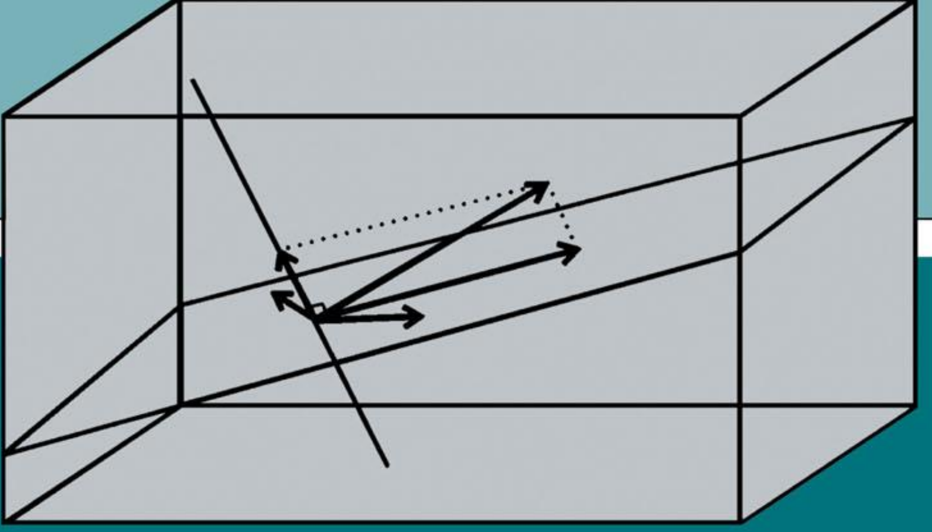
The Geometry of Multivariate Statistics



Thomas D. Wickens

The Geometry of Multivariate Statistics

Thomas D. Wickens
University of California, Los Angeles

Psychology Press
Taylor & Francis Group
NEW YORK AND LONDON

First published 1995 by Lawrence Erlbaum Associates, Inc.

Published 2014 by Psychology Press
711 Third Avenue, New York, NY 10017

and by Psychology Press
27 Church Road, Hove, East Sussex, BN3 2FA

Psychology Press is an imprint of the Taylor & Francis Group, an informa business

Library of Congress Cataloging-in-Publication Data
Wickens, Thomas D., 1942-
The Geometry of Multivariate Statistics / Thomas D. Wickens.
p. cm.
Includes index.
1. Multivariate analysis. 2. Vector Analysis. I. Title
QA278.W53 1994
519.5'35–dc20 94-4654
CIP

ISBN 13: 978-0-805-81656-3 (hbk)

Publisher's Note
The publisher has gone to great lengths to ensure the quality of this reprint but points out that some imperfections in the original may be apparent.

Contents

Preface

In simple terms, this little book is designed to help its reader think about multivariate statistics. I say "think" here because I have not written about how one programs the computer or calculates the test statistics. Instead I hope to help the reader understand in a broad and intuitive sense what the multivariate procedures do and how their results are interpreted.

There are many ways to develop multivariate statistical theory. The traditional approach is algebraic. Sets of observations are represented by matrices, linear combinations are formed from these matrices by multiplying them by coefficient matrices, and useful statistics are found by imposing various criteria of optimization on these combinations. Matrix algebra is the vehicle for these calculations. A second approach is computational. Many users of multivariate statistics find that they do not need to know the mathematical basis of the techniques as long as they can transform data into results. The computation can be done by a package of computer programs that somebody else has written. An approach to multivariate statistics from this perspective emphasizes how the computer packages are used, and is usually coupled with rules that allow one to extract the most important numbers from the output and interpret them.

Useful as both approaches are, particularly when combined, they overlook an important aspect of multivariate analysis. To apply it correctly, one needs a way to conceptualize the multivariate relationships among the variables. To some extent, the equations help. A linear combination explicitly defines a new variable, and a correlation matrix accurately expresses the pattern of association among the members of a set of variables. However, I have never found these descriptions sufficient, either for myself or when teaching others. Problems that involve many variables require a deeper understanding than is typically provided by the formal equations or the computer programs. Although knowing the algebra is helpful and a powerful computer program is almost essential, neither is sufficient without a good way to picture the variables.

Fortunately, a tool to develop this understanding is available. Multivariate statistical theory is fundamentally an application of the theory of linear algebra, and linear algebra has a strong geometric flavor. This spatial interpretation carries over to multivariate statistics and gives a concrete and pictorial form to multivariate relationships. The geometry lets one describe, more or less easily, the complex pattern of relationships among a set of variables. It gives a metaphor for the way that variables are combined. With a bit of practice, one develops an intuitive feel for how the multivariate methods work. However, rather unfortunately, I believe, this approach is ignored in the conventional treatment of multivariate statistics. To be sure, geometric references appear as asides in many texts and the metaphor motivates the terminology in several places—the use of the word "orthogonal" to mean "uncorrelated" is an example. However, except in a few domains, such as factor analysis (and even there not consistently), the understanding that a geometric representation gives is not exploited.

This book presents most important procedures of multivariate statistics geometrically. I have tried to develop the theory entirely this way. Even when computational equations are presented, they derive from the geometry instead of the algebra. I hope that this emphasis will give the reader a coherent picture into which all the multivariate techniques fit. In the interests of presenting a unified approach and to keep this book short, I have not covered either the algebraic basis of the methods or the computer tools that are available to carry it out. This omission does not indicate that I think that either algebra or computation is unimportant. I have concentrated on one aspect of multivariate statistics and have left the more mechanical parts to other sources. I expect that the book will often be used in tandem with either an algebraic or a computational study of the techniques, whichever is more compelling to a reader's needs and tastes. In this spirit, the book is an adjunct to, but not a substitute for, a more conventional treatment.

One feature of this book may seem curious. I do not refer to other books and readings. Two classes of references might have been expected. The first group contains references to other geometric treatments of multivariate statistics. This work is widely spread throughout many sources, but I have found none that directly follows on from this book. Geometric ideas pervade most treatments of linear algebra and some treatments of multivariate statistics, but are often given implicitly. The second missing class of citations is to multivariate statistics in nongeometric form. There are many such books, written at many different levels. References to all of them would be excessive, and a reference to one or two would be both arbitrary and restrictive. Moreover, I have noticed that, particularly with technical matter, the book that is the most understandable is the book one has used before. To send readers to a new source is often more confusing than helpful. Since my goal here is to present a way of thinking, and since I expect it to be used in combination with other approaches, I have not pinned things down tightly. A reader wishing to pursue these ideas can do so, with some thought, in any of the dozens of multivariate texts or hundreds of texts on linear algebra. The best start is to consult whichever of these books is familiar.

Finally, I want to acknowledge the many friends, colleagues, and students (in all permutations and combinations) who have read drafts of sections of this book. I have drawn freely on their comments and suggestions, although, perhaps foolishly, have not followed them all. I am no less grateful for their efforts if my memory is too poor and this preface is too short to list them all individually.

This page intentionally left blank

Chapter 1

Variable space and subject space

Multivariate statistics concerns the analysis of data in which several variables are measured on each of a series of individuals or subjects. The goal of the analysis is to examine the interrelationships among the variables: how they vary together or separately and what structure underlies them. These relationships are typically quite complex, and their study is made easier if one has a way to represent them graphically or pictorially. There are two complementary graphical representations, each of which contributes different insights. This chapter describes these two ways to view a set of multivariate data.

Any description of multivariate data starts with a representation of the observations and the variables. Consider an example. Suppose that one has ten observations of two variables, X and Y, as shown in Figure 1.1. For the ith subject, denote the scores by X_i and Y_i. Summary statistics for these data give the two means as

$$\overline{X} = 4.00 \qquad \text{and} \qquad \overline{Y} = 12.00,$$

their standard deviations as

$$s_{\mathsf{X}} = 2.06 \qquad \text{and} \qquad s_{\mathsf{Y}} = 4.55,$$

and their correlation as 0.904.

The first way to picture these data is as a *scatterplot*. One variable, here X, is assigned to the horizontal axis and the other variable, here Y, is assigned to the vertical axis. Each subject's scores are plotted as a point; thus, the first subject is plotted at the point (1, 4), the second subject at

i	X_i	Y_i
1	1	4
2	1	7
3	3	9
4	3	12
5	4	11
6	4	12
7	5	17
8	6	13
9	6	18
10	7	17

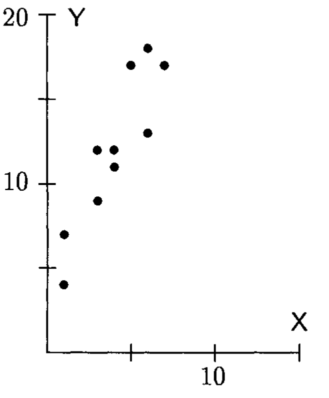


Figure 1.1: *Ten bivariate observations and their scatterplot.*

the point $(1, 7)$, and so on. This scatterplot is shown on the right of Figure 1.1. In it the axes are the variables, and each point expresses the data from a single subject. Several things are immediately clear from a glance at the scatterplot. First, the points do not cluster about the origin, so the means are nonzero. Second, the points are more spread out along the Y axis than along the X axis, so variable Y has greater variability than variable X. Third, high scores on one variable correspond closely with high scores on the other variable, so there is a substantial association between the variables. Finally, the connection between the variables does not bend and can be approximately represented by a straight line.

If one is only interested in how the individual values of X and Y go together, then the location of the origin of the scatterplot is unimportant. The X-Y relationship is the same, no matter where the axes are put. In most of multivariate statistics, the analysis of association is simplified by shifting the center of the plot to the origin. This shift is accomplished by subtracting the mean of each variable from every score, thereby creating new variables, here represented by lowercase letters,

$$x_i = X_i - \overline{X} \qquad \text{and} \qquad y_i = Y_i - \overline{Y}. \tag{1.1}$$

This operation is known as *centering* the variables. Subtracting the means $\overline{X} = 4$ and $\overline{Y} = 12$ from the scores in Figure 1.1 gives the new scores and scatterplot in Figure 1.2. The means of the centered scores are now zero, but the standard deviations and correlation are unchanged. Except for the position of the axes, the scatterplot is identical to the raw score plot. Centered variables are easier to work with than the raw scores, yet

i	x_i	y_i
1	−3	−8
2	−3	−5
3	−1	−3
4	−1	0
5	0	−1
6	0	0
7	1	5
8	2	1
9	2	6
10	3	5

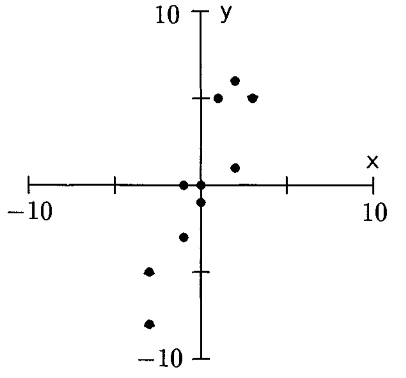
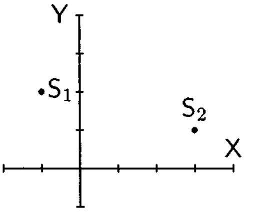


Figure 1.2: *Centered data from Figure 1.1.*

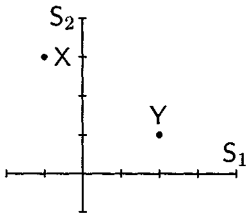
convey almost as much information. With a few exceptions, all the variables considered in these notes are centered.

The scatterplot is a very useful way to look at a set of data. It clearly shows the pattern of individual observations. Almost every multivariate analysis involves (or should involve) several of these plots. However, in one respect the very characteristic of the scatterplot that makes it useful also limits it. The scatterplot places emphasis on the observations, not on the variables as general entities. When one wants to talk about the variables, a different type of graph often gives a clearer picture.

The second way to plot a set of multivariate data exchanges the roles of subjects and variables from that in the scatterplot. Consider two bivariate observations: subject S_1 receives the score −1 on variable X and 2 on variable Y, and subject S_2 receives the scores 3 and 1. The scatterplot has an axis for each variable and a point for each subject:



In the new graph, there is an axis for each subject. Each variable is represented by a point, variable X by the point (−1, 3) and variable Y by the point (2, 1):



The two plots picture the data in different spaces. In the scatterplot, the axes are defined by the variables, so the plot is said to be located in *variable space*. The entities plotted are the observations, each of which is denoted by a point. In the new plot, the axes are defined by the observations or subjects, so it is said to be located in *subject space*. The entities plotted here are the variables themselves.

The trick now is to extend the subject-space plot to more than two observations. For the data in Figure 1.1, there is one axis for each subject, making ten axes in all. There are two points in this ten-dimensional space, one for X and one for Y. For the uncentered data in Figure 1.1 these points are

$$(1,1,3,3,4,4,5,6,6,7) \quad \text{and} \quad (4,7,9,12,11,12,17,13,18,17),$$

and for the centered data in Figure 1.2 they are

$$(-3,-3,-1,-1,0,0,1,2,2,3) \quad \text{and} \quad (-8,-5,-3,0,-1,0,5,1,6,5).$$

Here a problem arises. The data demand a ten-dimensional space, but ten-dimensional graph paper is hard to come by. On the face of it, subject space seems an impossible place to plot data or even to think about. However, the problem of visualization is vastly simplified because the plot contains only three objects. There is one landmark at the origin, one point for variable X, and another point for variable Y. To see the relationships among the variables, one need look only at the relative positions of these three points. Usually one can concentrate on the plane in subject space that they determine and can ignore the rest of the space. By dropping the original axes, the number of dimensions needed to draw the picture in subject space is no greater than the number of variables.

When drawing a picture of subject space, it is convenient to represent the variables not by points, as on a scatterplot, but by arrows, known as *vectors*, drawn from the origin (denoted by a boldface $\mathbf{0}$) to the points. There are only a few points in the space and the vectors help them stand out. To suggest their geometric nature, the vectors are indicated in this book by boldface letters with an arrow above them. Variable X becomes the vector $\vec{\mathbf{x}}$ and variable Y becomes the vector $\vec{\mathbf{y}}$. Figure 1.3 shows these vectors for the centered data in Figure 1.2.

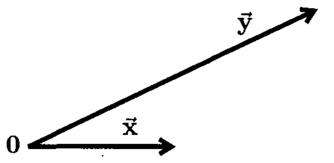


Figure 1.3: *The centered data of Figure 1.2 plotted in subject space.*

Two important properties of the variables are apparent in this plot (they were actually used to draw it). First, the lengths of the vectors indicate the variability of the corresponding variables. Variables X and Y have standard deviations of 2.06 and 4.55, and so $\vec{y}$ is $4.55/2.06 = 2.21$ times as long as $\vec{x}$. Second, the angle between the vectors measures how similar the variables are to each other. Vectors that represent highly correlated variables have a small angle between them, and vectors that represent uncorrelated variables are perpendicular. The correlation of two variables is equal to the cosine of the angle between their vectors, a relationship that is discussed further in Section 2.2. The correlation between X and Y is 0.904, which implies that the angle between $\vec{x}$ and $\vec{y}$ is 26°.

The differences between the scatterplot in Figure 1.2 and the vector plot in Figure 1.3 are substantial. Each type of plot has its uses. The scatterplot contains points describing the individual observations. These plots are the best places to look for effects that involve the individual subjects. For example, the straight-line character of the relationship between X and Y is readily apparent in the scatterplot. Such plots are essential tools for spotting outliers and specification errors in the statistical models. However, the characteristics of the variables are only implicit in the pattern of points in a scatterplot. When the data are in general conformity with the type of linear relationship studied in much of multivariate statistics, the detail in the individual points obscures the relationships among the variables. In the subject-space vector plot, this detail is eliminated and the variables are represented directly. By dropping the information about the individual observations, one obtains a far clearer picture of how the variables are related. This type of plot is the one to use when one wants to think about relationships among the variables.

Before turning to the geometry of the multivariate techniques, a review of the various entities introduced above helps to distinguish among them and to establish a consistent notation. Fundamentally, one wants to study the *variables*. One's substantive theories refer to them, either as abstract entities or with respect to a particular set of data. One speaks in general of a subject's age or a particular treatment as influencing the score on a particular test or the performance on a particular task. In this book, variables are represented by uppercase letters in a sans serif typeface. Examples are the variables X and Y above. Frequently several variables are differentiated by subscripts, numerical or otherwise, for example as X_1, X_2, and X_3, or as X_a and X_b. The dummy indices j and k are used to refer to these subscripts, for example, by writing X_j or Y_k. When discussing statistical testing, it is important to distinguish between abstract variables in the population, to which the theories apply, and their realization in a particular sample of data, but no typographical distinction is made between these concepts here.

