Calculate the generalized sample variance $|\mathbf{S}|$ for (a) the data matrix $\mathbf{X}$ in Exercise 3.1 and (b) the data matrix $\mathbf{X}$ in Exercise 3.2.

Step-by-step solution

Step 1 of 7

(a)

The 3×2 matrix is shown below:

$$X = \begin{bmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{bmatrix}$$

Firstly calculate the $\overline{\mathbf{x}}_i 's$ for each i as below:

$$\overline{x}_1 = \frac{9+5+1}{3}$$
$$= 5$$
$$\overline{x}_2 = \frac{3+1+2}{3}$$
$$= 2$$

Therefore, $\overline{x}l$ can be written as shown below:

$$\overline{x}l' = \begin{bmatrix} 5 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 5 & 5 & 5 \\ 2 & 2 & 2 \end{bmatrix}$$

$$X = \begin{pmatrix} 9 & 1 \\ 5 & 3 \\ 1 & 2 \end{pmatrix} \qquad \bar{x} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$$

Now, it is known that:

$$2S = (\bar{x} - \bar{x}1')(\bar{x} - \bar{x}1')' \text{ …… (*)}$$

$$d_i = y_i - \bar{x}_i \mathbf{1} \Leftrightarrow \boxed{\mathbf{d} = \mathbf{X} - \bar{x}\mathbf{1}}$$

First find $(\bar{x}' - \bar{x}1')$ as shown below:

$$d_i' d_k = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

$$\Rightarrow \boxed{d_i d_k' = n S_{ik}}$$

$$(\bar{x}' - \bar{x}1') = \begin{bmatrix} 9 & 5 & 1 \\ 1 & 3 & 2 \end{bmatrix} - \begin{bmatrix} 5 & 5 & 5 \\ 2 & 2 & 2 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 0 & -4 \\ -1 & 1 & 0 \end{bmatrix}$$

Therefore, (*) will become as follows:

$$\begin{pmatrix} 5 \\ 2 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \end{pmatrix} = \begin{pmatrix} 5 & 5 & 5 \\ 2 & 2 & 2 \end{pmatrix}$$

$$\bar{x} \cdot \mathbf{1}'$$

$$2S = (\bar{x} - \bar{x}1')(\bar{x} - \bar{x}1')'$$

$$2S = \begin{bmatrix} 4 & 0 & -4 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 4 & -1 \\ 0 & 1 \\ -4 & 0 \end{bmatrix}$$

$$2S = \begin{bmatrix} 32 & -4 \\ -4 & 2 \end{bmatrix}$$

$$S = \frac{1}{2} \begin{bmatrix} 32 & -4 \\ -4 & 2 \end{bmatrix}$$

$$S = \begin{bmatrix} 16 & -2 \\ -2 & 1 \end{bmatrix}$$

Hence, calculate generalized sample variance $|S|$ as shown below:

$$|S| = 16 \times 1 - (-2 \times -2)$$

$$= 16 - 4$$

$$= \boxed{12}$$

(b)

The 3×2 matrix is shown below:

$$X = \begin{bmatrix} 3 & 4 \\ 6 & -2 \\ 3 & 1 \end{bmatrix}$$

Firstly calculate the $\overline{x}_i$'s for each i as below:

$$\overline{x}_1 = \frac{3+3+3}{3}$$
$$= 4$$
$$\overline{x}_2 = \frac{4-2+1}{3}$$
$$1$$

Therefore, $\overline{x}1$ can be written as shown below:

$$\overline{x}1' = \begin{bmatrix} 4 \\ 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} 4 & 4 & 4 \\ 1 & 1 & 1 \end{bmatrix}$$

Now, it is known that:

$$2S = \left(\bar{x} - \bar{x}I'\right)\left(\bar{x} - \bar{x}I'\right)' \text{ ...... } (*)$$

First find $\left(\bar{x}' - \bar{x}I'\right)$ as shown below:

$$\left(\bar{x}' - \bar{x}I'\right) = \begin{bmatrix} 3 & 6 & 3 \\ 4 & -2 & 1 \end{bmatrix} - \begin{bmatrix} 4 & 4 & 4 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} -1 & 2 & -1 \\ 3 & -3 & 0 \end{bmatrix}$$

