The Matrix Algebra of Sample Statistics

James H. Steiger

Department of Psychology and Human Development
Vanderbilt University

The Matrix Algebra of Sample Statistics

Introduction

In this section, we show how matrix algebra can be used to express some common statistical formulas in a succinct way that allows us to derive some important results in multivariate analysis.

The Data Matrix

- Suppose we wish to discuss a set of sample data representing scores for n people on p variables.
- We can represent the people in rows and the variables in columns, or vice-versa.
- Placing the variables in columns seems like a more natural way to do things for the modern computer user, as most computer files for standard statistical software represent the "cases" as rows, and the variables as columns.
- Ultimately, we will develop the ability to work with both notational variations, but for the time being, we'll work with our data in "column form," i.e., with the variables in columns. Consequently, our standard notation for a data matrix is ${}_n\mathbf{X}_p$.

Converting to Deviation Scores

- Suppose $\mathbf{x}$ is an $n \times 1$ matrix of scores for n people on a single variable. We wish to transform the scores in $\mathbf{x}$ to *deviation score form*. (In general, we will find this a source of considerable convenience.)
- To accomplish the deviation score transformation, the arithmetic mean $\overline{x}_\bullet$, must be subtracted from each score in $\mathbf{x}$.

Converting to Deviation Scores

- Let $\mathbf{1}$ be a $n \times 1$ vector of ones. We will refer to such a vector on occasion as a "summing vector," for the following reason.
- Consider any vector x, for example a 3×1 column vector with the numbers 1, 2, 3. If we compute $\mathbf{1}'\mathbf{x}$, we are taking the sum of cross-products of a set of 1's with the numbers in x.
- In summation notation,

$$\mathbf{1}'\mathbf{x} = \sum_{i=1}^{n} 1_i x_i = \sum_{i=1}^{n} x_i$$

- So $\mathbf{1}'\mathbf{x}$ is how we express "the sum of the x's" in matrix notation.

Converting to Deviation Scores

Consequently,

$$\overline{x}_\bullet = (1/n)\mathbf{1}'\mathbf{x}$$

To transform $\mathbf{x}$ to deviation score form, we need to subtract $\overline{x}_\bullet$ from every element of $\mathbf{x}$. We can easily construct a vector with every element equal to $\overline{x}_\bullet$ by simply multiplying the scalar $\overline{x}_\bullet$ by a summing vector.

Converting to Deviation Scores

Consequently, if we denote the vector of deviation scores as $\mathbf{x}^*$, we have

$$
\begin{aligned}
\mathbf{x}^* &= \mathbf{x} - \mathbf{1}\bar{x}_\bullet \\
&= \mathbf{x} - \mathbf{1}\left(\frac{\mathbf{1}'\mathbf{x}}{n}\right) && (1) \\
&= \mathbf{x} - \frac{\mathbf{1}\mathbf{1}'}{n}\mathbf{x} \\
&= \mathbf{x} - \left(\frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x} \\
&= \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}'}{n}\right)\mathbf{x} && (2) \\
&= (\mathbf{I} - \mathbf{P})\,\mathbf{x} && (3) \\
\mathbf{x}^* &= \mathbf{Q}\mathbf{x} && (4)
\end{aligned}
$$

where

$$\mathbf{Q} = \mathbf{I} - \mathbf{P}$$

and

$$\mathbf{P} = \frac{\mathbf{1}\mathbf{1}'}{n}$$

Converting to Deviation Scores

1. You should study the above derivation carefully, making certain you understand all steps.
2. You should carefully verify that the matrix $\mathbf{11}'$ is an $n \times n$ matrix of 1's, so the expression $\mathbf{11}'/n$ is an $n \times n$ matrix with each element equal to $1/n$(Division of matrix by a non-zero scalar is a special case of a scalar multiple, and is perfectly legal).
3. Since $\mathbf{x}$ can be converted from raw score form to deviation score form by pre-multiplication with a single matrix, it follows that any *particular* deviation score can be computed with one pass through a list of numbers.
4. We would probably never want to compute deviation scores in practice using the above formula, as it would be inefficient. However, the formula does allow us to see some interesting things that are difficult to see using scalar notation (more about that later).
5. If one were, for some reason, to write a computer program using Equation 4, one would not need (or want) to save the matrix $\mathbf{Q}$, for several reasons. First, it can be very large! Second, no matter how large n is, the elements of $\mathbf{Q}$ take on only two distinct values. Diagonal elements of $\mathbf{Q}$ are always equal to $(n-1)/n$, and off-diagonal elements are always equal to $-1/n$. In general, there would be no need to store the numbers.

