have the sample $y_{1,0}, \ldots, y_{N,0}$, consisting of $N = 2^{m+1}n$ members.

4 Modified empirical covariances

Consider i.i.d random vectors

$$Z = (z_1, \ldots, z_m),\ Z_1 = (z_{11}, \ldots, z_{1m}),\ Z_2 = (z_{21}, \ldots, z_{2m}), \ldots.$$

Without loss of generality suppose that $Ez_1 = \ldots = Ez_m = 0$. Random vectors $Z, Z_1, Z_2, \ldots$ have common multidimensional distribution. These vectors are defined by i.i.d random vectors $(e_1, \ldots, e_m),\ (e_{11}, \ldots, e_{1m}),\ (e_{21}, \ldots, e_{2m}), \ldots,$ as follows

$$z_j = \sum_{t=1}^{m} a_{jt}e_t,\ z_{ij} = \sum_{t=1}^{m} a_{jt}e_{it},\ i \geq 1,\ 1 \leq j \leq m.$$

The random vector $(e_1, \dots, e_m)$ consists of i.i.d random components with zero mathematical expectation, single variance and finite fourth moment μ, the matrix $||a_{j,t}||_{j,t=1}^m$ consists of real numbers. The covariance $cov(z_t, z_s) = c(t,s)$ satisfies the equality

$$c(t,s) = \sum_{k=1}^{m} a_{tk} a_{sk}.$$

An unbiased estimate of the covariance $c(t,s)$ is

$$\hat{c}_n(t,s) = \frac{1}{n} \sum_{i=1}^{n} z_{it} z_{is}.$$

The estimate $\hat{c}_n(t,s)$ has the variance

$$Var\ \hat{c}_n(t,s) = \frac{1}{n} D z_t z_s,$$

where

$$Var\ z_t z_s = Var \sum_{j,k=1}^{m} a_{tj} a_{sk} e_j e_k = b(t,s) - c^2(t,s),$$

$$b(t,s) = E \left[\sum_{j=1}^{m} \sum_{k=1}^{m} a_{tj} a_{sk} e_j e_k \right]^2 = \sum_{j=1}^{m} \sum_{k=1}^{m} \sum_{r=1}^{m} \sum_{l=1}^{m} a_{tj} a_{tr} a_{sk} a_{sl} E e_j e_r e_k e_l.$$

From the random vector $(e_1, \dots, e_m)$ definition obtain that $E e_j e_r e_k e_l \neq 0$ in one of the following four cases:
a) $j = r = k = l,\ E e_j e_r e_k e_l = \mu$, b) $j = k \neq r = l,\ E e_j e_r e_k e_l = 1$,
c) $j = r \neq k = l,\ E e_j e_r e_k e_l = 1$, d) $j = l \neq k = r,\ E e_j e_r e_k e_l = 1$. So

$$b(t,s) = \mu \sum_{j=1}^{m} a_{tj}^2 a_{sj}^2 + \sum_{1 \leq j,r \leq m,\ j \neq r} a_{tj} a_{tr} a_{sj} a_{sr} + \sum_{1 \leq j,k \leq m,\ j \neq k} a_{tj}^2 a_{sk}^2 +$$

$$+ \sum_{1 \leq j,k \leq m,\ j \neq k} a_{tj} a_{tk} a_{sk} a_{sj}.$$

Then

$$Var\ \hat{c}_n(t,s) = \frac{(\mu - 3) d(t,s) + c^2(t,s) + c(t,t) c(s,s)}{n},$$

$$d(t,s) = \sum_{j=1}^{m} a_{tj}^2 a_{sj}^2 \leq c(t,t) c(s,s).$$

If the components of the random vector $(e_1, \dots, e_m)$ have normal distribution then $\mu = 3$ and so

$$Var\ \hat{c}_n(t,s) = \frac{c^2(t,s) + c(t,t) c(s,s)}{n}.$$

Remark 4.1 *A suggested technique of a covariance estimate and a calculation of its variance by means of previous section formulas may be generalized to deviations of a multidimensional regression function which is a polynomial function of integer multidimensional argument.*

References

[1] Rozanov Yu. A. (1971). *Probability Theory, Random Processes and Mathematical Statistics*, M. Science (In Russian).

[2] Orlov A. I. (2004). *Mathematics of Chance: Probability and Statistics - Main Facts: Tutorial*, MZ-Press, Moscow (In Russian).

Paper received on 18 March 2008; revised, 10 April 2008; accepted, 22 May 2008.