In a set of data, the variables are represented by a set of *scores* or *observations* that is to be analyzed. The scores are specific numbers, such as those in Figure 1.1. In their original uncentered values, the scores are represented by uppercase subscripted letters in an *italic* typeface, such as Y_1, Y_2, etc. The dummy subscript i is used to index the observations. When referring to a variable that is already subscripted, the subject subscript comes first, so variable X_j has the values X_{1j}, X_{2j}, ..., or in general X_{ij}. The centered values obtained by subtracting the mean of the variable are represented by lowercase letters with the same system of subscripts, such as y_i or x_{ij}.

The observations in a set of multivariate data vary on two dimensions. With p variables recorded from n subjects, the centered data form the array

$$\begin{matrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np}. \end{matrix}$$

This array can be thought of as a series of points in space either by slicing the table horizontally into the scores for each subject,

$$(x_{i1},\ x_{i2},\ \ldots,\ x_{ip}),$$

or by slicing it vertically into the observations for each variable,

$$(x_{1j},\ x_{1j},\ \ldots,\ x_{nj}).$$

The first set of points is plotted as a scatterplot in variable space, and the second set of points is plotted as vectors in subject space.

The observations of each variable, just described abstractly as a point in multidimensional space, are represented in either of two ways, both called vectors. The *algebraic vectors* are the columns of the data matrix and are conventionally denoted by a **boldface** lowercase letter, such as

$$\mathbf{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \qquad \text{or} \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

For typographical convenience, an algebraic vector is often written horizontally, that is, as its *transpose*. The transposition operation is denoted by a prime. By this convention, the vector $\mathbf{y}$ above is written either as

$$\mathbf{y} = [y_1,\, y_2,\, \ldots,\, y_n]' \qquad \text{or as} \qquad \mathbf{y}' = [y_1,\, y_2,\, \ldots,\, y_n].$$

The second type of vector is the *geometric vector*, in which the observations are treated as defining a directed line segment reaching from the origin to a point in subject space. The geometric vectors are denoted here by the same letter as their algebraic equivalent, with their geometric character emphasized by putting an arrow above them. For centered data, a lowercase boldface letter is used, for example, $\vec{\mathbf{y}}$ or $\vec{\mathbf{x}}_j$. When vectors describing uncentered data are needed, they are indicated by uppercase boldface letters such as $\vec{\mathbf{X}}_j$ or $\vec{\mathbf{Y}}$.

Having laid out these distinctions, it helps to blur them somewhat. There is a close association between a variable X, the algebraic vector $\mathbf{x}$ containing its scores, and the geometric vector $\vec{\mathbf{x}}$ that illustrates these scores. Each refers to the same entity in a slightly different way. Statements about the variable X refer broadly to the quantity being studied, and statements about the vector $\vec{\mathbf{x}}$ refer specifically to the geometric object that represents that variable. This correspondence is taken for granted below. It is common in much writing to impute the geometric properties of the vectors to the variables themselves. For example, one says that uncorrelated variables are orthogonal, which really means that the vectors that correspond to them are perpendicular. The book makes little use of the algebraic vectors, so the unqualified word "vector" refers to its geometric counterpart. However, when giving a particular instance of a geometric vector, its coordinates are written in algebraic form, for example, by writing $\vec{\mathbf{x}} = [1,\, 2,\, -3]'$.

The geometric picture is of minimal value with only two variables. Most useful applications involve at least three dimensions, and many require more than three. In this book vector configurations are described in words and represented by diagrams that attempt to portray three-dimensional pictures—for example, look at almost any figure in the latter two-thirds of the book. It is essential that the structure of these pictures be understood.

The reader is very strongly urged to construct these pictures in three dimensions and to make sure that the angular relationships among the vectors are clear. No equipment of great complexity is needed: a handful of pencils, pens, or sticks to stand for the vectors is sufficient, perhaps augmented by a sheet of paper or cardboard to show a plane.

A geometric treatment of multivariate statistics is founded on an understanding of how vectors are manipulated and combined. Most of the following chapter covers this material. How the vectors and the spaces that contain them are used to represent variables is described in Section 2.2, but without any discussion of their application to multivariate statistics. Readers who are familiar with vector operations can skim through this material fairly rapidly. The remaining chapters discuss the most important multivariate techniques. Much of the discussion focuses on multiple regression, a technique that is fairly simple, but in which almost all the interpretation problems of multivariate analysis are found.

Exercises

1. Two individuals take a two part test, one receiving scores of 5 and 3 on the two parts, the other, scores of 7 and 2. Represent these results as points in variable space and as vectors in subject space.

2. Take a pair of sticks or pointers, and position them as the vectors $\vec{\mathbf{x}} = [1,\ 0,\ 1]'$ and $\vec{\mathbf{y}} = [0,\ 1,\ 1]'$ in three-dimensional space. Put a sheet of paper or cardboard along the pointers, and describe their relationship in this two-dimensional plane.

Chapter 2

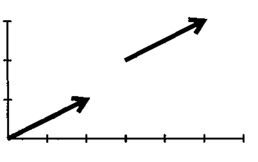
Some vector geometry

A geometric vector gives a concrete representation to a variable and to the algebraic vector of data that measures it. Before applying this geometric representation to the techniques of multivariate statistics, one needs to understand how to manipulate these vectors and combine them. The first section of this chapter describes these operations, and the second section considers the correspondence between vectors and variables. The final two sections describe the important concepts of vector spaces, linear dependence, and projection.

2.1 Elementary operations on vectors

Operations such as addition and multiplication that apply to algebraic vectors have their counterparts for geometric vectors.

Vectors. A vector is a directed line segment. It has two properties, its direction and its length. It can be started from any point. Both the vector that goes from the point (0, 0) to the point (2, 1) and the vector that goes from (3, 2) to (5, 3) move two units over and unit one up, and so are different instances of the same vector:



The standard position from which to start a vector is at the origin **0**. However, when combining several vectors or illustrating the relationships among them, it can be helpful to start them at different points. As long as neither its direction nor length is altered, one can freely slide a vector about a diagram without changing it.

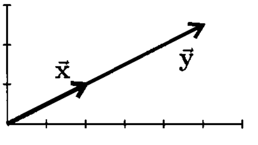
The length of a vector. A fundamental property of any vector is its length, which is denoted by placing vertical bars about it—the length of the vector $\vec{\mathbf{x}}$ is $|\vec{\mathbf{x}}|$. Algebraically, the length of a vector is found by using the Pythagorean theorem. The vector $\vec{\mathbf{x}} = [x_1,\ x_2]'$ has length $\sqrt{x_1^2 + x_2^2}$. More generally, the length of the n-dimensional vector $\vec{\mathbf{x}} = [x_1,\ x_2,\ \ldots,\ x_n]'$ is

$$|\vec{\mathbf{x}}| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}. \tag{2.1}$$

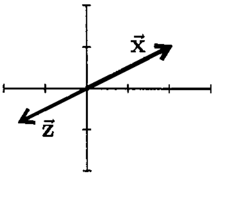
For example, the length of the vector $\vec{\mathbf{x}} = [2,1]'$ in the diagram above is

$$|\vec{\mathbf{x}}| = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.236.$$

Scalar multiplication. The simplest operation that changes one vector into another is multiplication by a number such as 2 or -7. In vector terminology, an ordinary number is called a *scalar*, so this type of multiplication is called *scalar multiplication*. In scalar multiplication of an algebraic vector, every component is multiplied by the same constant. For a geometric vector, the resultant vector is proportionally longer or shorter, but points in the same direction. If vector $\vec{\mathbf{y}}$ is 2.5 times the vector $\vec{\mathbf{x}}$, then it points in the same direction but is 2.5 times as long:



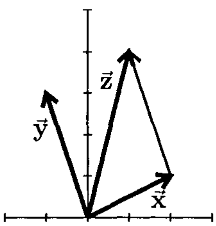
Multiplication by a negative number reverses the direction in which a vector points, but leaves it lying along the same line; the vector $\vec{\mathbf{z}} = -0.7\vec{\mathbf{x}}$ is shorter than $\vec{\mathbf{x}}$ and points in the opposite direction:



Addition. The simplest operation that combines a pair of vectors is addition. The sum of two vectors is obtained by adding their corresponding elements together. The sum of the algebraic vectors $\mathbf{x} = [2,\ 1]'$ and $\mathbf{y} = [-1,\ 3]'$ is the vector

$$\mathbf{z} = \mathbf{x} + \mathbf{y} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ 3 \end{bmatrix} = \begin{bmatrix} 2-1 \\ 1+3 \end{bmatrix} = \begin{bmatrix} 1 \\ 4 \end{bmatrix}.$$

Geometrically, the sum of two vectors is produced by moving one of the vectors to the end of the other and drawing the sum as a vector from the start of the first vector to the end of the second. To get vector $\vec{\mathbf{z}}$ above, $\vec{\mathbf{y}}$ is copied to the tip of $\vec{\mathbf{x}}$:

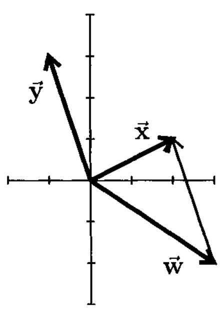


Which of the two vectors one puts first and which second makes no difference to the sum. Sums of more than two vectors are found in the same way by placing the tail of one at the tip of another to form a chain.

Subtraction. Vectors are subtracted by multiplying the vector to be subtracted by -1 to reverse its direction, then adding it to the first vector,

$$\vec{\mathbf{w}} = \vec{\mathbf{x}} - \vec{\mathbf{y}} = \vec{\mathbf{x}} + [(-1)\vec{\mathbf{y}}].$$

Geometrically, subtraction corresponds to going to the end of the first vector and moving in a direction opposite to that of the second vector:



Algebraically, one subtracts the vectors component by component to obtain the same result.

Linear combinations. A *linear combination* of two or more vectors involves both addition and scalar multiplication. The linear combination of the vectors $\vec{x}$ and $\vec{y}$ with scalar coefficients b_X and b_Y is the weighted sum

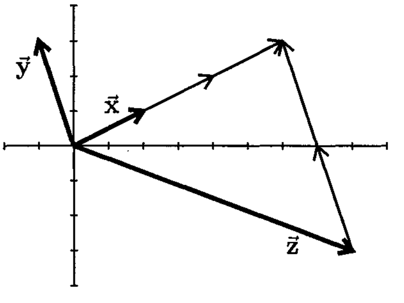
$$\vec{z} = b_X\vec{x} + b_Y\vec{y}.$$

Geometrically, this operation corresponds to going along $\vec{x}$ for a distance equal to b_X times its length, then turning in the direction of $\vec{y}$ and proceeding for b_Y times its length. A linear combination written with algebraic vectors has the same form. Both addition and subtraction are special cases of linear combination.

Using the vectors $\vec{x}$ and $\vec{y}$ from the example above, the linear combination $\vec{z} = 3\vec{x} - 2\vec{y}$ is the vector

$$\begin{aligned} \vec{z} &= 3\begin{bmatrix} 2 \\ 1 \end{bmatrix} - 2\begin{bmatrix} -1 \\ 3 \end{bmatrix} \\ &= \begin{bmatrix} 6 \\ 3 \end{bmatrix} + \begin{bmatrix} 2 \\ -6 \end{bmatrix} = \begin{bmatrix} 8 \\ -3 \end{bmatrix}. \end{aligned}$$

Thus, $\vec{z}$ is constructed by concatenating three copies of $\vec{x}$ and subtracting two copies of $\vec{y}$, giving a vector that goes from the origin to (8, −3):



More generally, a linear combination of the vectors $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p$ with coefficients b_j is the vector

$$\vec{y} = b_1\vec{x}_1 + b_2\vec{x}_2 + \cdots + b_p\vec{x}_p. \tag{2.2}$$

It is constructed by moving in the direction of $\vec{x}_1$ for b_1 times its length, then in the direction of $\vec{x}_2$ for b_2 times its length, and so forth. Linear combinations of this type are central to multivariate statistics.

The dot product and the angle between two vectors. The angle between two vectors is fundamental to multivariate geometry. Like the length of a vector, the angle $\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ between $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ is geometrically obvious. Simply shift the two vectors so they start at the same point and measure the angle between them. Algebraically this angle is calculated from the sum of the products of the coordinates (in matrix notation as the product $\mathbf{x}'\mathbf{y}$). This sum of products is important enough to be a concept in its own right. It is known as the *scalar product* or *dot product*, and usually denoted by a centered dot,

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n. \tag{2.3}$$

Geometrically, the dot product $\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}$ is equal to the product of three terms: the length of $\vec{\mathbf{x}}$, the length of $\vec{\mathbf{y}}$, and the cosine of the angle between them:

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = |\vec{\mathbf{x}}||\vec{\mathbf{y}}| \cos \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}). \tag{2.4}$$

Rearranging the terms of this equation lets one compute the cosine:

$$\cos \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\vec{\mathbf{x}} \cdot \vec{\mathbf{y}}}{|\vec{\mathbf{x}}||\vec{\mathbf{y}}|}. \tag{2.5}$$

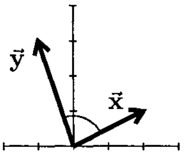
For the vectors $\vec{\mathbf{x}} = [2, 1]'$ and $\vec{\mathbf{y}} = [-1, 3]'$ used as examples above, the dot product is

$$\vec{\mathbf{x}} \cdot \vec{\mathbf{y}} = (2)(-1) + (1)(3) = 1,$$

and from Equation 2.5 the cosine of the angle between them is

$$\cos \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{1}{\sqrt{5}\sqrt{10}} = 0.141.$$

The cosine is translated to an angle by using either a table of cosines (such as Table 2.1, at the end of this chapter) or the inverse cosine key on a scientific calculator. Here $\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = 81.9°$. As a picture shows, the vectors lie at nearly a right angle:



Two angular relationships between vectors are particularly important. If $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ are scalar multiples of each other, so that $\vec{\mathbf{y}} = b\vec{\mathbf{x}}$, then they lie on the same line and are said to be *collinear*. The angle between them is either 0° or 180° and the cosine of the angle is either 1 or -1, respectively. In contrast, if $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ lie at right angles to each other, then they are said to be *orthogonal*, a state symbolized by the expression $\vec{\mathbf{x}} \perp \vec{\mathbf{y}}$. The angle between them is 90°. Equations 2.3–2.4 imply that when $\vec{\mathbf{x}} \perp \vec{\mathbf{y}}$, the dot product $\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} = \cos 90° = 0$. Calculating the dot product is the appropriate way to check whether two vectors are orthogonal.

A vector is collinear with itself and $\angle(\vec{\mathbf{x}},\vec{\mathbf{x}}) = 0°$. As $\cos 0° = 1$, Equation 2.4 says that the scalar product of a vector with itself is the square of its length:

$$\vec{\mathbf{x}}\cdot\vec{\mathbf{x}} = |\vec{\mathbf{x}}|^2 \qquad \text{or} \qquad |\vec{\mathbf{x}}| = \sqrt{\vec{\mathbf{x}}\cdot\vec{\mathbf{x}}}. \tag{2.6}$$

Of course, this result is equivalent to the component-wise formula of Equation 2.1.

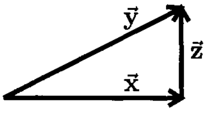
One frequently needs to find the dot product of a vector with a linear combination or the dot product of two linear combinations. These products can be simplified to contain only dot products of the individual vectors by distributing the product over the combination. For any three vectors $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$, and $\vec{\mathbf{z}}$ and scalars a and b, two relationships known as *distributive laws* hold:

$$\begin{aligned} \vec{\mathbf{x}}\cdot(a\vec{\mathbf{y}} + b\vec{\mathbf{z}}) &= a(\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}) + b(\vec{\mathbf{x}}\cdot\vec{\mathbf{z}}), \\ (a\vec{\mathbf{x}} + b\vec{\mathbf{y}})\cdot\vec{\mathbf{z}} &= a(\vec{\mathbf{x}}\cdot\vec{\mathbf{z}}) + b(\vec{\mathbf{y}}\cdot\vec{\mathbf{z}}). \end{aligned} \tag{2.7}$$

These rules generalize to combinations that involve more than two vectors and to dot products of one linear combination with another. For example,

$$(a\vec{\mathbf{u}} + b\vec{\mathbf{v}})\cdot(c\vec{\mathbf{x}} + d\vec{\mathbf{y}}) = ac(\vec{\mathbf{u}}\cdot\vec{\mathbf{x}}) + ad(\vec{\mathbf{u}}\cdot\vec{\mathbf{y}}) + bc(\vec{\mathbf{v}}\cdot\vec{\mathbf{x}}) + bd(\vec{\mathbf{v}}\cdot\vec{\mathbf{y}}).$$

Relationships in a right triangle. A *right triangle* is one that contains a 90° angle. The vectors corresponding to the sides surrounding this angle are orthogonal, making right triangles particularly important in the representation of statistical relationships. Suppose that $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$, and $\vec{\mathbf{z}}$ have a right-triangular relationship with $\vec{\mathbf{x}} \perp \vec{\mathbf{z}}$:



The Pythagorean Theorem relates the length of these vectors to each other:

$$|\vec{\mathbf{x}}|^2 + |\vec{\mathbf{z}}|^2 = |\vec{\mathbf{y}}|^2. \tag{2.8}$$

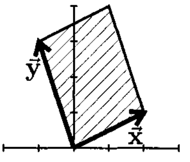
Thus, in a right triangle, when the lengths of any two vectors are known, the length of the third vector is easily found.