Example

Example (The Deviation Score Projection Operator)

Any vector of n raw scores can be converted into deviation score form by pre-multiplication by a "projection operator" $\mathbf{Q}$. Diagonal elements of $\mathbf{Q}$ are always equal to $(n-1)/n$, and off-diagonal elements are always equal to $-1/n$. Suppose we have the vector

$$\mathbf{x} = \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix}$$

Construct a projection operator $\mathbf{Q}$ such that $\mathbf{Qx}$ will be in deviation score form.

Solution

Example (Solution)

We have

$$\mathbf{Qx} = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 4 \\ 2 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ 0 \\ -2 \end{bmatrix}$$

Example

Example (Computing the *i*th Deviation Score)

An implication of the preceding result is that one can compute the *i*th deviation score as a single linear combination of the *n* scores in a list. For example, the 3rd deviation score in a list of 3 is computed as $[dx]_3 = -1/3x_1 - 1/3x_2 + 2/3x_3$.

Question. Does that surprise you?

Properties of the Deviation Score Operators

Let us now investigate the properties of the matrices $\mathbf{P}$ and $\mathbf{Q}$ that accomplish this transformation. First, we should establish an additional definition and result.

Definition. A matrix $\mathbf{C}$ is *idempotent* if $\mathbf{C}^2 = \mathbf{CC} = \mathbf{C}$.

Lemma. If $\mathbf{C}$ is idempotent and $\mathbf{I}$ is a conformable identity matrix, then $\mathbf{I} - \mathbf{C}$ is also idempotent. *Proof.* To prove the result, we need merely show that $(\mathbf{I} - \mathbf{C})^2 = (\mathbf{I} - \mathbf{C})$. This is straightforward.

$$\begin{aligned}(\mathbf{I} - \mathbf{C})^2 &= (\mathbf{I} - \mathbf{C})(\mathbf{I} - \mathbf{C}) \\ &= \mathbf{I}^2 - \mathbf{CI} - \mathbf{IC} + \mathbf{C}^2 \\ &= \mathbf{I} - \mathbf{C} - \mathbf{C} + \mathbf{C} \\ &= \mathbf{I} - \mathbf{C}\end{aligned}$$

□

Idempotency of **P**

Recall that **P** is an $n \times n$ symmetric matrix with each element equal to $1/n$. **P** is also idempotent, since:

$$\begin{aligned}
\mathbf{PP} &= \frac{\mathbf{11}'}{n}\frac{\mathbf{11}'}{n} \\
&= \frac{\mathbf{11}'\mathbf{11}'}{n^2} \\
&= \frac{\mathbf{1}\left(\mathbf{1}'\mathbf{1}\right)\mathbf{1}'}{n^2} \\
&= \frac{\mathbf{1}\left(n\right)\mathbf{1}'}{n^2} \\
&= \frac{\mathbf{11}'\left(n\right)}{n^2} \\
&= \mathbf{11}'\frac{n}{n^2} \\
&= \frac{\mathbf{11}'}{n} \\
&= \mathbf{P}
\end{aligned}$$

Some General Principles

The preceding derivation demonstrates some principles that are generally useful in reducing simple statistical formulas in matrix form:

1. Scalars can be "moved through" matrices to any position in the expression that is convenient.
2. Any expression of the form $\mathbf{x}'\mathbf{y}$ is a scalar product, and hence it is a scalar, and can be moved intact through other matrices in the expression. So, for example, we recognized that $\mathbf{1}'\mathbf{1}$ is a scalar and can be reduced and eliminated in the above derivation.

You may easily verify the following properties:

1. The matrix $\mathbf{Q} = \mathbf{I} - \mathbf{P}$ is also symmetric and idempotent. (Hint: Use a theorem we proved a few slides back.)
2. $\mathbf{Q1} = \mathbf{0}$ (Hint: First prove that $\mathbf{P1} = \mathbf{1}$.)

The Sample Variance

Since the sample variance S_X^2 is defined as the sum of squared deviations divided by $n-1$, it is easy to see that, if scores in a vector $\mathbf{x}$ are in deviation score form, then the sum of squared deviations is simply $\mathbf{x}^{*\prime}\mathbf{x}^{*}$, and the sample variance may be written

$$S_X^2 = 1/(n-1)\mathbf{x}^{*\prime}\mathbf{x}^{*} \tag{5}$$

If $\mathbf{x}$ is not in deviation score form, we may use the $\mathbf{Q}$ operator to convert it into deviation score form first. Hence, in general,

$$\begin{aligned} S_X^2 &= 1/(n-1)\mathbf{x}^{*\prime}\mathbf{x}^{*} \\ &= 1/(n-1)(\mathbf{Q}\mathbf{x})^{\prime}\mathbf{Q}\mathbf{x} \\ &= 1/(n-1)\mathbf{x}^{\prime}\mathbf{Q}^{\prime}\mathbf{Q}\mathbf{x}, \end{aligned}$$

since the transpose of a product of two matrices is the product of their transposes in reverse order.

