Linear Transformations and Linear Composites

I. Linear Transformations of Variables.

Means and Standard Deviations of Linear Transformations

A linear transformation takes the form of creating a new variable from the old variable using the equation for a straight line:

$$\text{new variable} = a + b^* (\text{old variable})$$

where *a* and *b* are mathmatical constants. What is the mean and the variance of the new variable? To solve this let X denote the old variable and assume that it has a mean of $\overline{X}$ and a variance of S_X^2. Let X* denote the new variable. Then

$$X^* = a + bX$$

The mean of X* is

$$\begin{aligned}\overline{X}^* &= \frac{\sum X^*}{n} \\ &= \frac{\sum (a + bX)}{n} \\ &= \frac{\sum a + \sum bX}{n} \\ &= \frac{\sum a + b\sum X}{n} \\ &= \frac{na + b\sum X}{n} \\ &= \frac{na}{n} + b\frac{\sum X}{n} \\ &= a + b\overline{X}\end{aligned}$$

so

$$\overline{X}^* = a + b\overline{X}\ .$$

The variance of the new variable, *X**, can be found in a similar way. The variance in *X** is simply the variance of the quantity (*a* + *bX*). So, we merely substitute and use some algebra:

$$\begin{aligned}S_{X^*}^2 &= \frac{\sum \left(X^* - \overline{X}^*\right)^2}{n} \\ &= \frac{\sum \left(a + bX - a - b\overline{X}\right)^2}{n} \\ &= \frac{\sum \left(bX - b\overline{X}\right)^2}{n}\end{aligned}$$

$$= \frac{\sum\left[b\left(X - \bar{X}\right)\right]^2}{n}$$

$$= \frac{\sum b^2\left(X - \bar{X}\right)^2}{n}$$

$$= b^2 \frac{\sum\left(X - \bar{X}\right)^2}{n}$$

$$= b^2 S_X^2$$

so

$$S_{X*} = b^2 S_X^2 .$$

One of the most important applications of linear transforms comes in standardization. To remove the effects of, say, school grade in a data set on reading ability, one could first sort the data by grade and then standardize the variables so that each grade has the same mean and the same standard deviation. One of the most frequently used methods of standardization is to *Z*-transform a variable so that the mean and standard deviation of the new variable are, resepctively, 0 and 1. The familiar formula for this is

$$Z = \frac{(X - \bar{X})}{s_X}$$

With a little bit of algebra, we can rework this formula to

$$Z = -\frac{\bar{X}}{s_X} + \left(\frac{1}{s_X}\right) X$$

This is the same as the linear equation

new variable = a + b * old variable.

The new variable is Z and the old variable is X. The value of a is

$$-\frac{\bar{X}}{s_X}$$

and the value of b is

$$\frac{1}{s_X} .$$

Hence the mean of Z must be

$$\bar{Z} = a + b\bar{X} = -\frac{\bar{X}}{s_X} + \left(\frac{1}{s_X}\right)\bar{X} = 0$$

and the variance of *Z* will be

$$s_Z^2 = b^2 s_X^2 = \left(\frac{1}{s_X}\right)^2 s_X^2 = 1.0 \ .$$

Transforming data to have a desired mean and/or standard deviation

The formulas given above may be used to demonstrate how to transform variables to have a desired mean and standard deviation. For example, suppose that we had raw scores on a newly developed MMPI scale and would like to express these scores in the customary metric of the MMPI, T scores with a mean of 50 and a standard deviation of 10.

Let X denote the variable in raw units with observed mean $\bar{X}$ and observed standard deviation s_X. Let $\bar{X}_d$ denote the desired mean and s_d denote the desired standard deviation. Taking the square root of equation given above for the variance of a transformed variable gives

$$\sqrt{s_d^2} = \sqrt{b^2 s_X^2}$$

so

$$b = \frac{s_d}{s_X} \ .$$

Thus the slope is simply the desired standard deviation divided by the observed standard deviation.