Therefore, (*) will become as follows:

$$2S=(\bar{x}-\bar{x}I')(\bar{x}-\bar{x}I')'$$

$$2S=\begin{bmatrix}-1 & 2 & -1\\ 3 & -3 & 0\end{bmatrix}\begin{bmatrix}-1 & 3\\ 2 & -3\\ -1 & 0\end{bmatrix}$$

$$2S=\begin{bmatrix}6 & -9\\ -9 & 18\end{bmatrix}$$

$$S=\frac{1}{2}\begin{bmatrix}6 & -9\\ -9 & 18\end{bmatrix}$$

$$S=\begin{bmatrix}3 & -9/2\\ -9/2 & 9\end{bmatrix}$$

Hence, calculate generalized sample variance $|S|$ as shown below:

$$\begin{aligned}|S|&=(3\times 9)-\left(-\frac{9}{2}\times-\frac{9}{2}\right)\\ &=\frac{27}{1}-\frac{81}{4}\\ &=\frac{108-81}{4}\\ &=\boxed{\frac{27}{4}}\end{aligned}$$

Result 3.2. The generalized variance is zero when, and only when, at least one deviation vector lies in the (hyper) plane formed by all linear combinations of the others—that is, when the columns of the matrix of deviations in (3-18) are linearly dependent.

Proof. If the columns of the deviation matrix $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ are linearly dependent, there is a linear combination of the columns such that

$$\begin{aligned}\mathbf{0} &= a_1 \operatorname{col}_1(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') + \cdots + a_p \operatorname{col}_p(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') \\ &= (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} \qquad \text{for some } \mathbf{a} \neq \mathbf{0}\end{aligned}$$

But then, as you may verify, $(n-1)\mathbf{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ and

$$(n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$$

so the same $\mathbf{a}$ corresponds to a linear dependency, $a_1 \operatorname{col}_1(\mathbf{S}) + \cdots + a_p \operatorname{col}_p(\mathbf{S}) = \mathbf{S}\,\mathbf{a} = \mathbf{0}$, in the columns of $\mathbf{S}$. So, by Result 2A.9, $|\mathbf{S}| = 0$.

In the other direction, if $|\mathbf{S}| = 0$, then there is some linear combination $\mathbf{S}\mathbf{a}$ of the columns of $\mathbf{S}$ such that $\mathbf{S}\mathbf{a} = \mathbf{0}$. That is, $\mathbf{0} = (n-1)\mathbf{S}\mathbf{a} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}$. Premultiplying by $\mathbf{a}'$ yields

$$0 = \mathbf{a}'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = L^2_{(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}}$$

and, for the length to equal zero, we must have $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = \mathbf{0}$. Thus, the columns of $(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$ are linearly dependent. ■

Example 3.10 (Creating new variables that lead to a zero generalized variance) Consider the data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & 9 & 10 \\ 4 & 12 & 16 \\ 2 & 10 & 12 \\ 5 & 8 & 13 \\ 3 & 11 & 14 \end{bmatrix}$$

where the third column is the sum of first two columns. These data could be the number of successful phone solicitations per day by a part-time and a full-time employee, respectively, so the third column is the total number of successful solicitations per day.

Show that the generalized variance $|\mathbf{S}| = 0$, and determine the nature of the dependency in the data.

We find that the mean corrected data matrix, with entries $x_{jk} - \bar{x}_k$, is

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}' = \begin{bmatrix} -2 & -1 & -3 \\ 1 & 2 & 3 \\ -1 & 0 & -1 \\ 2 & -2 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

The resulting covariance matrix is

$$\mathbf{S} = \begin{bmatrix} 2.5 & 0 & 2.5 \\ 0 & 2.5 & 2.5 \\ 2.5 & 2.5 & 5.0 \end{bmatrix}$$

We verify that, in this case, the generalized variance

$$|\mathbf{S}| = 2.5^2 \times 5 + 0 + 0 - 2.5^3 - 2.5^3 - 0 = 0$$

The following data matrix contains data on test scores, with x_1 = score on first test, x_2 = score on second test, and x_3 = total score on the two tests:

$$\mathbf{X} = \begin{bmatrix} 12 & 17 & 29 \\ 18 & 20 & 38 \\ 14 & 16 & 30 \\ 20 & 18 & 38 \\ 16 & 19 & 35 \end{bmatrix}$$

(a) Obtain the mean corrected data matrix, and verify that the columns arc linearly dependent. Specify an $\mathbf{a}' = [a_1, a_2, a_3]$ vector that establishes the linear dependence.