The angle between the non-orthogonal vectors in a right triangle is related to the ratios of the lengths of its sides by the elementary trigonometric functions—the sine, the cosine, the tangent, and their arcfunctions. Of these trigonometric relationships, the one most important for statistical vectors is the ratio of the lengths of the sides that meet at an angle; in the above triangle,

$$\cos \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{|\vec{\mathbf{x}}|}{|\vec{\mathbf{y}}|}. \tag{2.9}$$

The length of the hypotenuse, or longest side, appears in the denominator, so this ratio never exceeds unity.

Areas and volumes. Any set of vectors determines an area or a volume. In two dimensions, linking a pair of vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ with copies of themselves forms a parallelogram:



The area of this parallelogram depends on the vectors' lengths and on the angle between them. From elementary geometry, the area of the parallelogram is

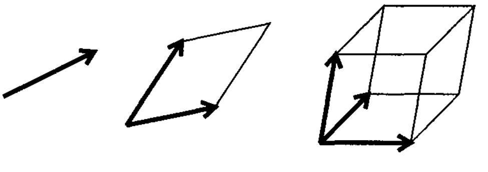
$$\text{Area} = |\vec{\mathbf{x}}||\vec{\mathbf{y}}| \sin \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = |\vec{\mathbf{x}}||\vec{\mathbf{y}}|\sqrt{1 - \cos^2 \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})}. \tag{2.10}$$

The area is large when the vectors are long or when they are nearly orthogonal. It is small when they are short or nearly parallel. For the two vectors just illustrated,

$$\text{Area} = \sqrt{5}\sqrt{10} \sin 81.9^\circ = 7.00.$$

A pair of collinear vectors make only a one-dimensional structure, and so the area they define is equal to zero.

This idea generalizes to any number of vectors. Think of sets vectors in one, two, and three dimensions:



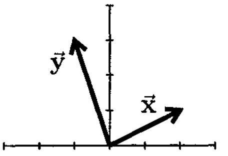
The single vector has only its length, the two vectors have the area just described, and the three vectors set out a volume. These three concepts are analogous in their respective spaces, and it is helpful to have a general notation that encompasses them and allows for more than three vectors. For this *generalized volume*, write $\text{vol}(\vec{\mathbf{x}})$ in one dimension, $\text{vol}(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ in two dimensions, $\text{vol}(\vec{\mathbf{x}}, \vec{\mathbf{y}}, \vec{\mathbf{z}})$ in three dimensions, and so forth.

Calculating the generalized volume is easy for one or two vectors and rather tedious for more. In one dimension $\mathrm{vol}(\vec{x}) = |\vec{x}|$, and in two dimensions Equation 2.10 applies. Beyond that, although the concept is well defined geometrically, the calculation is tedious in detail—it depends on the angles among all the vectors. Mathematically, the square of the volume equals a matrix quantity known as the determinant of the matrix of dot products of one vector with another. For example, with three vectors, the volume is

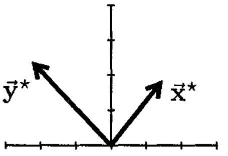
$$\mathrm{vol}(\vec{x}, \vec{y}, \vec{z}) = \sqrt{\begin{vmatrix} \vec{x}\cdot\vec{x} & \vec{x}\cdot\vec{y} & \vec{x}\cdot\vec{z} \\ \vec{x}\cdot\vec{y} & \vec{y}\cdot\vec{y} & \vec{y}\cdot\vec{z} \\ \vec{x}\cdot\vec{z} & \vec{y}\cdot\vec{z} & \vec{z}\cdot\vec{z} \end{vmatrix}}. \tag{2.11}$$

Most books on linear algebra contain instructions for calculating the determinant, and many computer packages either find it automatically or include it as an option.

Rotation of vectors. Another operation of importance to the geometry of multivariate statistics is the rotation of a set of vectors. The geometry of rotation is quite straightforward. A pair of vectors $\vec{x}$ and $\vec{y}$ is rotated through some angle θ by rigidly turning them through this angle without altering the angle between them. For example, a 25° counterclockwise rotation changes the vectors

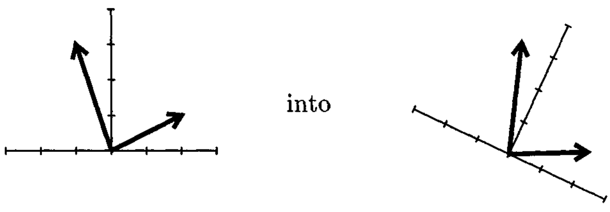


into

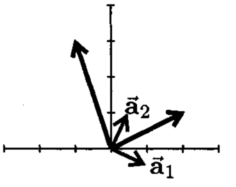


More generally, rotating a set of vectors changes their relationship to whatever coordinate system one is using, but does not alter either their lengths or the angular relationships among them. The effect is as if one was looking at the original vectors from a new viewpoint.

Algebraically, the components of a rotated set of vectors are linear combinations of the old components, the coefficients being functions of the angle of rotation. It is easiest to understand the transformation by thinking of the rotation in a different way. Rather than rotating the vectors within a single coordinate system, think of rotating the coordinate system through the same angle in the opposite direction. The 25° counterclockwise rotation of the vectors illustrated above becomes a clockwise rotation of the axes by 25°, changing the configuration



The new coordinate system is specified by writing unit vectors $\vec{\mathbf{a}}_1$, $\vec{\mathbf{a}}_2$, ..., that point along its new axes. For the rotation above, these vectors are



The vectors defining the new coordinate system have unit length and are mutually orthogonal, so for all $j \neq k$,

$$|\vec{\mathbf{a}}_j| = |\vec{\mathbf{a}}_k| = 1 \qquad \text{and} \qquad \vec{\mathbf{a}}_j \cdot \vec{\mathbf{a}}_k = 0. \tag{2.12}$$

To find a coordinate of a vector on the new axes, one takes its dot product with the appropriate axis-defining vector. When vector $\vec{\mathbf{x}}$ is rotated to the new vector $\vec{\mathbf{x}}^\star$, its coordinates are

$$\vec{\mathbf{x}}^\star = [\vec{\mathbf{a}}_1 \cdot \vec{\mathbf{x}},\ \vec{\mathbf{a}}_2 \cdot \vec{\mathbf{x}},\ \ldots,\ \vec{\mathbf{a}}_p \cdot \vec{\mathbf{x}}]'. \tag{2.13}$$

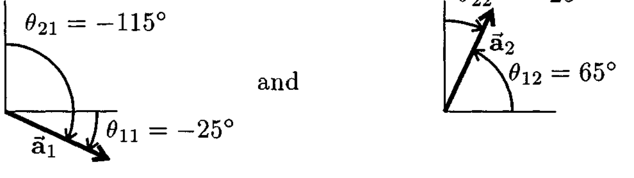
This equation can be justified from the notion of projection discussed in Section 2.5 (specifically, Equation 2.20).

To use Equation 2.13 to calculate $\vec{\mathbf{x}}^\star$, one must specify the $\vec{\mathbf{a}}_k$. The vector $\vec{\mathbf{a}}_k$ is located with respect to the original axes by the angles θ_{jk} between the jth original axis and the vector. Its coordinates are the cosines of these angles, known as *direction cosines*:

$$\vec{\mathbf{a}}_k = [\cos\theta_{1k},\ \cos\theta_{2k},\ \ldots,\ \cos\theta_{pk}]'. \tag{2.14}$$

These angles completely specify the new coordinate system and, via Equation 2.13, the algebraic form of the new vectors.

As an example, consider the 25° rotation of the vectors $\vec{\mathbf{x}} = [2,1]'$ and $\vec{\mathbf{y}} = [-1,3]$ shown above. The angles between the old and the new coordinates specify the rotated reference frame:



Taking the cosine of these angles gives the vectors

$$\begin{aligned}\vec{\mathbf{a}}_1 &= [\cos\theta_{11},\ \cos\theta_{21}]' = [\cos -25°,\ \cos -115°]' = [0.906,\ -0.423]',\\ \vec{\mathbf{a}}_2 &= [\cos\theta_{12},\ \cos\theta_{22}]' = [\cos 25°,\ \cos -65°]' = [0.423,\ 0.906]'.\end{aligned}$$

The coordinates of $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ in the rotated system are given by the dot products

$$\begin{aligned}\vec{\mathbf{x}}^\star &= [\vec{\mathbf{a}}_1\cdot\vec{\mathbf{x}},\ \vec{\mathbf{a}}_2\cdot\vec{\mathbf{x}}]' = [1.390,\ 1.752]',\\ \vec{\mathbf{y}}^\star &= [\vec{\mathbf{a}}_1\cdot\vec{\mathbf{y}},\ \vec{\mathbf{a}}_2\cdot\vec{\mathbf{y}}]' = [-2.174,\ 2.296]'.\end{aligned}$$

It is not necessary for the vectors $\vec{\mathbf{a}}_1$, $\vec{\mathbf{a}}_2$, $\ldots$, $\vec{\mathbf{a}}_p$ either to have unit length or to be mutually orthogonal for these calculations to work. The same principles can express vectors with respect to axes that have different scales or that do not meet perpendicularly. In an *oblique rotation*, the new coordinates are not orthogonal when expressed in the old system, and although the first part of Equations 2.12 holds, the second does not. Oblique coordinate systems are important in some areas of multivariate statistics, particularly factor analysis.

2.2 Variables and vectors

When vectors are used to represent variables in subject space, the statistical properties of the variables correspond to the geometry of the vectors in a simple way. The length of the vector $\vec{\mathbf{x}}$ representing the variable X is equal to the square root of the sum of its squared components (Equation 2.1). Since each x_i is the deviation of a score from the mean, the squared length is equal to the sum of the squared deviations, a quantity known as the *sum of squares*:

$$SS_{\mathsf{X}} = \sum_i x_i^2 = |\vec{\mathbf{x}}|^2. \tag{2.15}$$

Similarly, the dot product of two different vectors is the *sum of cross products*:

$$CP_{\mathsf{XY}} = \sum_i x_i y_i = \vec{\mathbf{x}}\cdot\vec{\mathbf{y}}. \tag{2.16}$$

Both these quantities play important roles in univariate and multivariate hypothesis testing.

When using vectors to visualize the relationship among several variables, a different interpretation of the length is more helpful. The standard deviation of a variable X (as the unbiased estimate of a population value) is

$$s_{\mathsf{X}} = \sqrt{\frac{\sum x_i^2}{N-1}} = \frac{\sqrt{\sum x_i^2}}{\sqrt{N-1}}.$$

Replacing the numerator by the length of $\vec{\mathbf{x}}$ shows immediately that the length of a vector is proportional to the standard deviation of the corresponding variable,

$$|\vec{\mathbf{x}}| = \sqrt{N-1}\, s_{\mathsf{X}}. \tag{2.17}$$

In most analyses, the constant of proportionality $\sqrt{N-1}$ is unimportant, since every vector is based on the same number of observations. One can treat the length of a vector as equal to the standard deviation of its variable. Vectors corresponding to variables with high variability are long, while those corresponding to variables that are nearly constant are short.

In many parts of multivariate analysis, one works with *standardized variables*, often known as z scores, that are both centered about their means and divided by their standard deviations. All vectors corresponding to standardized scores have the same length. Only their direction is important. Usually one ignores the factor of $\sqrt{N-1}$ and gives the standardized vectors unit length.

The angle $\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ between two vectors indicates the degree to which the corresponding variables vary together. Either this angle or a function of it measures their similarity. The cosine is the most useful function here. When the vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ point in the same direction, $\cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ is 1. As the vectors are turned away from each other, this value drops, falling to 0 when they are orthogonal and to -1 when they point in opposite directions. Thus, $\cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ measures the *correlation* between the variables X and Y. Indeed, $\cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})$ is equal to the conventional *Pearson correlation coefficient* r_{XY} between X and Y. Assembling the definitions of angle, dot product, and length from Equations 2.5, 2.3, and 2.6,

$$r_{\mathsf{XY}} = \cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{|\vec{\mathbf{x}}||\vec{\mathbf{y}}|} = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}}. \tag{2.18}$$

This is the familiar formula for a Pearson correlation coefficient. The match between angle and correlation is a nice result. Angles are easily comprehensible and correlations are only a little less so. The connection between them gives the geometric approach much of its power.

As an example, return to the ten centered observations in Figure 1.2. For these data, direct application of the scalar product formula (Equation 2.3) gives

$$|\vec{\mathbf{x}}|^2 = \vec{\mathbf{x}}\cdot\vec{\mathbf{x}} = (-3)^2 + (-3)^2 + (-1)^2 + \cdots + 2^2 + 3^2 = 38,$$

$$|\vec{\mathbf{y}}|^2 = \vec{\mathbf{y}}\cdot\vec{\mathbf{y}} = (-8)^2 + (-5)^2 + (-3)^2 + \cdots + 6^2 + 5^2 = 186,$$

and

$$\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} = (-3)(-8) + (-3)(-5) + (-1)(-3) + \cdots + (2)(6) + (3)(5) = 76.$$

Substituting into Equation 2.18 gives the correlation:

$$r_{\mathsf{XY}} = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{\sqrt{|\vec{\mathbf{x}}|^2|\vec{\mathbf{y}}|^2}} = \frac{76}{\sqrt{(38)(186)}} = 0.904.$$

This correlation is the cosine of the angle between the vectors, so using a calculator or Table 2.1,

$$\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \arccos(0.904) = 25.3^\circ.$$

Collinear vectors indicate variables that are perfectly correlated, either positively or negatively ($\cos 0^\circ = 1$ or $\cos 180^\circ = -1$, respectively), and uncorrelated variables are represented by orthogonal vectors ($\cos 90^\circ = 0$). The geometric terminology is commonly carried over from the vectors to the variables themselves, so that words like collinear or orthogonal are applied directly to variables. For example, unrelated (or uncorrelated) variables are commonly said to be orthogonal.

The correspondence between standard deviation and length and between correlation and angle allows subject-space diagrams to be constructed without actually plotting points in high-dimensional spaces. One simply draws vectors with their lengths proportional to the standard deviations and sets them at angles determined by their correlations. When more than two variables are involved, these diagrams are more comprehensible than are the variable-space scatterplots. With three variables, it is hard to keep the relationships implicit in a three-dimensional scatterplot in mind, let alone to plot them, but it is not hard to imagine three vectors in ordinary space and to illustrate them with pointers of any convenient sort. One can sketch the projection of three vectors on a two-dimensional sheet of paper and get an adequate representation of their relationship, a thing that is nearly impossible with a three-dimensional scatterplot.

Another advantage of the geometric approach should be noted. The angles between several variables accurately describe the patterns of correlations that exist among the corresponding variables. The constraints implied by the geometry are exactly those that apply to correlations. Not every pattern of correlations among three or more variables is possible. For example, when variables X and Y are correlated with $r_{\mathsf{XY}} = 0.8$, a third variable Z cannot simultaneously be positively correlated to X with $r_{\mathsf{XZ}} = 0.5$ and negatively correlated to Y with $r_{\mathsf{YZ}} = -0.5$. The impossibility of this configuration is not obvious from the numbers themselves, but is readily shown by the geometry. The correlation of 0.8 implies that the vectors $\vec{x}$ and $\vec{y}$ have a 37° angle between them. A little experimentation with sticks or pencils will show that one cannot position a third vector $\vec{z}$ such that it simultaneously makes a 60° angle with $\vec{x}$ and a 120° angle with $\vec{y}$ (the correlations of ± 0.5). Vectors $\vec{x}$ and $\vec{y}$ are simply too close together to be that different in their relationship to a third vector. For a set of correlations to be possible, the angular relationships must be realizable in physical space.