The Sample Covariance

The expression can be reduced further. Since $\mathbf{Q}$ is symmetric, it follows immediately that $\mathbf{Q}' = \mathbf{Q}$, and (remembering also that $\mathbf{Q}$ is idempotent) that $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}$. Hence

$$S_X^2 = 1/(n-1)\ \mathbf{x}'\mathbf{Q}\mathbf{x}$$

As an obvious generalization of the above, we write the matrix form for the covariance between two vectors of scores $\mathbf{x}$ and $\mathbf{y}$ as

$$S_{XY} = 1/(n-1)\mathbf{x}'\mathbf{Q}\mathbf{y}$$

A Surprising but Useful Result

Some times a surprising result is staring us right in the face, if we are only able to see it. Notice that the sum of cross products of deviation scores can be computed as

$$\begin{aligned} \mathbf{x}^{*\prime}\mathbf{y}^{*} &= (\mathbf{Q}\mathbf{x})^{\prime}(\mathbf{Q}\mathbf{y}) \\ &= \mathbf{x}^{\prime}\mathbf{Q}^{\prime}\mathbf{Q}\mathbf{y} \\ &= (\mathbf{x}^{\prime}\mathbf{Q})\,\mathbf{y} \\ &= \mathbf{x}^{\prime}(\mathbf{Q}\mathbf{y}) \\ &= \mathbf{x}^{\prime}\mathbf{y}^{*} \\ &= \mathbf{y}^{\prime}\mathbf{x}^{*} \end{aligned}$$

Because products of the form $\mathbf{QQ}$ or $\mathbf{QQ}^{\prime}$ can be collapsed into a single $\mathbf{Q}$, when computing the sum of cross products of deviation scores of two variables, one variable can be left in raw score form and the sum of cross products will remain the same! This surprising result is somewhat harder to see (and prove) using summation algebra.

A Standard Assumption

In what follows, we will generally assume, unless explicitly stated otherwise, that our data matrices have been transformed to deviation score form. (The $\mathbf{Q}$ operator discussed above will accomplish this simultaneously for the case of scores of n subjects on several, say p, variates.) For example, consider a data matrix ${}_n\mathbf{X}_p$, whose p columns are the scores of n subjects on p different variables. If the columns of $\mathbf{X}$ are in raw score form, the matrix $\mathbf{X}^* = \mathbf{QX}$ will have p columns of deviation scores.

Column Variate Form

- We shall concentrate on results in the case where $\mathbf{X}$ is in "column variate form," i.e., is an $n \times p$ matrix.
- Equivalent results may be developed for "row variate form" $p \times n$ data matrices which have the n scores on p variables arranged in p rows.
- The choice of whether to use row or column variate representations is arbitrary, and varies in books and articles, although column variate form is far more common.
- One must, ultimately, be equally fluent with either notation.

The Variance-Covariance Matrix

- Consider the case in which we have n scores on p variables. We define the *variance-covariance matrix* $\mathbf{S_{xx}}$ to be a symmetric $p \times p$ matrix with element s_{ij} equal to the covariance between variable i and variable j.
- Naturally, the ith diagonal element of this matrix contains the covariance of variable i with itself, i.e., its variance.
- As a generalization of our results for a single vector of scores, the variance-covariance matrix may be written as follows. First, for raw scores in column variate form:

$$\mathbf{S_{xx}} = 1/(n-1)\mathbf{X'QX}$$

- We obtain a further simplification if $\mathbf{X}$ is in deviation score form. In that case, we have:

$$\mathbf{S_{xx}} = 1/(n-1)\mathbf{X'X}$$

- Note that some authors use the terms "variance-covariance matrix" and "covariance matrix" interchangeably.

The Correlation Matrix

- For p variables in the data matrix X, the *correlation matrix* $\mathbf{R_{xx}}$ is a $p \times p$ symmetric matrix with typical element r_{ij} equal to the correlation between variables i and j.
- Of course, the diagonal elements of this matrix represent the correlation of a variable with itself, and are all equal to 1.
- Recall that all of the elements of the variance-covariance matrix $\mathbf{S_{xx}}$ are covariances, since the variances are covariances of variables with themselves. We know that, in order to convert s_{ij} (the covariance between variables i and j) to a correlation, we simply "standardize" it by dividing by the product of the standard deviations of variables i and j.
- This is very easy to accomplish in matrix notation.

The Correlation Matrix

- Specifically, let $\mathbf{D_{xx}} = \text{diag}(\mathbf{S_{xx}})$ be a diagonal matrix with ith diagonal element equal to the variance of the ith variable in $\mathbf{X}$.
- Then let $\mathbf{D}^{1/2}$ be a diagonal matrix with elements equal to standard deviations, and $\mathbf{D}^{-1/2}$ be a diagonal matrix with ith diagonal element equal to $1/s_i$, where s_i is the standard deviation of the ith variable.
- Then the correlation matrix is computed as:

$$\mathbf{R_{xx}} = \mathbf{D}^{-1/2}\mathbf{S_{xx}}\mathbf{D}^{-1/2}$$

- Let's verify this on the board.