(b) Obtain the sample covariance matrix $\mathbf{S}$, and verify that the generalized variance is zero. Also, show that $\mathbf{Sa} = \mathbf{0}$, so $\mathbf{a}$ can be resealed to be an eigenvector corresponding to eigenvalue zero.

(c) Verify that the third column of the data matrix is the sum of the first two columns. That is, show that there is linear dependence, with $a_1 = 1$, $a_2 = 1$, and $a_3 = -1$.

The data matrix is shown below:

$$X = \begin{bmatrix} 12 & 17 & 29 \\ 18 & 20 & 38 \\ 14 & 16 & 30 \\ 20 & 18 & 38 \\ 16 & 19 & 35 \end{bmatrix}$$

(a)

Firstly, calculate the $\overline{x}_i\text{'s}$ for each $i=1,2,3,4,5$ as below:

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{i=1}^{5} X_i \\
&= \frac{1}{5}\left(\begin{bmatrix}12\\17\\29\end{bmatrix}+\begin{bmatrix}18\\20\\38\end{bmatrix}+\begin{bmatrix}14\\16\\30\end{bmatrix}+\begin{bmatrix}20\\18\\38\end{bmatrix}+\begin{bmatrix}16\\19\\35\end{bmatrix}\right) \\
&= \frac{1}{5}\left(\begin{bmatrix}80\\90\\170\end{bmatrix}\right) \\
&= \begin{bmatrix}16\\18\\34\end{bmatrix}
\end{aligned}$$

Or,

$$\overline{x}' = [16,18,34]'$$

Now, calculate the each sample variances for i. For this, first calculate the $X_c = X - \overline{x}'$for sample variance. So,

$$\begin{aligned}
X_c &= X - \overline{x}' \\
&= \begin{bmatrix}12 & 17 & 29\\18 & 20 & 38\\14 & 16 & 30\\20 & 18 & 38\\16 & 19 & 35\end{bmatrix} - \begin{bmatrix}16 & 18 & 34\end{bmatrix} \\
&= \begin{bmatrix}-4 & -1 & -5\\2 & 2 & 4\\-2 & -2 & -4\\4 & 0 & 4\\0 & 1 & 1\end{bmatrix}
\end{aligned}$$

And, it is noticeable that:

$$Col_1(X_c) + Col_2(X_c) = Col_3(X_c)$$

Therefore "*a*" should be equal to:

$$[1,1,-1]'$$

That gives $X_c a = 0$

(b)

Now, calculate the sample covariance as shown below:

$$S = \frac{1}{n-1} X_c{}' X_c$$

$$= \frac{1}{5-1}\left(\begin{bmatrix} -4 & 2 & -2 & 4 & 0 \\ -1 & 2 & -2 & 0 & 1 \\ -5 & 4 & -4 & 4 & 1 \end{bmatrix} \times \begin{bmatrix} -4 & -1 & -5 \\ 2 & 2 & 4 \\ -2 & -2 & -4 \\ 4 & 0 & 4 \\ 0 & 1 & 1 \end{bmatrix}\right)$$

$$= \frac{1}{4}\begin{bmatrix} 40 & 12 & 52 \\ 12 & 10 & 22 \\ 52 & 22 & 74 \end{bmatrix}$$

$$= \begin{bmatrix} 10 & 3 & 13 \\ 3 & 10/4 & 22/4 \\ 13 & 22/4 & 74/4 \end{bmatrix}$$

$$= \boxed{\begin{bmatrix} 10 & 3 & 13 \\ 3 & 2.5 & 5.5 \\ 13 & 5.5 & 18.5 \end{bmatrix}}$$

Now verify for $|S|=0$ as shown below:

$$
\begin{aligned}
|S| &= 10\left((2.5\times 18.5)-(5.5)^2\right)(-1)^2+3\left((3\times 18.5)-(13\times 5.5)\right)(-1)^3 \\
&\quad +13\left((3\times 5.5)-(2.5\times 13)\right)(-1)^4 \\
&= 10(16)-3(-16)+13(-16) \\
&= 160+48+208 \qquad 160+48-208 \\
&= 0
\end{aligned}
$$

Hence proved

To show that $Sa=0$, do calculations a shown below:

$$
\begin{aligned}
Sa &= \begin{bmatrix} 10 & 3 & 13 \\ 3 & 2.5 & 5.5 \\ 13 & 5.5 & 18.5 \end{bmatrix} \times \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \\
&= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}
\end{aligned}
$$

It is also proved.

