APPLICATIONES MATHEMATICAE
28,2 (2001), pp. 143–149

Wiktor Oktaba and Joanna Tarasińska (Lublin)

ESTIMATION OF THE GENERALIZED VARIANCE IN A BIVARIATE NORMAL DISTRIBUTION FROM AN INCOMPLETE SAMPLE

Abstract. The aim of the paper is estimation of the generalized variance of a bivariate normal distribution in the case of a sample with missing observations. The estimator based on all available observations is compared with the estimator based only on complete pairs of observations.

1. Introduction. Let a random variable (y,z) have normal distribution with mean $\boldsymbol{\mu} = [\mu_1,\mu_2]'$ and variance-covariance matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_y^2 & \sigma_{yz} \\ \sigma_{yz} & \sigma_z^2 \end{bmatrix}$:

$$(y,z) \sim N_2\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \boldsymbol{\Sigma} \right). \tag{1}$$

Let $[\mathbf{y},\mathbf{z}]$ be a simple random sample of size k from the distribution (1). We are interested in estimation of the generalized variance, i.e. the determinant $|\boldsymbol{\Sigma}|$. The generalized variance is used in various statistical analyses concerning the covariance structure of the model.

The sample generalized variance

$$|\mathbf{S}| = \begin{vmatrix} \frac{1}{k-1}\sum_{i=1}^{k}(y_i-\overline{y})^2 & \frac{1}{k-1}\sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z}) \\ \frac{1}{k-1}\sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z}) & \frac{1}{k-1}\sum_{i=1}^{k}(z_i-\overline{z})^2 \end{vmatrix}, \tag{2}$$

where $\overline{y} = k^{-1}\sum_{i=1}^{k} y_i$, $\overline{z} = k^{-1}\sum_{i=1}^{k} z_i$, is very well investigated ([1], [7],

2000 *Mathematics Subject Classification*: 62F10, 62H12.
Key words and phrases: bivariate normal distribution, missing data, generalized variance, complete-case estimator, available-case estimator.

[5], [3], [4]). It is known for example that

$$\frac{|(k-1)\mathbf{S}|}{|\mathbf{\Sigma}|} = \chi^2_{k-1} \cdot \chi^2_{k-2},$$

where χ^2_{k-1} and χ^2_{k-2} are independently χ^2 distributed with $k-1$ and $k-2$ degrees of freedom, respectively. Thus

$$\text{(3)} \qquad \frac{k-1}{k-2}|\mathbf{S}| = \frac{1}{(k-1)(k-2)} \begin{vmatrix} \sum_{i=1}^{k}(y_i-\overline{y})^2 & \sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z}) \\ \sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z}) & \sum_{i=1}^{k}(z_i-\overline{z})^2 \end{vmatrix}$$

is an unbiased estimator of $|\mathbf{\Sigma}|$ and

$$\text{(4)} \qquad \operatorname{Var}\left(\frac{k-1}{k-2}|\mathbf{S}|\right) = \frac{2|\mathbf{\Sigma}|^2(2k-1)}{(k-1)(k-2)}.$$

2. Estimation of $|\mathbf{\Sigma}|$ in the case of missing observations. Let us consider an incomplete sample

$$\begin{bmatrix} y_1 & \dots & y_k & y_{k+1} & \dots & y_{k+p} & * & \dots & * \\ z_1 & \dots & z_k & * & \dots & * & z_{k+p+1} & \dots & z_{k+p+s} \end{bmatrix}',$$

where $*$ denotes an observation missing completely at random ([2], [6]). So, we have k complete pairs of observations, p additional observations of the y variable and s additional observations of the z variable. To simplify let us write the sample in the following form:

(5)

$\mathbf{y}_0$	$\mathbf{z}_0$
$\mathbf{y}_1$	$*$
$*$	$\mathbf{z}_2$

where $\mathbf{y}_0 = [\mathbf{y}_1, \dots, \mathbf{y}_k]'$, $\mathbf{z}_0 = [\mathbf{z}_1, \dots, \mathbf{z}_k]'$, $\mathbf{y}_1 = [\mathbf{y}_{k+1}, \dots, \mathbf{y}_{k+p}]'$, $\mathbf{z}_2 = [\mathbf{z}_{k+p+1}, \dots, \mathbf{z}_{k+p+s}]'$. Let us set

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{bmatrix}, \qquad \mathbf{z} = \begin{bmatrix} \mathbf{z}_0 \\ \mathbf{z}_2 \end{bmatrix}.$$

The question is: how should we estimate $|\mathbf{\Sigma}|$ using the additional information contained in the vectors $\mathbf{y}_1$ and $\mathbf{z}_2$ and is it worth doing? Perhaps the estimator based on complete pairs $[\mathbf{y}_0, \mathbf{z}_0]$ (complete-case estimator) is better?

