Multiple regression - Matrices

This handout will present various matrices which are substantively interesting and/or provide useful means of summarizing the data for analytical purposes. As we will see, means, standard deviations, and correlations are substantively interesting; other matrices are primarily useful for computational purposes.

Let $X_0 = 1$ for all cases. Let A and B be any two variables. We then define the following: (NOTE: To simplify the notation, when subscripting the X variables, we will refer to them only by number, e.g. s_{12} = sample covariance of X_1 and X_2.)

$$M_{AB} = \sum A_i B_i,$$

$$XP_{AB} = \sum (A_i - \bar{A})(B_i - \bar{B}) = M_{AB} - N\bar{A}\bar{B} = M_{AB} - M_{A0}M_{B0}/M_{00},$$

$$s_{AB} = \frac{XP_{AB}}{N-1} = \frac{MP_{AB} - N\bar{A}\bar{B}}{N-1} = \frac{\sum (A_i - \bar{A})(B_i - \bar{B})}{N-1},$$

$$r_{AB} = \frac{s_{AB}}{\sqrt{s_{AA}s_{BB}}} = \frac{s_{AB}}{s_A s_B} = \frac{XP_{AB}}{\sqrt{XP_{AA}XP_{BB}}} = \frac{1}{N-1}\sum \frac{(A_i - \bar{A})}{\sqrt{s_{AA}}}\ \frac{(B_i - \bar{B})}{\sqrt{s_{BB}}}$$

Note that each of these are symmetric, e.g. $M_{1Y} = M_{Y1}$, $s_{12} = s_{21}$. We will discuss each of these in turn:

1. **M_{AB}.** Let us compute the values of M for the variables X_0, X_1, X_2, Y.

M_{AB}	Formula: Σ (A * B)	Value
M_{00}	$\Sigma (X_0 * X_0) = \Sigma 1 = N$	20.0
$M_{01} = M_{10}$	$\Sigma (X_0 * X_1) = \Sigma X_1$	241.0
$M_{02} = M_{20}$	$\Sigma (X_0 * X_2) = \Sigma X_2$	253.0
$M_{0Y} = M_{Y0}$	$\Sigma (X_0 * Y) = \Sigma Y$	488.3
M_{11}	$\Sigma (X_1 * X_1) = \Sigma X_1^2$	3,285.0
$M_{12} = M_{21}$	$\Sigma (X_1 * X_2) = \Sigma X_1X_2$	2,999.0
$M_{1Y} = M_{Y1}$	$\Sigma (X_1 * Y) = \Sigma X_1Y$	6,588.3
M_{22}	$\Sigma (X_2 * X_2) = \Sigma X_2^2$	3,767.0
$M_{2Y} = M_{Y2}$	$\Sigma (X_2 * Y) = \Sigma X_2Y$	6448.9
M_{YY}	$\Sigma (Y * Y) = \Sigma Y^2$	13,742.3

Hence, to get the different possible values of M_{AB}, we multiply every variable (including X_0) by every variable. These numbers ought to look familiar. $M_{00} = N = 20$; The other numbers are the totals we got when we first presented the data. As we have seen, the different values of M_{AB} contain all the information we need for calculating regression models.

It is often convenient to present the values of M_{AB} in matrix form. We can write

$$\begin{array}{ccccc} & X_0 & X_1 & X_2 & Y \\ X_0 & 20.0 & & & \\ X_1 & 241.0 & 3{,}285.0 & & \\ X_2 & 253.0 & 2{,}999.0 & 3{,}767.0 & \\ Y & 488.3 & 6{,}588.3 & 6{,}448.9 & 13{,}742.27 \end{array}$$

or,

$$M = \begin{bmatrix} 20.0 & & & \\ 241.0 & 3{,}285.0 & & \\ 253.0 & 2{,}999.0 & 3{,}767.0 & \\ 488.3 & 6{,}588.3 & 6{,}448.9 & 13{,}742.27 \end{bmatrix}$$

Additional comments:

a. For symmetric matrices, it is common to not present either the elements above or the elements below the diagonal, since they are redundant.

b. The Matrix M/N (i.e. the M matrix with all elements divided by the sample size N) is sometimes called the augmented moment matrix. If you exclude the rows and columns for X_0 from the M matrix, then M/N is called the matrix of moments about the origin. The reason for this name may be clearer after we look at the covariance matrix.