The intercept may be found by substituting this expression into the equation for the mean of a transformed variable:

$$\bar{X}_d = a + b\bar{X} = a + \left(\frac{s_d}{s_X}\right)\bar{X}$$

so

$$a = \bar{X}_d - \left(\frac{s_d}{s_X}\right)\bar{X}$$

Putting these expressions for *a* and *b* together (plus doing a little algebra) gives the formula for the desired transformation:

$$X_d = \bar{X}_d + s_d\left(\frac{X - \bar{X}}{s_X}\right)$$

In plain English, to transform a variable to have a desired mean and a desired standard deviation, simply take the *Z*-transform of the original variable, multiply it by the desired standard deviation, and then add the desired mean. In the case of the MMPI, where we wanted scores with a mean of 50 and a standard deviation of 10, we would simply find the Z transform of the original score, multiply that by 10, and then add 50.

Covariance and Correlation of Two Linearly Transformed Variables

What is the covariance between two variables that have been linearly transformed? Here, let the old variables be X and Y and the new variables be denoted as, respectively, X* and Y*. Then the transformation will take the form

$$X^* = a + bX$$

and

$$Y^* = c + dY. .$$

The covariance is defined as

$$\text{cov}(X^*, Y^*) = \frac{\sum\left(X^* - \overline{X}^*\right)\left(Y^* - \overline{Y}^*\right)}{n}.$$

One again, substitute and do some algebra:

$$\begin{aligned}
&= \frac{\sum\left(a + bX - a - b\overline{X}\right)\left(c + dY - c - d\overline{Y}\right)}{n} \\
&= \frac{\sum\left(bX - b\overline{X}\right)\left(dY - d\overline{Y}\right)}{n} \\
&= \frac{\sum\left[b\left(X - \overline{X}\right)\right]\left[d\left(Y - \overline{Y}\right)\right]}{n} \\
&= \frac{\sum bd\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{n} \\
&= bd\frac{\sum\left(X - \overline{X}\right)\left(Y - \overline{Y}\right)}{n} \\
&= bd\ \text{cov}(X, Y)
\end{aligned}$$

so

$$\text{cov}(X^*, Y^*) = bd\ \text{cov}(X, Y).$$

Thus, a linear transformation will change the covariance only when both of the old variances are multiplied by something other than 1. If we simply add something to both old variables (i.e., let *a* and *c* be something other than 0, but make $b = d = 1$), then the covariance will not change.

Although a linear transformation *may* change the means and variances of variables and the covariances between variables, it will *never* change the correlation between variables. Consider *X** and *Y** as given above. We have already shown that the variances of these two variables are

$$S^2_{X^*} = b^2 S^2_X$$

and

$$S^2_{Y^*} = d^2 S^2_Y.$$

We have also demonstrated that the covariance between the two transformed variables is

$$\mathrm{cov}(X^*, Y^*) = bd\ \mathrm{cov}(X, Y).$$

The correlation between the transformed variables will be

$$\mathrm{corr}(X^*, Y^*) = \frac{\mathrm{cov}(X^*, Y^*)}{S_{X^*} S_{Y^*}}$$

Again, we substitute and perform some algebra:

$$\begin{aligned}\mathrm{corr}(X^*, Y^*) &= \frac{bd\ \mathrm{cov}(X, Y)}{\sqrt{b^2 S_X^2}\ \sqrt{d^2 S_Y^2}} \\ &= \frac{bd\,\mathrm{cov}(X, Y)}{\sqrt{b^2 d^2}\ S_X S_Y} \\ &= \frac{bd}{bd}\ \frac{\mathrm{cov}(X, Y)}{S_X S_Y} \\ &= \frac{\mathrm{cov}(X, Y)}{S_X S_Y}\end{aligned}$$

which is the correlation between X and Y.

II. Linear Composites.

Mean of a Linear Composite

Here, we wish to examine what happens when an entirely new variable is constructed as a linear function of several old variables. Let X_i denote the ith old variable and Y the new variable. We can make the case somewhat more general by assuming that we add a residual, U, that is actually a random number taken from a standard normal distribution with mean of 0 and standard deviation of 1. The equation for the new variable is

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots b_p X_p + uU.$$

(If this business of the random variable, U, is bothersome, then simply let the quantity u equal 0 in the equation and in all that follows. Nothing of substance will change)

We can now go through the same algebra that we used above in the transformation of an old variable into a new variable to calculate the variance of variable Y. The only trick here is to recall that variable U will have a mean of 0, and because it is random, will be uncorrelated with all the Xs.