Some other representations that are occasionally used to illustrate correlation lack the exact correspondence that exists between correlation and angle. In particular, the popular representation of correlation by Venn diagrams breaks down here. In this picture, variables are denoted by circles, and the square of their correlation is indicated by the amount that the circles overlap. Variables correlated with $r = 0.5$ would overlap by 25% of their area. However, correlations obey the axioms of angular measure, not those of area measure. Reasonable-looking Venn diagrams can be drawn for sets of variables that cannot exist. For example, a Venn diagram does not help one to see that the configuration of variables mentioned in the last paragraph is impossible—it is easy to draw two circles with a 64% overlap ($r = 0.8$) and a third circle that overlaps 25% with each of them ($r = \pm 0.5$). In other cases, (for example, with the suppressor relationships to be described in Section 5.4) Venn diagrams cannot be drawn for configurations of variables that are perfectly possible.

2.3 Vector spaces

Multivariate statistical theory makes heavy use of linear combinations to create new variables from old ones. These combinations have the form

$$\mathsf{Y} = b_1\mathsf{X}_1 + b_2\mathsf{X}_2 + \cdots + b_p\mathsf{X}_p.$$

When the coefficients b_j in this combination are allowed to vary freely, the number of new variables that can be created is infinite. To represent geometrically the range of possibilities implied by these combinations, it is helpful to introduce the idea of a vector space.

Suppose that one forms linear combinations from a set of vectors $\vec{x}_1$, $\vec{x}_2$, ..., $\vec{x}_p$. The set of vectors that can be created by these combinations is said to be *spanned* or *generated* by the original vectors. A member of this set is said to be in the *span* of the $\vec{x}_j$. For example, starting with any pair of vectors $\vec{x}_1$ and $\vec{x}_2$ that are not collinear, any other vector $\vec{y}$ in the same plane can be produced by a linear combination $\vec{y} = b_1\vec{x}_1 + b_2\vec{x}_2$ with appropriately selected coefficients b_1 and b_2. In particular, consider the vectors $\vec{\mathbf{x}}_1 = [1, -1]'$ and $\vec{\mathbf{x}}_2 = [2, 1]'$. These vectors are not collinear, and any point in the two-dimensional plane can be reached by combining them. As a little algebra shows, the general vector $\vec{\mathbf{y}} = [u, v]'$ equals the linear combination

$$\vec{\mathbf{y}} = \tfrac{1}{3}(u - 2v)\vec{\mathbf{x}}_1 + \tfrac{1}{3}(u + v)\vec{\mathbf{x}}_2.$$

For example, $[2, 2]' = -\frac{2}{3}\vec{\mathbf{x}}_1 + \frac{4}{3}\vec{\mathbf{x}}_2$ and $[-7, 4]' = -5\vec{\mathbf{x}}_1 - \vec{\mathbf{x}}_2$.

The collection of vectors spanned by a set of vectors has several properties that makes it useful to treat it as an entity in its own right. Such sets are known as *vector spaces.* In this book, vector spaces are denoted by the letter $\mathcal{V}$, with subscripts added to distinguish particular spaces. The relationships among vector spaces are denoted by the symbols of ordinary set notation. So $\vec{\mathbf{x}} \in \mathcal{V}$ means that the vector $\vec{\mathbf{x}}$ is a member of the vector space $\mathcal{V}$. For two vector spaces $\mathcal{V}_1$ and $\mathcal{V}_2$, the relationship $\mathcal{V}_1 \subset \mathcal{V}_2$ means that $\mathcal{V}_2$ contains every vector in $\mathcal{V}_1$ along with other vectors outside of $\mathcal{V}_1$. To allow for the possibility that $\mathcal{V}_1$ and $\mathcal{V}_2$ might be identical, one writes $\mathcal{V}_1 \subseteq \mathcal{V}_2$.

An important property of a vector space is that it is *closed* under linear combination. A linear combination of vectors chosen from a vector space cannot produce a vector that lies outside that space. More formally, for any vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_p$ from a vector space $\mathcal{V}$ and any scalar coefficients $b_1, b_2, \ldots, b_p$, the linear combination

$$\vec{\mathbf{y}} = b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2 + \cdots + b_p\vec{\mathbf{x}}_p$$

is always in $\mathcal{V}$. For the purposes of multivariate statistics, one can think of the vector spaces as consisting of lines, planes, solid three-dimensional spaces, and complete higher-dimensional units. Each of these spaces extends infinitely in all directions.

Many different sets of vectors can be used to span the same space. For example, ordinary two-dimensional space is spanned by any of the four sets of vectors

$$\vec{\mathbf{v}}_1 = [1,\ -1]', \qquad \vec{\mathbf{v}}_2 = [2,\ 1]',$$
$$\vec{\mathbf{a}}_1 = [1,\ -1]', \qquad \vec{\mathbf{a}}_2 = [1,\ 1]',$$
$$\vec{\mathbf{b}}_1 = [1,\ 0]', \qquad \vec{\mathbf{b}}_2 = [0,\ 1]',$$
$$\vec{\mathbf{v}}_1 = [1,\ -1]', \qquad \vec{\mathbf{v}}_2 = [2,\ 1]', \qquad \vec{\mathbf{v}}_3 = [3,\ 2]'.$$

Infinitely many other sets could also be used. Not all generating sets are equally useful, however. The final set is redundant, in that all three of its vectors are not needed. Any two of its members generate the same space as do the three. In contrast, the first three sets are minimal. Remove one vector from them and a smaller space is spanned. A minimal set of this type is known as a *basis* for the space. Any vector in the space can be written as a linear combination of the basis vectors.

Every basis for a vector space contains the same number of vectors. This number is a fundamental characteristic of the vector space, and is known as the *dimension* of the vector space. The dimension of the vector space $\mathcal{V}$ is denoted $\dim(\mathcal{V})$.

Among the most convenient bases for a vector space are those in which the vectors lie at right angles to each other. Such a basis is known as an *orthogonal basis.* The sets $\{\vec{\mathbf{a}}_1, \vec{\mathbf{a}}_2\}$ and $\{\vec{\mathbf{b}}_1, \vec{\mathbf{b}}_2\}$ above are both orthogonal bases, while the set $\{\vec{\mathbf{v}}_1, \vec{\mathbf{v}}_2\}$ is not. If all the vectors in an orthogonal basis have unit length, then it is called an *orthonormal basis.* The set $\{\vec{\mathbf{b}}_1, \vec{\mathbf{b}}_2\}$ is orthonormal. Although this set, with its vectors lying along the coordinate axes, is the most obvious orthonormal basis, there are many others. For any space of more than one dimension, there are an infinity of orthonormal bases, formed by rotations and reflections of the coordinates axes.

If one takes set of vectors from a vector space and uses them as generators, one creates a subset of the original vector space. This set is known as a *subspace.* Because the subspace is the span of a set of vectors, it is a vector space itself. Any linear combination of vectors in the subspace remains in the subspace. A set of vectors that is not closed is not a subspace. Strictly speaking a subspace may be identical to the original space, but the interesting subspaces are smaller—they are said to be *proper subspaces.* The subspaces mentioned in this book are proper. The dimension of a proper subspace is always less than that of its parent space.

Consider the ordinary three-dimensional vector space $\mathcal{V}_3$—the space in which one normally moves around. It can be generated by the standard orthonormal basis

$$[1,\ 0,\ 0]', \qquad [0,\ 1,\ 0]', \qquad \text{and} \qquad [0,\ 0,\ 1]'.$$

Within this space, the vector $\vec{\mathbf{x}}_1 = [1,\ 1,\ 1]'$ points diagonally away from the origin. The vector subspace $\mathcal{V}_1$ generated by this vector is a diagonal line through the origin that makes the same angle with the three axes and extends infinitely in both directions. This one-dimensional subspace contains every vector with three identical coordinates, $\vec{\mathbf{v}} = [b,\ b,\ b]'$. By introducing any second vector not collinear with $\vec{\mathbf{x}}_1$, one creates a two-dimensional subspace. For example, a planar subspace $\mathcal{V}_2$ within $\mathcal{V}_3$ is generated by the vectors

$$\vec{\mathbf{x}}_1 = [1,\ 1,\ 1]' \qquad \text{and} \qquad \vec{\mathbf{x}}_2 = [3,\ -2,\ 1]'.$$

The three vector spaces $\mathcal{V}_1$, $\mathcal{V}_2$, and $\mathcal{V}_3$ have an inclusive relationship in which each smaller space is contained within a larger space:

$$\mathcal{V}_1 \subset \mathcal{V}_2 \subset \mathcal{V}_3.$$

2.4 Linear dependence and independence

The fact that a set of p vectors may span a space that has fewer than p dimensions leads to an extension of the idea of collinearity. Collinear vectors lie along a single line, and thus the space that they span is no larger than the space spanned by one of them alone. Likewise the space spanned by a set of *multicollinear* vectors is less than it would be if no redundant vectors were present. The three vectors $\vec{\mathbf{v}}_1$, $\vec{\mathbf{v}}_2$, and $\vec{\mathbf{v}}_3$ above form a multicollinear set. Multicollinearity produces serious interpretation difficulties in multivariate statistics and is discussed further in Chapter 4.

It is not always obvious when a set of vectors is multicollinear. The problem is illustrated by returning to the subspace $\mathcal{V}_2$ generated by the vectors $\vec{\mathbf{x}}_1 = [1,\ 1,\ 1]'$ and $\vec{\mathbf{x}}_2 = [3,\ -2,\ 1]'$ above. Now suppose that the vector $\vec{\mathbf{x}}_3 = [-3,\ 7,\ 1]'$ is added to these two. Although it is not immediately apparent, $\vec{\mathbf{x}}_3$ is a member of the space $\mathcal{V}_2$ spanned by $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$. It adds nothing new, and the three vectors are multicollinear. Some algebra shows that $\vec{\mathbf{x}}_3 = 3\vec{\mathbf{x}}_1 - 2\vec{\mathbf{x}}_2$, so that any vector that can be reached by a linear combination of all three vectors can also be reached by $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ alone. For example, the vector $\vec{\mathbf{y}} = \vec{\mathbf{x}}_1 - 3\vec{\mathbf{x}}_2 + 2\vec{\mathbf{x}}_3$ can also be written by replacing $\vec{\mathbf{x}}_3$ to give

$$\vec{\mathbf{y}} = \vec{\mathbf{x}}_1 - 3\vec{\mathbf{x}}_2 + 2(3\vec{\mathbf{x}}_1 - 2\vec{\mathbf{x}}_2) = 7(\vec{\mathbf{x}}_1 - \vec{\mathbf{x}}_2).$$

Geometrically, the multicollinearity is illustrated by noting that all three vectors lie in the plane $\mathcal{V}_2$.

A formal criterion for multicollinearity or the lack of it uses the notion of linear combination. For a set of p vectors to span a p-dimensional space, no vector can lie in the space spanned by the other $p-1$ vectors. Consider the space spanned by the vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, $\ldots$, $\vec{\mathbf{x}}_p$, and suppose that it is multicollinear and that $\vec{\mathbf{x}}_p$ falls in the space spanned by the first $p-1$ of them. Because every vector in the space spanned by $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, $\ldots$, $\vec{\mathbf{x}}_{p-1}$ can be written as their linear combination, one can find coefficients $b_1, \ldots, b_{p-1}$ such that

$$\vec{\mathbf{x}}_p = b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2 + \cdots + b_{p-1}\vec{\mathbf{x}}_{p-1}.$$

A more symmetrical way to write this condition is to say that there is a solution to the equation

$$b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2 + \cdots + b_p\vec{\mathbf{x}}_p = 0. \tag{2.19}$$

For the three vectors above, this combination is

$$3\vec{\mathbf{x}}_1 - 2\vec{\mathbf{x}}_2 - \vec{\mathbf{x}}_3 = 0.$$

Of course, putting $b_1 = b_2 = \cdots = b_p = 0$ always solves Equation 2.19, but this solution puts no constraint on the vectors and is said to be trivial. When the $\vec{\mathbf{x}}_j$ are multicollinear, there is a solution to Equation 2.19 for which some of the b_j are nonzero. When every vector of a set is necessary, in the sense that none of them can be written as a linear combination of the others and that Equation 2.19 has only the trivial solution, then they are said to be *linearly independent.* A set of vectors that is not linearly independent is *linearly dependent.* The three vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, and $\vec{\mathbf{x}}_3$ are linearly dependent.

Although the criterion of Equation 2.19 gives a formal way to define linear dependence and independence, it is not always obvious how to tell when an equation that links the vectors can be found and when it cannot. A geometric approach to this problem uses the generalization of volume described in Section 2.1. When $\mathrm{vol}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_p) = 0$, the set of vectors has no volume within the p-dimensional space and falls entirely in a space with fewer than p dimensions.[1] The number of independent vectors in a set cannot exceed the dimension of the space that contains them—try to visualize a counterexample—so such a set of p vectors cannot be linearly independent. For example, the three-dimensional volume defined by the three vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, and $\vec{\mathbf{x}}_3$ above is zero, so they are linearly dependent. The two-dimensional volume of any pair of these vectors is nonzero, indicating that they are not collinear and span a plane.

2.5 Projection onto subspaces

The vector operation known as *projection* divides a vector into two orthogonal components, one lying entirely within a subspace of the original space and the other lying entirely outside that subspace. Many of the operations of multivariate statistics can be described as projection.

The geometry of projection is straightforward. Figure 2.1 shows three vectors, $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$, and $\vec{\mathbf{z}}$ in a three-dimensional space $\mathcal{V}$. A two-dimensional subspace $\mathcal{V}'$ cuts through $\mathcal{V}$. As drawn, the vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ lie above the plane and $\vec{\mathbf{z}}$ lies below it. For each of these vectors, one can find the *projection onto the subspace*, which is the part of the vector that is parallel to $\mathcal{V}'$. These projections are the vectors $\vec{\mathbf{x}}'$, $\vec{\mathbf{y}}'$, and $\vec{\mathbf{z}}'$ shown in the figure, all of which are in the plane $\mathcal{V}'$. The projection operation has a simple physical

[1] Algebraically, this condition means that the covariance matrix is singular.

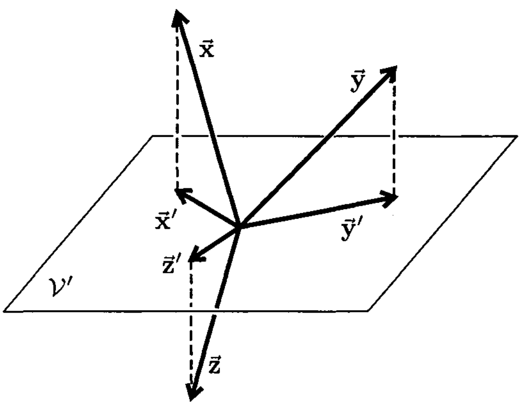


Figure 2.1: *Projection of three vectors $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$, and $\vec{\mathbf{z}}$ onto the vectors $\vec{\mathbf{x}}'$, $\vec{\mathbf{y}}'$, and $\vec{\mathbf{z}}'$ in the two-dimensional subspace $\mathcal{V}'$ of a three-dimensional space.*

analogy. Think of a light source lying very far off and directly above the plane $\mathcal{V}'$ (or below it in the case of $\vec{\mathbf{z}}$), so that its rays fall perpendicularly on the surface. The projections are the shadows cast by the vectors on the plane.

Because the projection $\vec{\mathbf{x}}'$ of the vector $\vec{\mathbf{x}}$ onto the subspace $\mathcal{V}'$ lies in $\mathcal{V}'$, it can be written as a linear combination of any set of basis vectors for $\mathcal{V}'$. When the subspace $\mathcal{V}'$ is one-dimensional, the projection is easy to calculate. Let $\vec{\mathbf{v}}$ be any vector in $\mathcal{V}'$. The projected vector $\vec{\mathbf{x}}'$ is a multiple of $\vec{\mathbf{v}}$, the multiplier being determined by the dot product of the two vectors and the length of $\vec{\mathbf{v}}$:

$$\vec{\mathbf{x}}' = \left(\frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{v}}}{|\vec{\mathbf{v}}|^2}\right)\vec{\mathbf{v}}. \tag{2.20}$$

This operation is often referred to as projecting $\vec{\mathbf{x}}$ onto the vector $\vec{\mathbf{v}}$, rather than onto the subspace generated by $\vec{\mathbf{v}}$. When $\mathcal{V}'$ is multidimensional, the calculation is somewhat more complicated. Specific formulae are discussed in Chapter 4 with those for multiple regression.