2. **XP_{AB}**. XP_{AB} gives the cross-product deviations from the means (which we have also referred to as SST and SP). In our current example:

XP_{AB}	Formula: $\Sigma (A - \bar{A})(B - \bar{B})$	Value
XP_{11}	$\Sigma (X_1 - \bar{X}_1) * (X_1 - \bar{X}_1) = SST_1$	380.95
$XP_{12} = XP_{21}$	$\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2) = SP_{12}$	-49.65
$XP_{1Y} = XP_{Y1}$	$\Sigma (X_1 - \bar{x}_1)(Y - \bar{Y}) = SP_{Y1}$	704.29
XP_{22}	$\Sigma (X_2 - \bar{X}_2)(X_2 - \bar{X}_2) = SST_2$	566.55
$XP_{2Y} = XP_{Y2}$	$\Sigma (X_2 - \bar{X}_2)(Y - \bar{Y}) = SP_{Y2}$	271.91
XP_{YY}	$\Sigma (Y - \bar{Y})(Y - \bar{Y}) = SST_Y$	1820.43

(We don't bother with X_0, since any cross-products involving it would equal 0). It is often convenient to present the XP_{AB} values in Matrix form. We can write

	X_1	X_2	Y
X_1	380.95		
X_2	-49.65	566.55	
Y	704.29	271.91	1820.43

or,

$$XP = \begin{bmatrix} 380.95 & & \\ -49.65 & 566.55 & \\ 704.29 & 271.91 & 1820.43 \end{bmatrix}$$

Incidentally, Hayes labels this as the SSCP matrix, for sums of squares and cross-products.

3. **s_{AB}**. s_{AB} is the covariance of variables A and B. When A = B, s_{AA} = the sample variance of A. s_{AA} can of course also be written as s_A^2. $s_{AB} = XP_{AB}/(N - 1)$. In matrix form, we can write

	X_1	X_2	Y
X_1	20.05		
X_2	-2.61	29.82	
Y	37.07	14.31	95.81

or,

$$s = \begin{bmatrix} 20.05 & & \\ -2.61 & 29.82 & \\ 37.07 & 14.31 & 95.81 \end{bmatrix}$$

Additional comments:

a. The s matrix is typically called the covariance matrix or the variance/covariance matrix. It is also sometimes called the Matrix of Moments about the mean, because the mean is subtracted from each variable.

b. The sample covariance gives us an indication of the association between two variables. If the covariance is zero, then there is no association. If the covariance is positive, that means that above-average values on one variable tend to be paired with above-average values on the other variable. The square roots of the diagonal elements of the s matrix (i.e. the standard deviations) give an idea of how closely clustered cases are about the mean.

c. The values for the covariances are dependent on the metrics of the variables. For example, income here is measured in thousands of dollars; if instead, income were measured in dollars, s_{YY} would be 1 million times larger, and all the other elements of the s matrix would increase by a factor of 1,000.

d. Because the values of the s matrix are so dependent on the metrics used, it is difficult to tell via just "eyeballing" how strong the association between variables is. For example, does the fact that $s_{Y1} = 37.07$ mean that there is a very strong link between education and income, or is it only a weak association? This difficulty in interpretation is one of the reasons that correlations between variables (discussed next) are often looked at.