As an alternative to the complete-case estimator we consider the available-case estimator which uses all the available values to estimate parameters in model (1). To estimate $|\mathbf{\Sigma}|$ we use the following sums:

$$(6)\quad \sum_{i=1}^{k+p}(y_i-\overline{y})^2,\quad \sum_{i=1}^{k}(z_i-\overline{z})^2+\sum_{i=k+p+1}^{k+p+s}(z_i-\overline{z})^2,\quad \sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z})$$

where $\overline{y}$ and $\overline{z}$ are the arithmetic means of elements of $\mathbf{y}$ and $\mathbf{z}$, respectively. Each of these sums, multiplied by a suitable constant, is a better unbiased estimator of σ_y^2 , σ_z^2 , σ_{yz} than the complete-case estimators

$$\frac{1}{k-1}\sum_{i=1}^{k}(y_i-\overline{y}_0)^2,\quad \frac{1}{k-1}\sum_{i=1}^{k}(z_i-\overline{z}_0)^2,\quad \frac{1}{k-1}\sum_{i=1}^{k}(y_i-\overline{y}_0)(z_i-\overline{z}_0),$$

where $\overline{y}_0$ and $\overline{z}_0$ are the means of $\mathbf{y}_0$ and $\mathbf{z}_0$.

Let us consider the following estimate of $|\mathbf{\Sigma}|$:

$$(7)\qquad E=a\cdot\sum_{i=1}^{k+p}(y_i-\overline{y})^2\cdot\Big[\sum_{i=1}^{k}(z_i-\overline{z})^2+\sum_{i=k+p+1}^{k+p+s}(z_i-\overline{z})^2\Big] - b\cdot\Big(\sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z})\Big)^2$$

where a and b are constants (depending on k,p,s) giving unbiasedness of E. To determine a and b and then to calculate the variance of E we use the results of Wilks [8]. He considered the following random variables for the incomplete sample (5):

$$\xi_0=\frac{1}{k+p}\sum_{i=1}^{k+p}(y_i-\overline{y})^2,\quad \eta_0=\frac{1}{k+s}\Big(\sum_{i=1}^{k}(z_i-\overline{z}\Big)^2+\sum_{i=k+p+1}^{k+p+s}(z_i-\overline{z})^2\Big),$$

$$\zeta_0=\frac{1}{k}\sum_{i=1}^{k}(y_i-\overline{y})(z_i-\overline{z}),$$

and found the moment generating function

$$\varphi(\gamma,\delta,\varepsilon)=E(e^{\gamma\xi_0+\delta\eta_0+\varepsilon\zeta_0}),$$

which can be used for finding joint moments of (ξ_0,η_0,ζ_0):

$$E(\xi_0^h\eta_0^k\zeta_0^l)=M(h,k,l)=\frac{\partial^h\partial^k\partial^l}{\partial\gamma^h\partial\delta^k\partial\varepsilon^l}\varphi(\gamma,\delta,\varepsilon)\Big|_{\gamma=\delta=\varepsilon=0}.$$

We have used $\varphi(\gamma,\delta,\varepsilon)$ to obtain the required moments of sums (6). All computations were done by using Maple V. The values of a and b are

$$a = \frac{2(k-1)+c+c^2+(k-1+c)^2}{(k+p-1)(k+s-1)[k-1+c^2+(k-1+c)^2]-2(k-1+c)^2},$$

$$b = \frac{(k+p-1)(k+s-1)+2(k-1+c)}{(k+p-1)(k+s-1)[k-1+c^2+(k-1+c)^2]-2(k-1+c)^2},$$

where $c = \frac{ps}{(k+p)(k+s)}$. When $s = 0$, a and b have a simpler form:

$$a = \frac{k+1}{(k-1)(k^2-k+pk-2)}, \quad b = \frac{k+p+1}{(k-1)(k^2-k+pk-2)}.$$