Unless the vector whose projection is being taken already lies in the subspace $\mathcal{V}'$, the projection $\vec{\mathbf{x}}'$ is not the same as the original vector $\vec{\mathbf{x}}$. The portion of $\vec{\mathbf{x}}$ that is not in $\mathcal{V}'$ is found by subtracting the projection from the original vector:

$$\vec{\mathbf{x}}'_{\perp} = \vec{\mathbf{x}} - \vec{\mathbf{x}}'. \tag{2.21}$$

This vector lies entirely outside $\mathcal{V}'$ and is orthogonal to the projection $\vec{\mathbf{x}}'$. In this way, the original vector is divided into two orthogonal additive

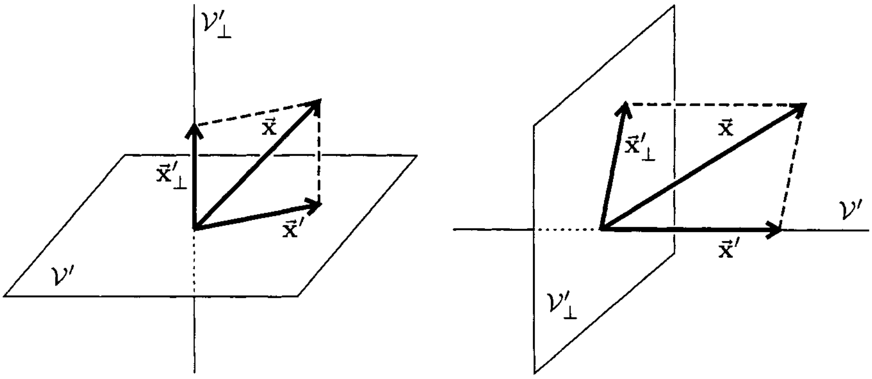


Figure 2.2: *The projection of a vector* $\vec{x}$ *onto a subspace* $\mathcal{V}'$ *and onto the orthogonal complement* $\mathcal{V}'_\perp$ *of* $\mathcal{V}'$. *In the left panel* $\mathcal{V}'$ *is two-dimensional and in the right panel it is one-dimensional.*

portions:

$$\vec{x} = \vec{x}' + \vec{x}'_\perp. \tag{2.22}$$

The dashed lines in Figure 2.1 are these vectors. This decomposition into orthogonal components is an essential part of much of multivariate statistics. In particular, multiple regression, discussed in the next chapters, is an instance of this operation.

Projections of vectors are made with respect to a subspace—the subspace $\mathcal{V}'$ in Figure 2.1. The operation also creates a second subspace $\mathcal{V}'_\perp$ that contains vectors such as $\vec{x}'_\perp$. Let $\mathcal{V}$ denote the overall vector space within which $\mathcal{V}'$ is a subspace. The *orthogonal complement* of $\mathcal{V}'$ in $\mathcal{V}$ is the set of all vectors in $\mathcal{V}$ that are orthogonal to those in $\mathcal{V}'$. In this book, orthogonal complements are denoted by a $\perp$ in the subscript, often with an additional symbol to show the defining subspace—for example, the orthogonal complement of a space $\mathcal{V}_X$ is denoted $\mathcal{V}_{\perp X}$, both symbols without primes. Except for the trivial vector $\vec{0}$ with no length, a subspace and its orthogonal complement have no common members. Figure 2.2 shows two examples of orthogonal complements, both embedded in a three-dimensional space. In the left panel the defining space $\mathcal{V}'$ is two dimensional, indicated by the plane, and so $\mathcal{V}'_\perp$ is a one-dimensional vertical line. In the right panel $\mathcal{V}'$ is unidimensional—the line from left to right—and $\mathcal{V}'_\perp$ is a two-dimensional plane.

The orthogonal complement of a subspace is itself a subspace. For example, the line and the plane that form the orthogonal complements in

Figure 2.2 are like any other line or plane. This fact is important, since that means that vector operations can be carried on naturally within the orthogonal complement. The characteristic of $\mathcal{V}'_\perp$ that makes it a vector space is that it is closed under linear combination. Intuitively, the closed character is obvious. Combining vectors from the plane $\mathcal{V}'_\perp$ in the right panel of Figure 2.2 cannot produce a vector outside the plane. This result is easily proved more formally. Suppose that $\vec{\mathbf{x}}$ is any vector in $\mathcal{V}'$ and $\vec{\mathbf{y}}$ and $\vec{\mathbf{z}}$ are members of $\mathcal{V}'_\perp$. By definition $\vec{\mathbf{y}}$ and $\vec{\mathbf{z}}$ are orthogonal to $\vec{\mathbf{x}}$, so their dot product with $\vec{\mathbf{x}}$ is zero. Now take any linear combination $\vec{\mathbf{w}} = a\vec{\mathbf{y}} + b\vec{\mathbf{z}}$. For $\mathcal{V}'_\perp$ to be a subspace, this new vector must also be in $\mathcal{V}'_\perp$, and so it must be orthogonal to $\vec{\mathbf{x}}$. Using the distributive law (Equations 2.7), the dot product is

$$\begin{aligned} \vec{\mathbf{w}}\cdot\vec{\mathbf{x}} &= (a\vec{\mathbf{y}} + b\vec{\mathbf{z}})\cdot\vec{\mathbf{x}} \\ &= a(\vec{\mathbf{y}}\cdot\vec{\mathbf{x}}) + b(\vec{\mathbf{z}}\cdot\vec{\mathbf{x}}) \\ &= a\times 0 + b\times 0 = 0. \end{aligned}$$

Thus, $\vec{\mathbf{w}} \perp \vec{\mathbf{x}}$ and $\mathcal{V}'_\perp$ is closed under linear combination, as required.

A subspace and its orthogonal complement have a symmetrical relationship in the large vector space. Just as $\mathcal{V}'_\perp$ is the orthogonal complement of $\mathcal{V}'$ within the overall space $\mathcal{V}$, so also $\mathcal{V}'$ is the orthogonal complement of $\mathcal{V}'_\perp$. One should think of the two subspaces as splitting the direction and length of vectors in $\mathcal{V}$ into two parts, one in each subspace, in such a way that between them $\mathcal{V}$ is completely represented. Every vector in $\mathcal{V}$ has a part (possibly the null vector $\vec{\mathbf{0}}$) in $\mathcal{V}'$ and a part (also possibly null) in $\mathcal{V}'_\perp$. The split is analogous to distinguishing up-down movement from horizontal movement in normal three-dimensional space.

A subspace and its orthogonal complement divide up the dimensions of the original space. Two-dimensional horizontal movement and one-dimensional vertical movement give a three-dimensional position. The dimensions of a subspace and its orthogonal complement sum to that of the enclosing space:

$$\dim(\mathcal{V}') + \dim(\mathcal{V}'_\perp) = \dim(\mathcal{V}). \tag{2.23}$$

The splitting of the dimensions between a projection and its orthogonal complement plays a very important part in the construction of statistical tests, described in Chapter 6.

Exercises

1. Let $\vec{\mathbf{x}} = [3,\ 2]'$, $\vec{\mathbf{y}} = [-1,\ 1]'$, and $\vec{\mathbf{z}} = [-2,\ -1]'$. Draw a picture of the vectors, and calculate the following quantities. Where the result is a vector, include it in the picture.

a. $\vec{x} + \vec{y}$

b. $\vec{x} - \vec{z}$

c. $2\vec{x} - 3\vec{y} + 4\vec{z}$

d. $|\vec{z}|$

e. $\vec{x} \cdot \vec{y}$

f. $\angle(\vec{y}, \vec{z})$

g. $\text{vol}(\vec{y}, \vec{z})$

2. Suppose that $\vec{u}$ goes from the point $(-1, 3)$ to the point $(3, 7)$ and $\vec{v}$ goes from point $(0, 2)$ to $(4, 0)$. Draw a picture and find $\angle(\vec{u}, \vec{v})$?

3. Calculate the angle between the two vectors in Problem 1.2.

4. Suppose that $\vec{x}$, $\vec{y}$, and $\vec{z}$ are orthonormal vectors and that

$$\vec{u} = a\vec{x} + b\vec{y} \qquad \text{and} \qquad \vec{v} = a\vec{x} + b\vec{z}.$$

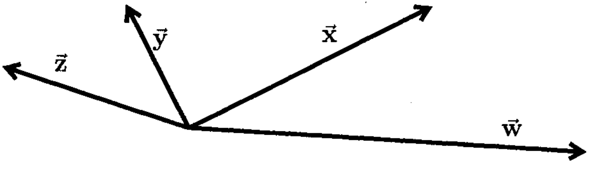
Moreover, let $\vec{u}$ and $\vec{v}$ have unit length, and suppose that$\angle(\vec{u}, \vec{v}) = 45°$. Find a and b. Calculations similar to these are used in the factor analysis models of Section 9.4.

5. Consider the following small set of bivariate data:

X	10	5	1	6	7	3	4	5	1	8
Y	2	4	4	2	4	5	4	5	6	4

Calculate the means, standard deviations, and correlation between the variables, and plot them as a scatterplot and as centered vectors.

6. From the following vector diagram, find the standard deviations of the variables and the correlations between them. Assume that the standard deviation of X equals 5.



7. Use vector constructions to determine which of the two correlation matrices below represent a realizable configuration of variables (i.e., which can be constructed in three-dimensional space).

$$\begin{bmatrix} 1.0 & 0.6 & 0.7 \\ 0.6 & 1.0 & 0.8 \\ 0.7 & 0.8 & 1.0 \end{bmatrix} \qquad \begin{bmatrix} 1.0 & -0.6 & -0.7 \\ -0.6 & 1.0 & -0.8 \\ -0.7 & -0.8 & 1.0 \end{bmatrix}$$

8. Draw the subspace of two-dimensional space spanned by the vectors

a. $\vec{1} = [1, 1]'$

b. $\vec{x}_1 = [2, 1]'$

c. $\vec{x}_1 = [2, 1]'$ and $\vec{x}_2 = [3, 1]'$

d. $\vec{x}_1 = [2, 1]'$ and $\vec{x}_3 = [-4, -2]'$

9. Show that the three vectors $\vec{x} = [3, 1, 2]'$, $\vec{y} = [1, 1, 2]'$, and $\vec{z} = [5, 3, 6]'$ are linearly dependent by finding the coefficients that make them satisfy Equation 2.19. Hint: Some trial and error may be necessary. The coefficients are small integers.

10. Which of the following pairs of vectors are orthogonal bases for a two-dimensional subspace of three-dimensional space. Is any pair an orthonormal basis?

a. $\vec{x}_1 = [1, 2, 3]'$ and $\vec{x}_2 = [2, 1, 1]'$

b. $\vec{x}_1 = [1, 2, 3]'$ and $\vec{x}_3 = [-2, -4, -6]'$

c. $\vec{x}_1 = [1, 2, 3]'$ and $\vec{x}_4 = [1, 1, -1]'$

d. $\vec{x}_1 = [1, 2, 3]'$ and $\vec{x}_5 = [1, -5, 3]'$

11. For the following pairs of vectors, find the vector $\vec{w}$ that is the projection of $\vec{y}$ onto $\vec{x}$. Also find $\vec{z} = \vec{y} - \vec{w}$? Draw a diagram.

a. $\vec{x} = [2, 1]'$ and $\vec{y} = [3, 1]'$

b. $\vec{x} = [-1, 3]'$ and $\vec{y} = [2, -6]'$

c. $\vec{x} = [1, 2, 3]'$ and $\vec{y} = [2, 1, 1]'$

d. $\vec{x} = [1, 1, -1]'$ and $\vec{y} = [1, 2, 3]'$

12. What is the dimension of the orthogonal complement (in three-dimensional space) of the space spanned by the vectors in Problem 9?

Cosines to angles

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	90°	89°	89°	88°	88°	87°	87°	86°	85°	85°
0.10	84°	84°	83°	83°	82°	81°	81°	80°	80°	79°
0.20	78°	78°	77°	77°	76°	76°	75°	74°	74°	73°
0.30	73°	72°	71°	71°	70°	70°	69°	68°	68°	67°
0.40	66°	66°	65°	65°	64°	63°	63°	62°	61°	61°
0.50	60°	59°	59°	58°	57°	57°	56°	55°	55°	54°
0.60	53°	52°	52°	51°	50°	49°	49°	48°	47°	46°
0.70	46°	45°	44°	43°	42°	41°	41°	40°	39°	38°
0.80	37°	36°	35°	34°	33°	32°	31°	30°	28°	27°
0.90	26°	24°	23°	22°	20°	18°	16°	14°	11°	8°
1.00	0°									

Angles to cosines

	0°	1°	2°	3°	4°	5°	6°	7°	8°	9°
0°	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
10°	0.98	0.98	0.98	0.97	0.97	0.97	0.96	0.96	0.95	0.95
20°	0.94	0.93	0.93	0.92	0.91	0.91	0.90	0.89	0.88	0.87
30°	0.87	0.86	0.85	0.84	0.83	0.82	0.81	0.80	0.79	0.78
40°	0.77	0.75	0.74	0.73	0.72	0.71	0.69	0.68	0.67	0.66
50°	0.64	0.63	0.62	0.60	0.59	0.57	0.56	0.54	0.53	0.52
60°	0.50	0.48	0.47	0.45	0.44	0.42	0.41	0.39	0.37	0.36
70°	0.34	0.33	0.31	0.29	0.28	0.26	0.24	0.22	0.21	0.19
80°	0.17	0.16	0.14	0.12	0.10	0.09	0.07	0.05	0.03	0.02
90°	0.00									

Table 2.1: *A short table for converting between cosines and angles. The rows give the first digits and the columns the final digit; for example, a cosine of 0.31 corresponds to an angle of 72°, as found in row 0.30 and column 0.01 of the top table. Negative cosines correspond to angles between 90° and 180° and are found from the relationship* $\cos(\theta) = -\cos(180° - \theta)$; *for example,* $\cos(130°) = -\cos(50°) = -0.64$. *Similarly, if* $\cos(\theta) = -0.55$, *then* $180° - \theta = 57°$, *and so* $\theta = 123°$. *The cosine function is symmetrical about zero, so that the cosines of positive and negative angles are identical.*

Chapter 3

Bivariate regression

Linear regression is the simplest statistical procedure that gains from geometric visualization. In turn, bivariate regression is the simplest version of this procedure. Although all visualization in bivariate regression is easily done with a scatterplot in variable space, the subject-space picture is worth examining, particularly the aspects of it that do not change when one introduces multiple predictors. As a bonus, the geometry gives a natural way to develop the regression formulae.

3.1 Selecting the regression vector

A regression problem starts with two observed variables X and Y, the *predictor* and the *outcome*. Suppose that an imperfect linear relationship exists between these variables and that one wants to use this relationship to predict Y from X. The *regression equation* gives this prediction. The predicted value of Y is a new variable $\widehat{\mathsf{Y}}$ that is a linear function of X:

$$\widehat{\mathsf{Y}} = a + b\mathsf{X}. \tag{3.1}$$

The coefficients a and b are chosen to minimize the discrepancy between $\widehat{\mathsf{Y}}$ and Y.

In vector terms, the three variables X, Y, and $\widehat{\mathsf{Y}}$ are represented by three vectors $\vec{\mathbf{X}}$, $\vec{\mathbf{Y}}$, and $\vec{\widehat{\mathbf{Y}}}$. Since the variables may have non-zero means, these vectors are denoted by uppercase letters. For most purposes, it suffices to work with centered vectors $\vec{\mathbf{x}}$, $\vec{\mathbf{y}}$, and $\vec{\widehat{\mathbf{y}}}$, obtained by subtracting the means. In the centered regression, the constant a is zero, and the vector counterpart to Equation 3.1 is

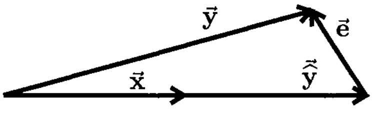
$$\vec{\widehat{\mathbf{y}}} = b\vec{\mathbf{x}}. \tag{3.2}$$

This section and the following one discuss this centered form of regression, and uncentered scores are treated in Section 3.3.

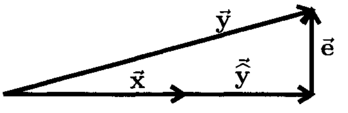
As Equation 3.2 shows, the prediction $\hat{\vec{\mathbf{y}}}$ is a scalar multiple of the predictor $\vec{\mathbf{x}}$. The two vectors are collinear. Unless the correlation between X and Y is perfect (which makes the problem trivial), the vector $\vec{\mathbf{y}}$ points in a different direction from $\vec{\mathbf{x}}$ and $\hat{\vec{\mathbf{y}}}$. Making $\hat{\vec{\mathbf{y}}}$ as similar to $\vec{\mathbf{y}}$ as possible amounts to choosing the proper value of b. This choice is helped by introducing an *error vector* to express the difference between the outcome and the predication:

$$\vec{\mathbf{e}} = \vec{\mathbf{y}} - \hat{\vec{\mathbf{y}}}. \tag{3.3}$$

It is easiest to visualize $\vec{\mathbf{e}}$ drawn not from the origin, but from the tip of $\hat{\vec{\mathbf{y}}}$ to the tip of $\vec{\mathbf{y}}$ (remember that the direction and length define a vector and that it can be started anywhere):



The regression prediction $\hat{\vec{\mathbf{y}}}$ is the vector that has the smallest error, that is, the vector for which $\vec{\mathbf{e}}$ is shortest. Clearly, the vector $\hat{\vec{\mathbf{y}}}$ that is closest to $\vec{\mathbf{y}}$ is the one whose tip is directly below $\vec{\mathbf{y}}$:



This configuration occurs when $\vec{\mathbf{e}} \perp \hat{\vec{\mathbf{y}}}$.