(c)

It is known that whenever the columns of the mean corrected data matrix are linearly independent then equation can be written as shown below:

$$\begin{aligned}(n-1)Sa &= (X - I\,\bar{x}')' \times (X - I\,\bar{x}')a \\ &= (X - I\,\bar{x}')0 \\ &= 0\end{aligned}$$

And,

$$Sa = 0$$

Here,

$$a' = [1,1,-1]'$$

Therefore,

$$\begin{aligned}Sa &= \begin{bmatrix} 10 & 3 & 13 \\ 3 & 2.5 & 5.5 \\ 13 & 5.5 & 18.5 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}\end{aligned}$$

The coefficients reveal that for all *j*:

$$1(x_{j1} - \bar{x}_1) + 1(x_{j2} - \bar{x}_2) + (-1)(x_{j3} - \bar{x}_3) = 0$$

In addition, the third variable is actually the sum of first and second variable. So, it can conclude that the data matrix satisfy a linear constraint with $c = 0$

Example 3.11 (Illustrating the relation between $|\mathbf{S}|$ and $|\mathbf{R}|$) Let us illustrate the relationship in (3-21) for the generalized variances $|\mathbf{S}|$ and $|\mathbf{R}|$ when $p = 3$. Suppose

$$\underset{(3\times 3)}{\mathbf{S}} = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}$$

Then $s_{11} = 4$, $s_{22} = 9$, and $s_{33} = 1$. Moreover,

$$\mathbf{R} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{2}{3} \\ \frac{1}{2} & \frac{2}{3} & 1 \end{bmatrix}$$

Using Definition 2A.24, we obtain

$$|\mathbf{S}| = 4\begin{vmatrix} 9 & 2 \\ 2 & 1 \end{vmatrix}(-1)^2 + 3\begin{vmatrix} 3 & 2 \\ 1 & 1 \end{vmatrix}(-1)^3 + 1\begin{vmatrix} 3 & 9 \\ 1 & 2 \end{vmatrix}(-1)^4$$

$$= 4(9 - 4) - 3(3 - 2) + 1(6 - 9) = 14$$

$$|\mathbf{R}| = 1\begin{vmatrix} 1 & \frac{2}{3} \\ \frac{2}{3} & 1 \end{vmatrix}(-1)^2 + \frac{1}{2}\begin{vmatrix} \frac{1}{2} & \frac{2}{3} \\ \frac{1}{2} & 1 \end{vmatrix}(-1)^3 + \frac{1}{2}\begin{vmatrix} \frac{1}{2} & 1 \\ \frac{1}{2} & \frac{2}{3} \end{vmatrix}(-1)^4$$

$$= \left(1 - \tfrac{4}{9}\right) - \left(\tfrac{1}{2}\right)\left(\tfrac{1}{2} - \tfrac{1}{3}\right) + \left(\tfrac{1}{2}\right)\left(\tfrac{1}{3} - \tfrac{1}{2}\right) = \tfrac{7}{18}$$

It then follows that

$$14 = |\mathbf{S}| = s_{11}s_{22}s_{33}|\mathbf{R}| = (4)(9)(1)\left(\tfrac{7}{18}\right) = 14 \qquad \text{(check)}$$

■

Consider the data matrix

$$\mathbf{X} = \begin{bmatrix} -1 & 3 & -2 \\ 2 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$$

(a) Calculate the matrix of deviations (residuals). $\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$. Is this matrix of full rank? Explain.

(b) Determine **S** and calculate the generalized sample variance |**S**|. Interpret the latter geometrically.

(c) Using the results in (b), calculate the total sample variance. [See (3-23).]