For a complete sample ($p = s = 0$) we have the known values

$$a = b = \frac{1}{(k-1)(k-2)}$$

(see (3)). The variance of E is

$$\begin{aligned}\mathrm{Var}(E) = {} & a^2(k+p)^2(k+s)^2[M(2,2,0)-M(1,1,0)^2] \\ & + b^2k^4[M(0,0,4)-M(0,0,2)^2] \\ & - 2abk^2(k+p)(k+s)[M(1,1,2)-M(1,1,0)\cdot M(0,0,2)].\end{aligned}$$

We do not give here the expressions for the moments $M(h,k,l)$ because they are long and complicated (especially $M(2,2,0)$, $M(0,0,4)$, $M(1,1,2)$). We are interested in comparing the estimator E given by (7) and the estimator E_0 based on complete pairs of observations:

$$E_0 = \frac{1}{(k-1)(k-2)}\Big[\sum_{i=1}^{k}(y_i-\overline{y}_0)^2\cdot\sum_{i=1}^{k}(z_i-\overline{z}_0)^2-\Big(\sum_{i=1}^{k}(y_i-\overline{y}_0)(z_i-\overline{z}_0)\Big)^2\Big].$$

When $s = 0$ we get a simple equation

$$\mathrm{Var}(E) - \mathrm{Var}(E_0) = \frac{-2p\sigma_y^4\sigma_z^4(k+1)[A\varrho^4+B\varrho^2+C]}{(k-2)(k-1)(k^2+pk-k-2)^2}, \tag{8}$$

where $A = 4(k+1)(k-2)+2pk$, $B = -2(k^2-4)(k+p+1)-4pk$, $C = (k-2)(k^2-1)+p(k^2-k+2)$ and ϱ is the correlation coefficient between y and z.

Superiority of one estimator over the other depends on ϱ^2, k, p, namely E is better when $\varrho^2 < f(k,p)$ and E_0 is better when $\varrho^2 > f(k,p)$, where $f(k,p)$ is the smaller root of the quadratic equation $Ax^2+Bx+C = 0$. Analysing $f(k,p)$ we can state the following simple corollary:

Corollary 1. *If $\varrho^2 \le 0.3$ than E is better than E_0 for each $k > 3$ and for each $p > 0$. If $\varrho^2 \ge 0.5$ then E_0 is better than E for each $k \ge 3$ and for each $p > 0$.*

The case $s = 0$ can be applied to the situation when getting an observation of one variable (for example z) is much more difficult or expensive than for the other (y). Suppose we have k complete pairs of observations. The question is: how large is p_0, the number of additional observations of y that cause at least the same decrease of variance of E as one additional complete pair? Using Maple V we get the following answer:

Corollary 2. • *If* $|\varrho| \leq 0.3$ *and* $k \geq 10$ *then* $p_0 = 3$.
• *If* $|\varrho| \leq 0.5$ *and* $k \geq 10$ *then* $p_0 = 5$.
• *If* $|\varrho| \leq 0.5$ *and* $k \geq 20$ *then* $p_0 = 3$.

When $s > 0$ then the difference $\mathrm{Var}(E) - \mathrm{Var}(E_0)$ is not so simple as in (8) and we do not give here the long expression for that. Let us only state that $\mathrm{Var}(E)$ is symmetric in p and s, that is,

$$\mathrm{Var}(E)_{(k,p,s)} = \mathrm{Var}(E)_{(k,s,p)}.$$

In Tables 1, 2, 3 and 4 we give the values of $\mathrm{Var}(E)/\mathrm{Var}(E_0)$ for various k, p, s and ϱ. The upper value in the tables is for $|\varrho| = 0.3$, the middle one for $|\varrho| = 0.5$ and the lower one for $|\varrho| = 0.8$.

So the estimator E can be either much better or much worse than E_0. E is not recommended when $|\varrho|$ is greater than 0.5. Unfortunately E has one disadvantage: theoretically it can have a negative value. We tried to estimate how often it can happen using Maple V simulation. We generated 1000 samples from a bivariate normal distribution with $\mu_1 = \mu_2 = 0$, $\sigma_y^2 = \sigma_z^2 = 1$, $\varrho = 0.5$ for different k, p, s. The results of this simulation in Table 5 show that the probability of getting negative values of E is small.