The orthogonality of $\vec{\mathbf{e}}$ and $\vec{\mathbf{x}}$ lets one calculate b. When b is chosen correctly, the dot product $\vec{\mathbf{x}}\cdot\vec{\mathbf{e}}$ is zero. Now use Equations 3.2 and 3.3 to replace $\vec{\mathbf{e}}$ by $\vec{\mathbf{y}} - b\vec{\mathbf{x}}$, giving

$$\vec{\mathbf{x}} \cdot (\vec{\mathbf{y}} - b\vec{\mathbf{x}}) = 0.$$

Applying the distributive law for dot products to the difference gives

$$\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} - b(\vec{\mathbf{x}}\cdot\vec{\mathbf{x}}) = 0,$$

or

$$\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} - b|\vec{\mathbf{x}}|^2 = 0.$$

The optimal regression coefficient is obtained by solving this equation:

$$b = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{|\vec{\mathbf{x}}|^2}. \tag{3.4}$$

Translated into algebraic notation, Equation 3.4 is the usual formula for the regression coefficient:

$$b = \frac{\sum x_i y_i}{\sum x_i^2}.$$

This equation has been obtained through the geometry, without recourse to such techniques as the differential calculus.

The centered score vectors for the data in Figure 1.2 are

$$\vec{\mathbf{x}} = [-3,\ -3,\ -1,\ -1,\ 0,\ 0,\ 1,\ 2,\ 2,\ 3]'$$

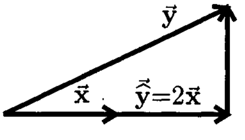
and

$$\vec{\mathbf{y}} = [-8,\ -5,\ -3,\ 0,\ -1,\ 0,\ 5,\ 1,\ 6,\ 5]'.$$

For these scores, $|\vec{\mathbf{x}}|^2 = 38$, $\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} = 76$, and the regression coefficient is

$$b = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{|\vec{\mathbf{x}}|^2} = \frac{76}{38} = 2.00.$$

The prediction vector $\hat{\vec{\mathbf{y}}}$ is exactly twice as long as $\vec{\mathbf{x}}$:



Replacing the dot product with its equivalent expression in terms of the angle between the predictor and outcome vector (Equation 2.4) shows that the regression coefficient equals the correlation between the two variables scaled by the lengths of the two vectors:

$$b = \cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})\,\frac{|\vec{\mathbf{y}}|}{|\vec{\mathbf{x}}|} = r_{\mathsf{xy}}\frac{|\vec{\mathbf{y}}|}{|\vec{\mathbf{x}}|}. \tag{3.5}$$

Substituting this expression into the regression equation and rearranging terms gives

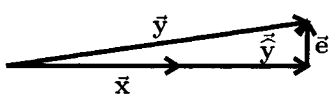
$$\frac{\hat{\vec{\mathbf{y}}}}{|\vec{\mathbf{y}}|} = \cos\angle(\vec{\mathbf{x}}, \vec{\mathbf{y}})\,\frac{\vec{\mathbf{x}}}{|\vec{\mathbf{x}}|}.$$

The prediction vector is always shorter relative to the vector being predicted than is the predictor relative to its length—the cosine is always less than one in absolute value. In statistical terms, the prediction is less variable than the predictor, substantially so when the association between the variables is low. This phenomenon, known as *regression to the mean*, is a characteristic of prediction under uncertainty and gives the technique its name.

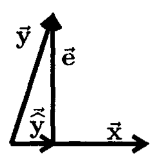
The decomposition of $\vec{y}$ into the regression vector $\vec{\hat{y}}$ and the error vector $\vec{e}$ is an instance of projection into subspaces, as discussed in Section 2.5, particularly Equation 2.22. In this view, $\mathcal{V}$ is the full subject space containing both $\vec{x}$ and $\vec{y}$. The space denoted $\mathcal{V}'$ in Section 2.5 is the one-dimensional subspace $\mathcal{V}_X$ spanned by $\vec{x}$. The vector $\vec{\hat{y}}$ is the projection of $\vec{y}$ onto this subspace. Its length is given by Equation 2.20, which combines Equations 3.2 and 3.4. The orthogonal complement of $\mathcal{V}_X$ in $\mathcal{V}$ is the subspace $\mathcal{V}_{\perp X}$ of vectors orthogonal to $\vec{x}$, and the vector $\vec{e}$ is the projection of $\vec{y}$ onto this subspace. One can think of $\vec{\hat{y}}$ as the part of $\vec{y}$ in common with $\vec{x}$, and of $\vec{e}$ as the portion unrelated to $\vec{x}$. Although this interpretation of regression as projection seems unnecessarily elaborate here, it provides a connection to multiple regression.

3.2 Measuring goodness of fit

Although the regression predictor $\vec{\hat{y}}$ agrees with $\vec{y}$ as well as possible, the actual agreement of these vectors can be either good or poor. A numerical measure is needed to assess the quality of the fit. A comparison of the subject-space picture when the fit is good and when it is poor suggests two indices. When the fit is good, the angle between $\vec{y}$ and $\vec{\hat{y}}$ is small and $\vec{\hat{y}}$ is long:



In contrast, when the fit is poor, the angle is large and $\vec{\hat{y}}$ is small:



Either the angle between the vectors or their relative lengths can be used to measure the quality of the fit, and both have important counterparts when more predictors are used.

First consider the angle between the vectors $\vec{y}$ and $\vec{\hat{y}}$. Although this angle can be expressed directly, it is more common to use its cosine, the correlation coefficient $R = \cos \angle(\vec{y}, \vec{\hat{y}})$. The uppercase letter here conventionally refers to the correlation of a variable with its prediction. The way that the regression predictor is chosen means that $\angle(\vec{y}, \vec{\hat{y}})$ is always less that $90°$, so R is always between zero and one, with large values indicating that the fit is good. In the single-predictor case R is equal to the absolute value of r_{xy}, but this equality is not generally true when there are more predictors.

The second measure of the quality of the fit is the length of the vector $\vec{\hat{y}}$. Where the fit is good, $\vec{\hat{y}}$ is long, and where the fit is poor, $\vec{\hat{y}}$ is short. Because the actual length of $\vec{\hat{y}}$ depends on the scale of the picture, it is expressed relative to the length of the outcome vector $\vec{y}$. When the fit is good, the two vectors are of nearly equal length, although unless the vectors are collinear, $\vec{\hat{y}}$ is always shorter than $\vec{y}$. When the fit is poor, $|\vec{\hat{y}}|$ is much less than $|\vec{y}|$. In the extreme case when $\vec{x} \perp \vec{y}$ and no prediction can be done at all, the projection of $\vec{y}$ onto $\vec{x}$ is the null vector and $|\vec{\hat{y}}|$ is zero.

The two measures, angle and length, are two ways of looking at the same concept, and they are closely related. The vectors $\vec{y}$, $\vec{\hat{y}}$, and $\vec{e}$ form a right triangle, with the 90° angle between $\vec{\hat{y}}$ and $\vec{e}$. The cosine of an angle in a right triangle is the ratio of the leg adjacent to the angle to the hypotenuse (Equation 2.9), so the correlation is the same as the relative length of $\vec{\hat{y}}$:

$$R = \cos \angle(\vec{y}, \vec{\hat{y}}) = \frac{|\vec{\hat{y}}|}{|\vec{y}|}.$$

Turning this equation about, the correlation is the proportion by which the regression vector is shorter than the vector that it predicts:

$$|\vec{\hat{y}}| = R|\vec{y}|. \tag{3.6}$$

This equation is another manifestation of regression to the mean.

The division of the variability into predicted and unexplained parts follows naturally from the geometry. The lengths of the sides of a right triangle are related by the Pythagorean theorem (Equation 2.8), so that

$$|\vec{\hat{y}}|^2 + |\vec{e}|^2 = |\vec{y}|^2.$$

The squared length of a vector is the sum of squares of the corresponding variable (Equation 2.15), so this equation expresses an additive relationship on the sums of squares. Using the sums-of-squares notation and naming the parts gives the *partition of the sums of squares* for regression:

$$SS_{\text{regression}} + SS_{\text{residual}} = SS_{\text{total}}. \tag{3.7}$$

From the link between the length of $\vec{\hat{y}}$ and the correlation in Equation 3.6,

$$SS_{\text{regression}} = |\vec{\hat{y}}|^2 = R^2|\vec{y}|^2 = R^2 SS_{\text{total}}. \tag{3.8}$$

Combining this equation with the partition of the sums of squares in Equation 3.7 gives a comparable expression for the residual sum of squares:

$$SS_{\text{residual}} = |\vec{e}|^2 = (1 - R^2) SS_{\text{total}}. \tag{3.9}$$

These quantities have an important role in statistical testing, a topic that is discussed in Chapter 6.

These results, like the basic regression prediction equation in the last section, can also be derived algebraically. Here, they arise naturally from the geometry.

3.3 Means and the regression intercept

The analysis in Sections 3.1 and 3.2 used centered scores from which the mean had been subtracted, not raw scores. This restriction simplifies the description and is consistent with the fact that most of multivariate statistics concerns the relationship among the variables, not their overall level. However, the original mean of most variables is not zero, and it is helpful to see how the means fit into the geometry.

The key point here is that the mean of a variable and the variation of the observations about that mean are geometrically orthogonal. Because of this orthogonality, variation in one domain can be examined separately from that in the other. In regression problems, the mean can be fitted separately from the regression weights.

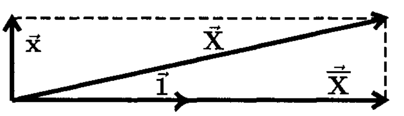
To start out, consider the geometry of an uncentered variable X. The ith observation of this variable is the sum of the mean and the deviation about that mean:

$$X_i = \overline{X} + x_i.$$

Correspondingly, the uncentered vector $\vec{\mathbf{X}}$ is the sum of a vector $\vec{\overline{\mathbf{X}}}$ representing the mean and the centered vector $\vec{\mathbf{x}}$ representing the variability about this common point:

$$\vec{\mathbf{X}} = \vec{\overline{\mathbf{X}}} + \vec{\mathbf{x}}.$$

Graphically, $\vec{\mathbf{X}}$ is resolved into $\vec{\overline{\mathbf{X}}}$ and $\vec{\mathbf{x}}$:



The same mean $\overline{X}$ applies to every observation, so the vector $\vec{\overline{\mathbf{X}}}$ has the algebraic form $[\overline{X}, \overline{X}, \ldots, \overline{X}]'$. It is a scalar multiple of the vector $\vec{\mathbf{1}} = [1, 1, \ldots, 1]'$ that has unit components along each axis:

$$\vec{\overline{\mathbf{X}}} = \overline{X}\vec{\mathbf{1}}.$$

The two vectors $\vec{\overline{\mathbf{X}}}$ and $\vec{\mathbf{x}}$ are orthogonal. Conceptually, this orthogonality is very satisfactory, as one would like the measure of central tendency

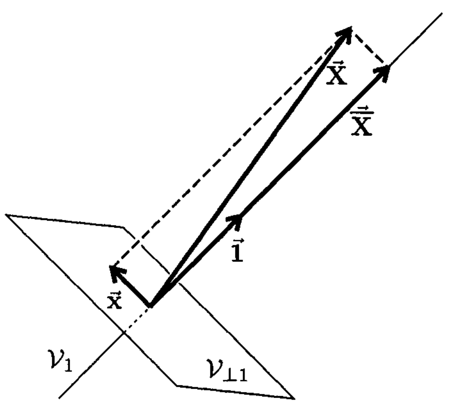


Figure 3.1: *An uncentered vector $\vec{\mathbf{X}}$ is decomposed into the mean vector $\vec{\vec{\mathbf{X}}}$ and the deviations $\vec{\mathbf{x}}$ about the mean, lying in the subspaces $\mathcal{V}_1$ and $\mathcal{V}_{\perp 1}$, respectively.*

to be distinct from how the scores vary about the center. Formally, the orthogonality is easy to prove. Since $\vec{\vec{\mathbf{X}}}$ is a scalar multiple of $\vec{\mathbf{1}}$, one need only show that $\vec{\mathbf{1}} \perp \vec{\mathbf{x}}$, a fact established by calculating the dot product:

$$\begin{aligned} \vec{\mathbf{1}}\cdot\vec{\mathbf{x}} &= \sum x_i \\ &= \sum (X_i - \overline{X}) \\ &= \sum X_i - N\overline{X}. \end{aligned}$$

Substituting the definition of the mean, $\overline{X} = (\sum X_i)/N$, one finds that

$$\vec{\mathbf{1}}\cdot\vec{\mathbf{x}} = \sum X_i - N(\sum X_i)/N = 0.$$

The orthogonality of $\vec{\vec{\mathbf{X}}}$ and $\vec{\mathbf{x}}$ makes the mean more appropriate as a measure of central tendency for this type of analysis than, say, the median or the mode.

Treating the removal of the mean as projection into orthogonal subspaces, as illustrated in Figure 3.1, makes the operation more general. The partition of the scores into the mean vector $\vec{\vec{\mathbf{X}}}$ and a centered vector $\vec{\mathbf{x}}$ is an orthogonal decomposition with respect to $\mathcal{V}_1$. The uncentered vector $\vec{\mathbf{X}}$ lies in the N-dimensional subject space described in Chapter 1. This space contains the subspace $\mathcal{V}_1$ generated by the vector $\vec{\mathbf{1}}$. Vectors in $\mathcal{V}_1$ have the same value in every component and represent constants. The mean $\vec{\overline{\mathbf{X}}}$ is the projection of $\vec{\mathbf{X}}$ onto $\mathcal{V}_1$. The orthogonal complement $\mathcal{V}_{\perp 1}$ of $\mathcal{V}_1$ in the overall space contains vectors whose elements have no common component. It contains only the variation about the mean. Removing the mean vector $\vec{\overline{\mathbf{X}}}$ from $\vec{\mathbf{X}}$ projects it onto the subspace $\mathcal{V}_{\perp 1}$ and creates the vector $\vec{\mathbf{x}}$. This projection argument is the geometric way to interpret the centering of the scores.

Now turn to the regression equation. In vector terms, the unstandardized regression equation is

$$\vec{\hat{\mathbf{Y}}} = a\vec{\mathbf{1}} + b\vec{\mathbf{X}}.$$

As in the centered regression problem, the coefficients a and b are chosen to put $\vec{\hat{\mathbf{Y}}}$ as close to $\vec{\mathbf{Y}}$ as possible. The trick here is to partition the uncentered space into the orthogonal subspaces $\mathcal{V}_1$ and $\mathcal{V}_{\perp 1}$ and to minimize the error separately in each subspace. The best-fitting vector $\vec{\hat{\mathbf{Y}}}$ is the sum of the best-fitting vectors in the two subspaces. The criterion vector $\vec{\mathbf{Y}}$ is the sum of a mean $\vec{\overline{\mathbf{Y}}}$ in $\mathcal{V}_1$ and a deviation $\vec{\mathbf{y}}$ in $\mathcal{V}_{\perp 1}$:

$$\vec{\mathbf{Y}} = \vec{\overline{\mathbf{Y}}} + \vec{\mathbf{y}} = \overline{Y}\vec{\mathbf{1}} + \vec{\mathbf{y}}.$$

The projections of the prediction vector $\vec{\hat{\mathbf{Y}}}$ in the two subspaces are

$$\begin{aligned} \vec{\hat{\mathbf{Y}}} &= a\vec{\mathbf{1}} + b\vec{\mathbf{X}} \\ &= a\vec{\mathbf{1}} + b(\vec{\overline{\mathbf{X}}} + \vec{\mathbf{x}}) \\ &= (a + b\overline{X})\vec{\mathbf{1}} + b\vec{\mathbf{x}}. \end{aligned}$$

To maximize the agreement between $\vec{\hat{\mathbf{Y}}}$ and $\vec{\mathbf{y}}$ in the centered space $\mathcal{V}_{\perp 1}$, one makes $b\vec{\mathbf{x}}$ agree as closely as possible with $\vec{\mathbf{y}}$. This problem is identical to the centered regression problem discussed in the last section. Agreement in $\mathcal{V}_1$ is even easier, since the space is unidimensional. A perfect match occurs when $(a + b\overline{X})\vec{\mathbf{1}}$ is equal to $\overline{Y}\vec{\mathbf{1}}$, which happens when one sets

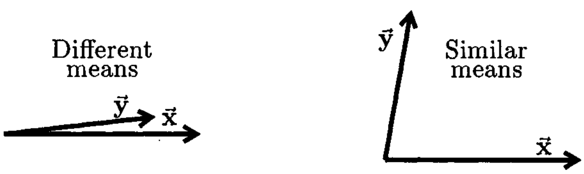
$$a = \overline{Y} - b\overline{X}. \tag{3.10}$$

This assignment is the conventional equation for the intercept in bivariate regression. Once again, it follows from the geometry.