4. **r_{AB}.** r_{AB} is the <u>correlation</u> between variables A and B. As noted above, $r_{AB} = s_{AB}/s_A s_B$. There are several properties about correlations worth noting:

a. r can range from -1 to 1. The larger the absolute value of r is, the stronger the association is between the two variables. Hence, r provides a more intuitive means than s for looking at association.

b. the correlation of any variable with itself is 1, e.g. $r_{AA} = s_{AA}/s_A s_A = s_A^2/s_A^2 = 1$.

c. Another way of thinking of r is that it is the covariance of the <u>standardized</u> variables. Let $A' = (A - \overline{A})/s_A$, $B' = (B - \overline{B})/s_B$ (Note that both A' and B' appear in one of our formulas for r). That is, let A' and B' be the <u>z-score transformations</u> of A and B. As we showed before, any Z score has a mean of zero and a variance of 1. Ergo,

$$r_{A'B'} = \frac{s_{A'B'}}{\sqrt{s_{A'A'} s_{B'B'}}} = s_{A'B'} = r_{AB}$$

Let us now compute the correlations of the variables. Keep in mind that $r_{11} = r_{22} = r_{YY} = 1$.

r_{AB}	Formula: $s_{AB}/\sqrt{}\,(s_{AA}s_{BB})$	Value
$r_{12} = r_{21}$	$-2.61/\sqrt{(20.05 * 29.82)}$	-.1067
$r_{1Y} = r_{Y1}$	$37.07/\sqrt{(20.05 * 95.81)}$	.8458
$r_{2Y} = r_{Y2}$	$14.31/\sqrt{(29.82 * 95.81)}$	.2677

In matrix form, we can write this as

	X_1	X_2	Y
X_1	1.00		
X_2	-.11	1.00	
Y	.85	.27	1.00

or,

$$r = \begin{bmatrix} 1.00 & & \\ -.11 & 1.00 & \\ .85 & .27 & 1.00 \end{bmatrix}$$

From the correlation matrix, it is clear that education (X_1) is much more strongly correlated with income (Y) than is job experience (X_2).

5. **Alternative formulas**. It is very common for computer programs to report the correlations, standard deviations, and means for all the variables, along with the sample size. From this information, you can construct any of the matrices we have just talked about.

$$s_{AB} = r_{AB} s_A s_B,$$

$$XP_{AB} = s_{AB}(N - 1) = r_{AB} s_A s_B(N - 1),$$

$$M_{AB} = XP_{AB} + N\bar{A}\bar{B} = s_{AB}(N - 1) + N\bar{A}\bar{B} = r_{AB} s_A s_B(N - 1) + N\bar{A}\bar{B}$$

Example. To illustrate this, let us compute s_{11}, s_{Y1}, XP_{12}, M_{Y2} using only the correlations, means, standard deviations, and sample size.

$s_{11} = r_{11} * s_1 * s_1 = 1 * 4.478 * 4.478 = 20.05$,
$s_{Y1} = r_{Y1} * s_Y * s_1 = .8458 * 9.788 * 4.478 = 37.07$,
$XP_{12} = r_{12} * s_1 * s_2 * (N - 1) = -.1067 * 4.478 * 5.46 * 19 = -49.57$,
$M_{Y2} = r_{Y2} * s_y * s_2 * (N - 1) + N\,\bar{Y}\,\bar{X}_2$
$\quad = .2677 * 9.788 * 5.46 * 19 + 20 * 24.415 * 12.65 = 6448.82$

6. **Counting rules.** Let L = # of X variables + # of Y variables (Not counting X_0). In this case, L = 3. Note that

A. The M matrix has (L + 1)(L + 2)/2 = 10 unique elements (i.e. diagonal and sub-diagonal elements).

B. The XP matrix has (L)(L + 1)/2 = 6 unique elements

C. The s matrix has (L)(L + 1)/2 = 6 unique elements

D. The correlation matrix has (L)(L - 1) = 3 unique (i.e. sub-diagonal) elements

E. There are L means (i.e. 3)

F. There are L standard deviations (i.e. 3)

Note further that, if you know

A. The correlations (3 pieces of information), means (3), standard deviations (3), and sample size (1 piece of information), (10 pieces of information altogether), or

B. The covariance matrix (6 elements), the means (3), and sample size (1) (10 pieces of information), or

C. The XP Matrix (6 elements), means (3), and sample size (1) (10 pieces of information altogether), or

D. The M matrix (10 elements)

THEN you have all the information you need to work any regression problem.