The 3×3 matrix is shown below:

$$X = \begin{bmatrix} -1 & 3 & -2 \\ 2 & 4 & 2 \\ 5 & 2 & 3 \end{bmatrix}$$

Comment

(a)

Matrix of deviations can be represented by $X - I\,\overline{x}'$

Firstly calculate the $\overline{x}_i$'s for each $i = 1,2,3$ as below:

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{3} X_i \\ &= \frac{1}{3}\left(\begin{bmatrix} -1 \\ 3 \\ -2 \end{bmatrix} + \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix} + \begin{bmatrix} 5 \\ 2 \\ 3 \end{bmatrix}\right) \\ &= \frac{1}{3}\left(\begin{bmatrix} 6 \\ 9 \\ 3 \end{bmatrix}\right) \\ &= \begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix} \end{aligned}$$

And,

$$\overline{x}' = \begin{bmatrix} 2 & 3 & 1 \end{bmatrix}$$

Now put the value of $\overline{x}'$ in above equation. Here, $X - I\,\overline{x}'$ can be calculated as shown below:

$$\begin{aligned} X - I\,\overline{x}' &= \begin{bmatrix} -1-2 & 3-3 & -2-1 \\ 2-2 & 4-3 & 2-1 \\ 5-2 & 2-3 & 3-1 \end{bmatrix} \\ &= \begin{bmatrix} -3 & 0 & -3 \\ 0 & 1 & 1 \\ 3 & -1 & 2 \end{bmatrix} \end{aligned}$$

The deviation (column) vectors are given below:

$$\begin{aligned} d_1' &= [-3, 0, 3] \\ d_2' &= [0, 1, -1] \\ d_3' &= [-3, 1, 2] \end{aligned}$$

Since $d_1 = d_2 = d_3$, the matrix of deviations is not of full rank.

(b)

Here, it is known that:

$$2S = (\bar{x} - I\,\bar{x}_1')(\bar{x} - I\,\bar{x}_1')' \text{ ...... (*)}$$

Therefore, (*) will become as follows:

$$2S = (\bar{x} - \bar{x}I')(\bar{x} - \bar{x}I')'$$

$$2S = \begin{bmatrix} -3 & 0 & -3 \\ 0 & 1 & 1 \\ 3 & -1 & 2 \end{bmatrix} \begin{bmatrix} -3 & 0 & 3 \\ 0 & 1 & -1 \\ -3 & 1 & 2 \end{bmatrix}$$

$$2S = \begin{bmatrix} 18 & -3 & 15 \\ -3 & 2 & -1 \\ 15 & -1 & 14 \end{bmatrix}$$

$$S = \frac{1}{2}\begin{bmatrix} 18 & -3 & 15 \\ -3 & 2 & -1 \\ 15 & -1 & 14 \end{bmatrix}$$

$$S = \begin{bmatrix} 9 & -3/2 & 15/2 \\ -3/2 & 1 & -1/2 \\ 15/2 & -1/2 & 7 \end{bmatrix}$$

Now verify for $|S| = 0$ as shown below:

$$\begin{aligned}
|S| &= 9\begin{vmatrix} 1 & -1/2 \\ -1/2 & 7 \end{vmatrix}(-1)^2 + \left(-\frac{3}{2}\right)\begin{vmatrix} -3/2 & -1/2 \\ 15/2 & 7 \end{vmatrix}(-1)^3 + \frac{15}{2}\begin{vmatrix} -3/2 & 1 \\ 15/2 & 1/2 \end{vmatrix}(-1)^4 \\
&= 9\left(7 - \frac{1}{4}\right) + \frac{3}{2}\left(-\frac{21}{2} + \frac{15}{4}\right) + \frac{15}{2}\left(\frac{3}{4} - \frac{15}{2}\right) \\
&= \frac{243}{4} - \frac{81}{8} - \frac{405}{8} \\
&= 0
\end{aligned}$$

It can say that the 3-deviation vectors lie in a 2-dimensional space. The 3-dimensional volume enclosed by the deviation vectors zero.

(c)

The total sample variance can be calculated as shown below:

$$S_{11}+S_{22}+\ldots+S_{pp}$$

Therefore, the total sample variance is:

$$9+1+7=\boxed{17}$$