3.4 The difference between two means

Linear regression can be used to solve a problem seemingly quite different from prediction: the assessment of the difference between the means of two groups of scores. Suppose that one wishes to see whether the average value of a variable Y is different in two groups. The most common way to treat this problem is to use a t test, as described in most elementary statistics texts. Although this test is completely satisfactory, an approach to the problem through regression is also valuable, for it introduces some techniques that are needed to analyze more complex designs.

In the regression context, one examines the difference between the means of a pair of groups by looking at the similarity of the score vector $\vec{y}$ to a vector $\vec{x}$ that represents the grouping of the observations. If Y has different means in the two groups, then the grouping vector $\vec{x}$ and the score vector $\vec{y}$ carry similar information and are angularly close to each other. If the means do not differ, then the grouping vector $\vec{x}$ has little in common with the score vector $\vec{y}$ and the angle between them is substantial:



When the two groups have the same mean, $\vec{x}$ and $\vec{y}$ are orthogonal. The magnitude of the relationship, that is, the size of the difference in group means relative to the within-group variability, is measured by the angle between the vectors, by the cosine of this angle, or by the square of this cosine. A test of the statistical significance here is the same as it is for regression (discussed in Chapter 6) and is equivalent to a t test.

The analysis largely follows the regression procedure described above. The only new part is the grouping vector $\vec{x}$. To express the grouping, one creates a variable, all of whose variability is between the groups. This variable takes one value for all the observations in one group and a different value for all the observations in the other group. This group-coding variable is known as a *dummy variable*, and the corresponding vector $\vec{x}$ is a *dummy vector*. With two groups, the dummy variable X is constructed by assigning a single value to all scores in one group and a different value to all scores in the other group. One common assignment is to set X to 1 in one group and -1 in the other, another is to use the numbers 0 and 1. Any other two numbers can be used with the same ultimate effect.

Table 3.1 shows an example with one group of two scores and one group of four scores. The dummy variable X is defined using 1 and -1, and the

	Raw		Centered	
Group	Y_i	X_i	y_i	x_i
1	2	-1	-3	$-\frac{4}{3}$
	3	-1	-1	$-\frac{4}{3}$
2	5	1	0	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	6	1	1	$\frac{2}{3}$
	7	1	2	$\frac{2}{3}$
Mean	5	$\frac{1}{3}$	0	0

Table 3.1: *Two groups of scores and a dummy coding variable.*

uncentered vectors are

$$\vec{\mathbf{X}} = [-1,\ -1,\ 1,\ 1,\ 1,\ 1]' \qquad \text{and} \qquad \vec{\mathbf{Y}} = [2,\ 3,\ 5,\ 6,\ 6,\ 7]'.$$

Variable Y is centered by subtracting the common mean of 5 from each score, thereby projecting $\vec{\mathbf{Y}}$ onto $\mathcal{V}_{\perp 1}$ to give $\vec{\mathbf{y}}$. Similarly, X is centered by subtracting $\frac{1}{3}\vec{\mathbf{1}}$ from $\vec{\mathbf{X}}$. The resulting centered vectors are

$$\vec{\mathbf{x}} = [-\tfrac{4}{3},\ -\tfrac{4}{3},\ \tfrac{2}{3},\ \tfrac{2}{3},\ \tfrac{2}{3},\ \tfrac{2}{3}]' \qquad \text{and} \qquad \vec{\mathbf{y}} = [-3,\ -1,\ 0,\ 1,\ 1,\ 2]'.$$

The lengths of $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$ are 2.309 and 4.000, respectively, and $\vec{\mathbf{x}}\cdot\vec{\mathbf{y}} = 8.000$. Applying Equation 2.5, the cosine of the angle between them is

$$R = \cos \angle(\vec{\mathbf{x}}, \vec{\mathbf{y}}) = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{|\vec{\mathbf{x}}||\vec{\mathbf{y}}|} = \frac{8.000}{2.309 \times 4.000} = 0.866.$$

The angle itself is 30°, which is small enough to indicate an appreciable difference between the groups.

In the two-group problem, the choice of the numbers that code the dummy variables is unimportant. The dummy-vector space $\mathcal{V}_x$ is one dimensional. Consequently, every possible centered dummy vector has the same orientation, except possibly for a reflection through the origin. In this unidimensional space, the prediction $\hat{\vec{\mathbf{y}}}$ does not depend on how the groups are coded in the dummy vector $\vec{\mathbf{x}}$. It is fairly easy to see why $\mathcal{V}_{\mathsf{X}}$ is unidimensional here. First note that any dummy variable X can be written as a linear combination of two variables U_1 and U_2 that take the value 1 for subjects in one group and 0 for those in the other group. The individual

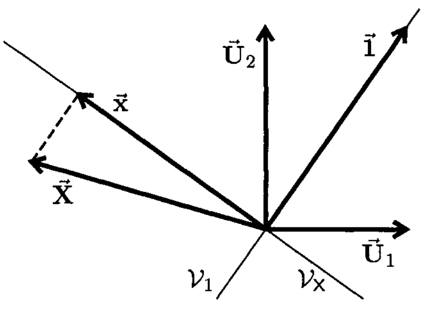


Figure 3.2: *The two-dimensional space $\mathcal{V}_\mathrm{U}$ containing the dummy variable coding for the difference between groups for the data in Table 3.1.*

scores for these variables are

$$U_{i1} = \begin{cases} 1, & \text{if subject } i \text{ is in group 1,} \\ 0, & \text{if subject } i \text{ is in group 2,} \end{cases}$$

and

$$U_{i2} = \begin{cases} 0, & \text{if subject } i \text{ is in group 1,} \\ 1, & \text{if subject } i \text{ is in group 2.} \end{cases}$$

The corresponding vectors $\vec{\mathbf{U}}_1$ and $\vec{\mathbf{U}}_2$ are orthogonal (one of the two variables is zero for every observation), and their squared lengths are equal to the number of subjects in the individual group. Figure 3.2 shows the configuration of these vectors for the data in Table 3.1. The dummy vector used in Table 3.1 is the difference of these vectors:

$$\vec{\mathbf{X}} = \vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_1.$$

The space $\mathcal{V}_\mathrm{U}$ that contains all possible dummy vectors is generated by the two vectors $\vec{\mathbf{U}}_1$ and $\vec{\mathbf{U}}_2$. The vector $\vec{\mathbf{1}} = \vec{\mathbf{U}}_1 + \vec{\mathbf{U}}_2$ is a member of this space, and so $\mathcal{V}_1$ is a subspace of $\mathcal{V}_\mathrm{U}$. However, vectors in this space do not discriminate between the groups and so are not useful dummy vectors. The subspace $\mathcal{V}_\mathrm{X}$ in which discrimination between the groups takes place is the orthogonal complement of $\mathcal{V}_1$ in $\mathcal{V}_\mathrm{U}$. Since $\mathcal{V}_\mathrm{U}$ is two dimensional (it is generated by $\vec{\mathbf{U}}_1$ and $\vec{\mathbf{U}}_2$) and $\mathcal{V}_1$ is one dimensional (it is generated by $\vec{\mathbf{1}}$), the subspace $\mathcal{V}_\mathrm{X}$ is one dimensional. The unidimensionality of $\mathcal{V}_\mathrm{X}$ means that dummy vectors $\vec{\mathbf{x}} \in \mathcal{V}_\mathrm{X}$ can differ only their length, not (except for 180° rotation) in their angular relationship to $\vec{\mathbf{y}}$.

Although the choice of dummy coding has no impact on the two-group analysis, it is important in the multigroup analysis of variance. The space of vectors that discriminate among more than two groups is multidimensional, and the form of the coding influences the interpretation of the analysis. Appropriate coding schemes and their relationship to useful hypotheses are discussed in Chapter 8.

Exercises

1. Calculate the regression of Y onto X for the data in Problem 2.5. Illustrate the result with a diagram of the centered vectors.

2. Draw a diagram of the uncentered predictor space comparable to Figure 3.1 for the regression in Problem 1. Show the vectors $\vec{1}$, $\vec{X}$ and $\vec{x}$ in their proper angular relationships.

3. Two independent groups contain the following scores:

Group 1	5	4	3	7	4	5	7		
Group 2	11	10	8	5	11	11	7	10	8

a. Use a t test to investigate the difference between the means of these groups.

b. Construct the dummy variable $X = U_1 - U_2$, and calculate R^2 for the hypothesis of equal group means.

c. Illustrate the regression analysis with subject-space pictures.

Chapter 4

Multiple regression

In univariate regression, one predictor variable X is used to estimate one outcome variable Y. In *multiple regression* the same goal of predicting a single variable Y remains, but several regression predictors X_1, X_2, ..., X_p are used. The outcome variable Y is estimated by a linear combination that uses all the X_j:

$$\widehat{Y} = a + b_1 X_1 + b_2 X_2 + \cdots + b_p X_p,$$

or in geometric form,

$$\vec{\widehat{\mathbf{Y}}} = a\vec{\mathbf{1}} + b_1 \vec{\mathbf{X}}_1 + b_2 \vec{\mathbf{X}}_2 + \cdots + b_p \vec{\mathbf{X}}_p.$$

Using many predictors lets a more accurate match to $\vec{\mathbf{Y}}$ be made. As in the univariate case, the core of the problem is fitting the centered equation:

$$\vec{\widehat{\mathbf{y}}} = b_1 \vec{\mathbf{x}}_1 + b_2 \vec{\mathbf{x}}_2 + \cdots + b_p \vec{\mathbf{x}}_p.$$

The first task in multiple regression is to use a set of data to assign values to the b_j and, if necessary, to a. The second, and often more difficult, part of the analysis is to investigate how the relationship among the X_j helps or hinders the prediction operation. Here the vector representation shows its power.

4.1 The geometry of prediction

Begin with the picture in variable space. Figure 4.1 shows a two-predictor multiple regression as a scatterplot in this space. The observations form a swarm of points in a three-dimensional space. The regression equation

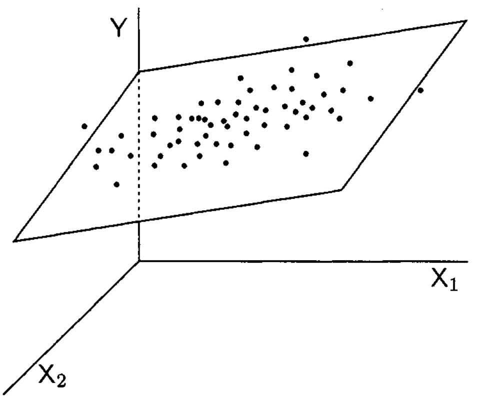


Figure 4.1: *The regression of* Y *onto* X_1 *and* X_2 *as a scatterplot in variable space.*

defines a plane that passes through this swarm, chosen to minimize the vertical deviations of the points about it. Unfortunately, although the idea is simple, the amount of detail in this picture makes it hard to read. The relationship of the individual points to the plain is unclear, and the relationships among the variables are particularly hard to comprehend. These difficulties get worse when there are more than two predictors.

In contrast, the subject-space picture emphasizes the variables and is easier to use. Figure 4.2 shows a two-predictor regression in subject space. The left panel shows the three-dimensional configuration of vectors, and the right panel shows the vectors in the subspace $\mathcal{V}_{\mathsf{X}}$ spanned by $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$. This subspace is two-dimensional and is known as the *regression space.* The prediction $\vec{\hat{\mathbf{y}}} = b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2$ is a linear combination of the predictors, and so lies in this regression space. Among all vectors in $\mathcal{V}_{\mathsf{X}}$, the vector $\vec{\hat{\mathbf{y}}}$ is the closest to $\vec{\mathbf{y}}$. However, since $\vec{\mathbf{y}}$ is not in $\mathcal{V}_{\mathsf{X}}$, perfect prediction is impossible. Some error remains, represented by a vector $\vec{\mathbf{e}} = \vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}$ lying outside $\mathcal{V}_{\mathsf{X}}$, in what is known as the *error space.* The criterion for the best-fitting prediction remains the same as it was in the one-predictor case: to minimize the length of this error vector $\vec{\mathbf{e}}$. Geometric intuition tells what to do here: pick b_1 and b_2 so that the tip of $\vec{\hat{\mathbf{y}}}$ is directly under the tip of $\vec{\mathbf{y}}$.[1] This choice breaks $\vec{\mathbf{y}}$ into two orthogonal components, $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$.

[1] Imagine that you can move about the floor of a room ($\mathcal{V}_{\mathsf{X}}$) and are trying to get close to a point on the ceiling (the tip of $\vec{\mathbf{y}}$). You go to where the point is directly above you.

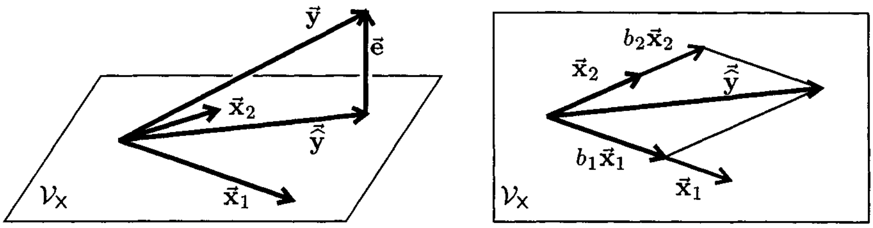


Figure 4.2: *Two-predictor regression in subject space. On the left, a three-dimensional representation; on the right the construction of $\vec{\hat{y}}$ in the two-dimensional regression space $\mathcal{V}_X$ spanned by $\vec{x}_1$ and $\vec{x}_2$.*

The picture in Figure 4.2 is like the one-predictor case in many respects, but differs from it in an important way. With one predictor, the space spanned by $\vec{x}$ is unidimensional, and $\vec{\hat{y}}$ and $\vec{x}$ are collinear. With two or more linearly independent predictors, $\vec{\hat{y}}$ need not be collinear with any of the $\vec{x}_j$. Clearly this is the case in Figure 4.2. The fit of $\vec{\hat{y}}$ to $\vec{y}$ is better than the fit of either predictor alone. The greater freedom of the space $\mathcal{V}_X$ almost always makes the bivariate prediction better than a univariate prediction. Only in a few special cases are they the same, such as when one predictor happens to lie along $\vec{\hat{y}}$ in $\mathcal{V}_X$ or when one predictor is orthogonal both to the other predictor and to $\vec{y}$.

The orthogonality of the decomposition of $\vec{y}$ into $\vec{\hat{y}}$ and $\vec{e}$ is the key to finding the regression coefficients. Consider the two-predictor case. Because $\vec{e}$ is orthogonal to $\mathcal{V}_X$, it is orthogonal to both $\vec{x}_1$ and $\vec{x}_2$. Now express the assertion $\vec{x}_1 \perp \vec{e}$ as a dot product, substitute the definitions of $\vec{e}$ and of $\vec{\hat{y}}$, and distribute the sum:

$$\begin{gathered}\vec{x}_1 \cdot \vec{e} = 0, \\ \vec{x}_1 \cdot (\vec{y} - \vec{\hat{y}}) = 0, \\ \vec{x}_1 \cdot [\vec{y} - (b_1\vec{x}_1 + b_2\vec{x}_2)] = 0, \\ \vec{x}_1 \cdot \vec{y} - b_1(\vec{x}_1 \cdot \vec{x}_1) - b_2(\vec{x}_1 \cdot \vec{x}_2) = 0.\end{gathered}$$

A companion equation derives from the relationship $\vec{x}_2 \perp \vec{e}$. Together they make a pair of simultaneous equations, known as the *normal equations*:[2]

$$\begin{aligned} b_1(\vec{x}_1 \cdot \vec{x}_1) + b_2(\vec{x}_1 \cdot \vec{x}_2) &= \vec{x}_1 \cdot \vec{y}, \\ b_1(\vec{x}_1 \cdot \vec{x}_2) + b_2(\vec{x}_2 \cdot \vec{x}_2) &= \vec{x}_2 \cdot \vec{y}. \end{aligned} \tag{4.1}$$

[2]The word *normal* here does not refer to the normal distribution.