Table 1. $k = 10$

p / s	2	5	10	15
0	0.910 0.937 1.392	0.824 0.875 1.770	0.740 0.814 2.135	0.690 0.778 2.348
2	0.827 0.887 1.898	0.745 0.837 2.377	0.666 0.787 2.830	0.620 0.757 3.091
5		0.669 0.798 2.948	0.595 0.759 3.488	0.552 0.735 3.798
10			0.526 0.731 4.112	0.486 0.713 4.472
15				0.447 0.700 4.861

Table 2. $k = 20$

p / s	5	10	15	20
0	0.897 0.917 1.331	0.830 0.863 1.552	0.782 0.825 1.710	0.747 0.796 1.829
5	0.800 0.853 1.852	0.738 0.811 2.193	0.693 0.781 2.434	0.660 0.759 2.613
10		0.677 0.777 2.612	0.634 0.752 2.906	0.602 0.734 3.125
15			0.593 0.732 3.239	0.562 0.717 3.486
20				0.532 0.704 3.753

Table 3. $k = 50$

p / s	10	20	30	40	50
0	0.916 0.928 1.216	0.857 0.877 1.372	0.812 0.838 1.489	0.778 0.801 1.580	0.750 0.758 1.653
10	0.836 0.868 1.570	0.779 0.826 1.821	0.737 0.795 2.010	0.704 0.770 2.156	0.677 0.751 2.273
20		0.724 0.790 2.141	0.683 0.764 2.380	0.651 0.743 2.565	0.625 0.726 2.713
30			0.643 0.740 2.656	0.611 0.722 2.870	0.587 0.708 3.041
40				0.581 0.706 3.107	0.556 0.694 3.296
50					0.532 0.682 3.499

Table 4. $k = 100$

p / s	20	40	60	80	100
0	0.917 0.927 1.198	0.858 0.875 1.340	0.813 0.836 1.446	0.779 0.806 1.530	0.751 0.781 1.596
20	0.837 0.866 1.533	0.780 0.823 1.772	0.738 0.791 1.951	0.705 0.766 2.091	0.678 0.746 2.202
40		0.725 0.786 2.080	0.684 0.759 2.311	0.652 0.737 2.491	0.626 0.720 2.634
60			0.643 0.735 2.581	0.612 0.716 2.790	0.587 0.701 2.957
80				0.581 0.700 3.022	2.557 0.687 3.208
100					0.532 0.675 3.408

Table 5. The number of negative values of E (per 1000 samples)

$k = 10$					$k = 20$		$k = 50$	$k = 100$
$p = 5$ $s = 0$	$p = 5$ $s = 5$	$p = 10$ $s = 0$	$p = 10$ $s = 5$	$p = 10$ $s = 10$	$p = 10$ $s = 10$	$p = 20$ $s = 20$	$p = 50$ $s = 50$	$p = 100$ $s = 100$
3	18	10	21	40	0	6	0	0

References

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, Wiley, New York, 1958.

[2] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

[3] W. Oktaba, *Densities of determinant ratios, their moments and some simultaneous confidence intervals in the multivariate Gauss–Markoff model*, Appl. Math. 40 (1995), 47–54.

[4] —, *Asymptotically normal confidence intervals for a determinant in a generalized multivariate Gauss–Markoff model*, ibid., 55–59.

[5] C. R. Rao, *Linear Statistical Inference and Its Applications*, Wiley, New York, 1973.

[6] D. B. Rubin, *Inference and missing data*, Biometrika 63 (1976), 581–592.

[7] M. S. Srivastava and C. G. Khatri, *An Introduction to Multivariate Statistics*, North-Holland, New York, 1979.

[8] S. S. Wilks, *Moments and distributions of estimates of population parameters from fragmentary samples*, Ann. Math. Statist. 3 (1932), 163–195.

Department of Mathematical Statistics
Institute of Applied Mathematics
Academy of Agriculture
Akademicka 13
20-934 Lublin, Poland
E-mail: johata@ursus.ar.lublin.pl

Received on 28.10.1999;
revised version on 27.9.2000 (1508)