Consider the mean. The mean of Y is the mean of

$$a + b_1 X_1 + b_2 X_2 + \ldots b_p X_p + uU.$$

$$= \frac{\sum_{i=1}^{N}(a + b_1X_{1i} + b_2X_{2i} + \ldots b_pX_{pi} + uU)}{N}$$

$$= \frac{Na + b_1\sum_{i=1}^{N}X_{1i} + b_2\sum_{i=1}^{N}X_{2i} + \ldots b_p\sum_{i=1}^{N}X_{pi} + u\sum_{i=1}^{N}U_i}{N}$$

$$\frac{Na}{N} + b_1\frac{\sum_{i=1}^{N}X_{1i}}{N} + b_2\frac{\sum_{i=1}^{N}X_{2i}}{N} + \ldots b_p\frac{\sum_{i=1}^{N}X_{pi}}{N} + u\frac{\sum_{i=1}^{N}U_i}{N} =$$

$$= a + b_1\bar{X}_1 + b_2\bar{X}_2 + \ldots b_p\bar{X}_p + u\bar{U}\ .$$

$$= a + b_1\bar{X}_1 + b_2\bar{X}_2 + \ldots b_p\bar{X}_p + 0.$$

so

$$\bar{Y} = a + b_1\bar{X}_1 + b_2\bar{X}_2 + \ldots\ b_p\bar{X}_p\ .$$

Variance of a Linear Composite

Similar logic will write the variance of *Y* as a function of the variables on the right side of the equation. We will not go through the elaborate algebra, but instead give the result:

$$S_Y^2 = \sum_{i=1}^{P}\sum_{j=1}^{P} b_i\, b_j\, \mathrm{cov}\left(X_i, X_j\right) + u^2$$

Note that the term u^2 is NOT included in this summation.

For example, suppose that the new variable is a linear composite of three variables, or

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + uU\ .$$

Then

$$S_Y^2 = b_1^2S_{X_1}^2 + b_2^2S_{X_2}^2 + b_3^2S_{X_3}^2 + 2b_1b_2\,\mathrm{cov}\left(X_1, X_2\right) + 2b_1b_3\,\mathrm{cov}\left(X_1, X_3\right) + 2b_2b_3\,\mathrm{cov}\left(X_2, X_3\right) + u^2$$

(Recall, here, that $\mathrm{cov}\left(X_i, X_i\right) = S_{X_i}^2$).

Covariance of Two Linear Composites

With similar algebra, it can be shown that the covariances between any two linear composites can be written in terms of the *b*s, and the covariances among the *X*s and the *U*s. Let

$$Y_1 = a_1 + b_{11}X_1 + b_{12}X_2 + \ldots b_{1p}X_p + u_1U_1$$

and

$$Y_2 = a_2 + b_{21}X_1 + b_{22}X_2 + \ldots b_{2p}X_p + u_2U_2\ .$$

Then

$$\mathrm{cov}(Y_1, Y_2) = \sum_{i=1}^{P}\sum_{j=1}^{P} b_{1i} b_{2j} \mathrm{cov}(X_i, X_j) + u_1 u_2 \mathrm{cov}(U_1, U_2)$$

If all the variables are standardized, the *b*s become β coefficients, all of the covariances become correlations, and all variances become 1.0. Then

$$S_Y^2 = 1 = \sum_{i=1}^{P}\sum_{j=1}^{P} \beta_i \beta_j \mathrm{corr}(X_i, X_j) + u^2$$

and

$$\mathrm{cov}(Y_1, Y_2) = \mathrm{corr}(Y_1, Y_2) = \sum_{i=1}^{P}\sum_{j=1}^{P} \beta_{1i} \beta_{2j} \mathrm{corr}(X_i, X_j) + u_1 u_2 \mathrm{corr}(U_1, U_2).$$