The Covariance Matrix

Given ${}_n\mathbf{X}_m$ and ${}_n\mathbf{Y}_p$, two data matrices in deviation score form. The *covariance matrix* $\mathbf{S_{xy}}$ is a $m \times p$ matrix with element s_{ij} equal to the covariance between the ith variable in $\mathbf{X}$ and the jth variable in $\mathbf{Y}$. $\mathbf{S_{xy}}$ is computed as

$$\mathbf{S_{xy}} = 1/(n-1)\mathbf{X}'\mathbf{Y}$$

Variance of a Linear Combination

Earlier, we developed a summation algebra expression for evaluating the variance of a linear combination of variables. In this section, we derive the same result using matrix algebra.

We first note the following result.

Lemma. Given $\mathbf{X}$, a data matrix in column variate deviation score form. For any linear composite $\mathbf{y} = \mathbf{Xb}$, $\mathbf{y}$ will also be in deviation score form.

Proof. The variables in $\mathbf{X}$ are in deviation score form if and only if the sum of scores in each column is zero, i.e., $\mathbf{1}'\mathbf{X} = \mathbf{0}'$. But if $\mathbf{1}'\mathbf{X} = \mathbf{0}'$, then for any linear combination $\mathbf{y} = \mathbf{Xb}$, we have, immediately,

$$\begin{aligned} \mathbf{1}'\mathbf{y} &= \mathbf{1}'\mathbf{Xb} \\ &= (\mathbf{1}'\mathbf{X})\,\mathbf{b} \\ &= \mathbf{0}'\mathbf{b} \\ &= 0 \end{aligned}$$

Since, for any $\mathbf{b}$, the linear combination scores in $\mathbf{y}$ sum to zero, it must be in deviation score form. □

Variance of a Linear Combination

We now give a result that is one of the cornerstones of multivariate statistics. *Theorem*. Given $\mathbf{X}$, a set of n deviation scores on p variables in column variate form, having variance-covariance matrix $\mathbf{S_{xx}}$. The variance of any linear combination $\mathbf{y} = \mathbf{Xb}$ may be computed as

$$S_y^2 = \mathbf{b}'\mathbf{S_{xx}}\mathbf{b} \tag{6}$$

Proof. Suppose $\mathbf{X}$ is in deviation score form. Then, by a previous Lemma, so must $\mathbf{y} = \mathbf{Xb}$, for any $\mathbf{b}$. From the formula for the sample variance, we know that

$$\begin{aligned} S_y^2 &= 1/(n-1)\ \mathbf{y}'\mathbf{y} \\ &= 1/(n-1)\ \ (\mathbf{Xb})'(\mathbf{Xb}) \\ &= 1/(n-1)\ \mathbf{b}'\mathbf{X}'\mathbf{Xb} \\ &= \mathbf{b}'\ \left[1/(n-1)\ \mathbf{X}'\mathbf{X}\right]\mathbf{b} \\ &= \mathbf{b}'\mathbf{S_{xx}}\mathbf{b} \end{aligned}$$

□

Variance of a Linear Combination

- This is a very useful result, as it allows to variance of a linear composite to be computed directly from the variance-covariance matrix of the original variables.
- This result may be extended immediately to obtain the variance-covariance *matrix* of a set of linear composites in a matrix $\mathbf{Y} = \mathbf{XB}$. The proof is not given as, it is a straightforward generalization of the previous proof.

Variance-Covariance Matrix of Several Linear Combinations

A beautiful thing about matrix algebra is the way formulas generalize to the multivariate case.

Theorem. Given $\mathbf{X}$, a set of n deviation scores on p variables in column variate form, having variance-covariance matrix $\mathbf{S_{xx}}$. The variance-covariance matrix of any set of linear combinations $\mathbf{Y} = \mathbf{XB}$ may be computed as

$$\mathbf{S_{YY}} = \mathbf{B}'\mathbf{S_{xx}}\mathbf{B} \tag{7}$$

Covariance Matrix of Two Sets of Linear Combinations

Theorem. Given $\mathbf{X}$ and $\mathbf{Y}$, two sets of n deviation scores on p and q variables in column variate form, having covariance matrix $\mathbf{S_{xy}}$. The covariance matrix of any two sets of linear combinations $\mathbf{W} = \mathbf{XB}$ and $\mathbf{M} = \mathbf{YC}$ may be computed as

$$\mathbf{S_{wm}} = \mathbf{B}'\mathbf{S_{xy}}\mathbf{C} \tag{8}$$