III. Transformations and Linear Composites in Matrix Algebra

Transformations of variables can be economically written using matrix algebra. Let X denote the old variable and Y denote the new variable. We have seen that the transformation for the ith individual takes the form

$$Y_i = a + bX_i$$

Now let **x** denote a column vector of old variable values and **y** a column vector of new variable values. Equation ( ) above may now be written as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ Y_N \end{pmatrix} = \begin{pmatrix} a \\ a \\ \cdot \\ a \end{pmatrix} + \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ X_N \end{pmatrix} b$$

or

$$\mathbf{y} = \mathbf{a} + \mathbf{x}b$$

If we wish to make a new variable as a linear composite of several old variables, then let **X** denote a matrix of the old variable values. The rows of **X** correspond to the observations and the columns to the variables. Let **b** denote a column vector of weights. The equation becomes

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ Y_N \end{pmatrix} = \begin{pmatrix} a \\ a \\ \cdot \\ a \end{pmatrix} + \begin{pmatrix} X_{11} & X_{12} & . & X_{1p} \\ X_{21} & X_{22} & . & X_{2p} \\ . & . & . & . \\ X_{N1} & X_{N2} & . & X_{Np} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ . \\ b_p \end{pmatrix}$$

or

$$\mathbf{y} = \mathbf{a} + \mathbf{X}\mathbf{b}.$$

A more general formulation permits a linear transformation of one set of variables into another, new set of variables. That is, instead of a column vector of *Y*s, there is now a matrix of *Y*s. Let X_{ij} denote the score of the ith person on the jth old variable and let Y_{ij} denote the score of the ith person on the jth new variable. Let a_j denote the constant for the jth variable, and let b_{ij} denote the weight used to multiply the ith *X* variable for the jth *Y* variable. The transformation is

$$\begin{pmatrix} Y_{11} & Y_{12} & . & Y_{1q} \\ Y_{21} & Y_{22} & . & Y_{2q} \\ . & . & . & . \\ . & . & . & . \\ Y_{N1} & Y_{N2} & . & Y_{Nq} \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & . & a_q \\ a_1 & a_2 & . & a_q \\ . & . & . & . \\ . & . & . & . \\ a_1 & a_2 & . & a_q \end{pmatrix} + \begin{pmatrix} X_{11} & X_{12} & . & X_{1p} \\ X_{21} & X_{22} & . & X_{2p} \\ . & . & . & . \\ . & . & . & . \\ X_{N1} & X_{N2} & . & X_{Np} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & . & b_{1q} \\ b_{21} & b_{22} & . & b_{2q} \\ . & . & . & . \\ . & . & . & . \\ b_{p1} & b_{p2} & . & b_{pq} \end{pmatrix}$$

or

$$\mathbf{Y} = \mathbf{A} + \mathbf{XB} .$$

The general case for the mean and the variance-covariance matrix of the transformed variables can now be written. Let

$$\begin{pmatrix} \bar{Y}_1 \\ \bar{Y}_2 \\ . \\ . \\ \bar{Y}_q \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ . \\ . \\ a_q \end{pmatrix} + \begin{pmatrix} b_{11} & b_{21} & . & b_{p1} \\ b_{12} & b_{22} & . & b_{p2} \\ . & . & . & . \\ . & . & . & . \\ b_{1q} & b_{2q} & . & b_{pq} \end{pmatrix} \begin{pmatrix} \bar{X}_1 \\ \bar{X}_2 \\ . \\ . \\ \bar{X}_p \end{pmatrix}$$

or

$$\bar{\mathbf{y}} = \mathbf{a} + \mathbf{B}^t\bar{\mathbf{x}} .$$

Likewise, the covariance matrix may be written in a general form. Let $\mathbf{C_{ij}}$ denote a covariance matrix between the **i** variables (the rows of the matrix) and the **j** variables (the columns). The variance-covariance matrix for the new variables is

$$\mathbf{C_{yy}} = \mathbf{B}^t\mathbf{C_{xx}}\mathbf{B}$$

and the covariance matrix between the transformed variances and the original variables is

$$\mathbf{C_{yx}} = \mathbf{B}^t\mathbf{C_{xx}} .$$

Once again, the transformation of several *X* variables into a single *Y* variable is a special instance of this equation where **B** becomes a column vector. If one transforms a single *X* into a single *Y*, then **B** is a column vector and matrix $\mathbf{C_{xx}}$ becomes a scalar equal to the variance of X.