Solving this pair of simultaneous linear equations gives the regression coefficients:

$$\begin{aligned} b_1 &= \frac{(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{y}})(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{x}}_2) - (\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}})(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_2)}{(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_1)(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{x}}_2) - (\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_2)^2}, \\ b_2 &= \frac{(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}})(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_1) - (\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{y}})(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_2)}{(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_1)(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{x}}_2) - (\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_2)^2}. \end{aligned} \tag{4.2}$$

Selecting the regression vector to minimize the error has an important consequence in the multivariate case. No other vector in $\mathcal{V}_X$ makes a smaller angle with $\vec{\mathbf{y}}$ than does $\vec{\hat{\mathbf{y}}}$. The picture makes this fact evident—any vector in the plane of $\mathcal{V}_X$ in Figure 4.2 that does not point along $\vec{\hat{\mathbf{y}}}$ is farther away from $\vec{\mathbf{y}}$. This characteristic of the regression vector is not apparent in the univariate case, where $\vec{\hat{\mathbf{y}}}$ is constrained to lie along $\vec{\mathbf{x}}$. With a multidimensional regression space, one does better.

The analysis depicted in Figure 4.2 uses two predictors. With more than two predictors, the same general representation works, although the geometry is embedded in a space of more than three dimensions, which is more difficult to appreciate in our three-dimensional world. However, after a little practice, it is surprisingly easy to conceptualize a four- or five-dimensional structure in a loose way, using the three-dimensional picture as a guide. As in the simpler case, the predictors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$ span the regression space $\mathcal{V}_X$ and $\vec{\mathbf{y}}$ lies outside this space. The vector $\vec{\hat{\mathbf{y}}}$ is the projection of $\vec{\mathbf{y}}$ onto $\mathcal{V}_X$, and the error of prediction $\vec{\mathbf{e}}$ is the projection of $\vec{\mathbf{y}}$ onto the orthogonal complement $\mathcal{V}_{\perp X}$ of $\mathcal{V}_X$. Treated as an optimization problem, $\vec{\hat{\mathbf{y}}}$ is placed as close to $\vec{\mathbf{y}}$ as possible by minimizing $\vec{\mathbf{e}}$.

Computationally, finding the projection $\vec{\hat{\mathbf{y}}}$ based on many predictors involves solving a set of simultaneous equations that generalizes Equations 4.1. These normal equations derive from the orthogonality of the regression space $\mathcal{V}_X$ and the error $\vec{\mathbf{e}}$, just as they did with two variables. Because $\vec{\mathbf{x}}_j$ and $\vec{\mathbf{e}}$ are orthogonal,

$$\vec{\mathbf{x}}_j \cdot \vec{\mathbf{e}} = \vec{\mathbf{x}}_j \cdot (\vec{\mathbf{y}} - \vec{\hat{\mathbf{y}}}) = 0,$$

or more symmetrically,

$$\vec{\mathbf{x}}_j \cdot \vec{\hat{\mathbf{y}}} = \vec{\mathbf{x}}_j \cdot \vec{\mathbf{y}}. \tag{4.3}$$

Now substitute the regression equation for $\vec{\hat{\mathbf{y}}}$ and distribute the dot product to give the normal equation

$$b_1(\vec{\mathbf{x}}_j \cdot \vec{\mathbf{x}}_1) + b_2(\vec{\mathbf{x}}_j \cdot \vec{\mathbf{x}}_2) + \cdots + b_p(\vec{\mathbf{x}}_j \cdot \vec{\mathbf{x}}_p) = \vec{\mathbf{x}}_j \cdot \vec{\mathbf{y}}. \tag{4.4}$$

Using each $\vec{\mathbf{x}}_j$ in turn, a system of p linear equations in p unknowns is produced. The coefficients of $\vec{\hat{\mathbf{y}}}$ are found by solving this system. Numerically, the best way to solve the normal equations is to use matrix algebra. For one's geometric intuition, it suffices to know that they can be solved.

An uncentered multiple regression problem is analyzed in the same way that the comparable univariate problem was treated in Section 3.3. The uncentered regression equation has a term $a\vec{1}$ lying in the space $\mathcal{V}_1$,

$$\hat{\vec{\mathbf{Y}}} = a\vec{\mathbf{1}} + b_1\vec{\mathbf{X}}_1 + b_2\vec{\mathbf{X}}_2 + \cdots + b_p\vec{\mathbf{X}}_p.$$

The centered regression in $\mathcal{V}_{\perp 1}$ determines the coefficients b_1, b_2, $\ldots$, b_p, and the coefficient a is chosen to make the projections of $\vec{\mathbf{Y}}$ and $\hat{\vec{\mathbf{Y}}}$ onto $\mathcal{V}_1$ identical. The estimate of the regression constant is

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X} - \cdots - b_p\overline{X}_p. \tag{4.5}$$

4.2 Measuring goodness of fit

The solution to the normal equations gives the regression prediction $\widehat{\mathsf{Y}}$, but this equation may provide either a good or a poor fit to Y. There are two aspects to measuring the quality of the fit. The first is to characterize the fit with a descriptive index of the quality of the prediction. The second is to apply sampling theory to decide whether the observed association is real or accidental. The descriptive problem is discussed in this section, and the inferential problem is deferred until Chapter 6.

As in the single-predictor case, the quality of the fit of $\hat{\vec{\mathbf{y}}}$ to $\vec{\mathbf{y}}$ is measured either by the angle between the two vectors or by their relative lengths:

$$R = \cos\angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}}) = \frac{|\hat{\vec{\mathbf{y}}}|}{|\vec{\mathbf{y}}|}. \tag{4.6}$$

This quantity is known as the *multiple correlation coefficient.* Typically R is subscripted with the outcome variable and the predictors. For example, when using the variables X and Z to predict Y, one writes the multiple correlation coefficient as $R_{\mathsf{Y}\cdot\mathsf{XZ}}$. Except in the degenerate case when all but one predictor is useless, the angle between $\vec{\mathbf{y}}$ and $\hat{\vec{\mathbf{y}}}$ is smaller than the angle between any of the $\vec{\mathbf{x}}_j$ and $\vec{\mathbf{y}}$. As the geometry shows, minimizing $|\vec{\mathbf{e}}|$ is comparable to minimizing this angle and, thus, to maximizing the multiple correlation coefficient.

The value of the multiple correlation coefficient can be calculated from its angular definition. To apply Equation 4.6, one needs the lengths of $\vec{\mathbf{y}}$ and $\hat{\vec{\mathbf{y}}}$. Although one can calculate the second of these lengths as $\sqrt{\hat{\vec{\mathbf{y}}}\cdot\hat{\vec{\mathbf{y}}}}$, an expression in terms of the original vectors is more convenient. Start with the dot product, substitute the definition of $\vec{\hat{\mathbf{y}}}$ for its first occurrence, and distribute the sum to get

$$\begin{aligned}\vec{\hat{\mathbf{y}}}\cdot\vec{\hat{\mathbf{y}}} &= (b_1\vec{\mathbf{x}}_1 + b_2\vec{\mathbf{x}}_2 + \cdots + b_p\vec{\mathbf{x}}_p)\cdot\vec{\hat{\mathbf{y}}} \\ &= b_1(\vec{\mathbf{x}}_1\cdot\vec{\hat{\mathbf{y}}}) + b_2(\vec{\mathbf{x}}_2\cdot\vec{\hat{\mathbf{y}}}) + \cdots + b_p(\vec{\mathbf{x}}_p\cdot\vec{\hat{\mathbf{y}}}).\end{aligned}$$

The normal equations in the form of Equation 4.3 let one replace the terms in parentheses by the dot product of the predictors with the outcome, giving

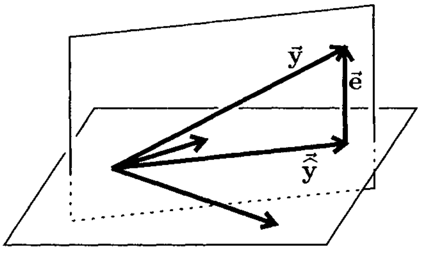
$$|\vec{\hat{\mathbf{y}}}|^2 = b_1(\vec{\mathbf{x}}_1\cdot\vec{\mathbf{y}}) + b_2(\vec{\mathbf{x}}_2\cdot\vec{\mathbf{y}}) + \cdots + b_p(\vec{\mathbf{x}}_p\cdot\vec{\mathbf{y}}).$$

Thus, the square of the multiple correlation coefficient is

$$R^2 = \frac{|\vec{\hat{\mathbf{y}}}|^2}{|\vec{\mathbf{y}}|^2} = \frac{b_1(\vec{\mathbf{x}}_1\cdot\vec{\mathbf{y}}) + b_2(\vec{\mathbf{x}}_2\cdot\vec{\mathbf{y}}) + \cdots + b_p(\vec{\mathbf{x}}_p\cdot\vec{\mathbf{y}})}{\vec{\mathbf{y}}\cdot\vec{\mathbf{y}}}. \tag{4.7}$$

Like all the equations developed from geometric principles, these equations have their counterparts in the algebraic analysis.

The relative lengths of the vectors in multiple regression have the same interpretation as they do with a single predictor. Indeed, when one extracts the plane that contains $\vec{\mathbf{y}}$, $\vec{\hat{\mathbf{y}}}$, and $\vec{\mathbf{e}}$, these vectors have the same right-triangular configuration that they do in the one-predictor regression:



The identification of vector lengths with sums of squares still holds:

$$|\vec{\mathbf{y}}|^2 = SS_{\text{total}}, \tag{4.8}$$

$$|\vec{\hat{\mathbf{y}}}|^2 = SS_{\text{regression}} = R^2 SS_{\text{total}}, \tag{4.9}$$

$$|\vec{\mathbf{e}}|^2 = SS_{\text{residual}} = (1 - R^2)SS_{\text{total}}. \tag{4.10}$$

These expressions give the same decomposition of the sum of squares that appeared in the one-predictor regression (Equation 3.7):

$$SS_{\text{total}} = SS_{\text{regression}} + SS_{\text{residual}}.$$

Uncentered			Centered		
X_1	X_2	Y	X_1	X_2	Y
2	3	15	−3	−4	5
2	5	16	−3	−2	6
4	4	15	−1	−3	5
4	7	10	−1	0	0
5	5	13	0	−2	3
5	8	9	0	1	−1
5	9	8	0	2	−2
6	8	7	1	1	−3
7	7	8	2	0	−2
7	10	5	2	3	−5
8	11	4	3	4	−6

Table 4.1: *Scores for a bivariate regression.*

Table 4.1 and Figure 4.3 show the procedures of the last two sections applied in a two-predictor regression. Raw scores for three variables X_1, X_2, and Y are given on the left of the table. The variables are centered by subtracting the means, which are 5, 7, and 10, to give the scores in the right side of the table. The dot products of the vectors with themselves are 38, 64, and 174, and their lengths are the square roots of these numbers:

$$|\vec{x}_1| = 6.16, \qquad |\vec{x}_2| = 8.00, \qquad \text{and} \qquad |\vec{y}| = 13.19.$$

The dot products of the vectors are

$$\vec{x}_1 \cdot \vec{x}_2 = 40, \qquad \vec{x}_1 \cdot \vec{y} = -73, \qquad \text{and} \qquad \vec{x}_2 \cdot \vec{y} = -100.$$

The correlation between the two predictors is

$$\cos \angle(\vec{x}_1, \vec{x}_2) = \frac{\vec{x}_1 \cdot \vec{x}_2}{|\vec{x}_1||\vec{x}_2|} = 0.811.$$

To draw a picture, the correlation is converted to an angle of 35.8°. Similarly, $\angle(\vec{x}_1, \vec{y}) = 153.9°$ and $\angle(\vec{x}_2, \vec{y}) = 161.4°$. The placement of these vectors is sketched in Figure 4.3. From Equations 4.2, the regression coefficients are $b_1 = -0.807$ and $b_2 = -1.058$, so the prediction vector for the centered variables is

$$\hat{\vec{y}} = -0.807\vec{x}_1 - 1.058\vec{x}_2.$$

The length of this vector is

$$|\hat{\vec{y}}| = \sqrt{b_1(\vec{x}_1 \cdot \vec{y}) + b_2(\vec{x}_2 \cdot \vec{y})} = \sqrt{164.7} = 12.83,$$

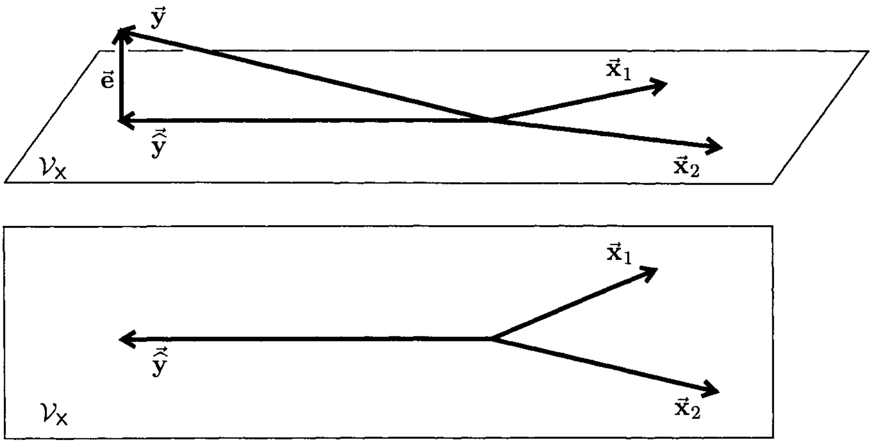


Figure 4.3: *The configuration of vectors for the data in Table 4.1. The upper panel is three dimensional and the lower panel shows the regression space $\mathcal{V}_{\mathsf{X}}$.*

and the multiple correlation coefficient is

$$R = \cos \angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}}) = \frac{|\hat{\vec{\mathbf{y}}}|}{|\vec{\mathbf{y}}|} = \frac{12.83}{13.19} = 0.973.$$

The corresponding angle is 13.3°. The vector $\vec{\mathbf{y}}$ lies very close to the plane spanned by $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$, indicating that accurate prediction is possible.

In the uncentered problem, the regression coefficients b_1 and b_2 remain the same. The constant is obtained from Equation 4.5, and is

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2 = 10 - (-0.807\times 5) - (-1.058\times 7) = 21.441.$$

The complete regression equation in vector form is

$$\hat{\vec{\mathbf{Y}}} = 21.441\vec{\mathbf{1}} - 0.807\vec{\mathbf{X}}_1 - 1.058\vec{\mathbf{X}}_2.$$

The multiple correlation coefficients for the uncentered and centered regressions are identical.

4.3 Interpreting a regression vector

If one intends only to use a multiple regression predictor to estimate values of the variable Y for some new sets of predictors X_j, then one can find the prediction equation without worrying about what it means. Such uses are rare; more often, one also wants to assign meaning to the predictor $\widehat{\mathsf{Y}}$—to say what it represents and to describe how it is related to the X_j. Indeed, despite the prediction-oriented terminology, in many applications of multiple regression the interpretation is the only point of the analysis and actual predictions are never made. However, interpreting a multiple regression variable (or any derived multivariate linear combination) is less obvious than it seems, particularly when the component variables are substantially correlated. The geometric representation is very helpful in working out the relationships among the variables. Both this section and the next chapter concerned the task of interpretation.

One approach to assigning meaning to a created variable such as $\widehat{\mathsf{Y}}$ is to look at the linear combination that forms it; another is to look at its angular relationship to the variables that make it up. Superficially, these approaches seem similar, but they are not the same. As the example at the end of this section shows, they can lead to different conclusions.

Although looking at the coefficients of a linear combination seems a natural way to understand the regression prediction, it is the less useful of the two procedures. The individual coefficients depend on the entire configuration of variables and are rarely meaningful in isolation. The most informative way to use the individual coefficients is to look at the pattern they make and see if it suggests how the new variable (if it has intrinsic meaning) was formed. For example, a linear combination in which all the variables have coefficients with approximately the same value suggests that some sort of sum or average is operating. Similarly, a combination in which two variables figure with approximately equal and opposite sign suggests that the derived variable contrasts values of the two variables, although here some study of the relationships with the other variables is still needed. An interpretation of a linear combination in this way is only useful when the pattern as a whole makes sense and when it points to a mechanism by which the new variable was generated. It is a risky way to interpret one or two coefficients within a larger equation.

Interpretation is usually helped by rounding the coefficients to simplify the equation. For example, suppose that one has obtained the regression equation

$$\hat{\vec{\mathbf{y}}} = 0.23\vec{\mathbf{x}}_1 + 0.18\vec{\mathbf{x}}_2 + 0.27\vec{\mathbf{x}}_3.$$

All the coefficients here are positive and have roughly the same magnitude. The combination is approximated by an equation with equally-weighted predictors:

$$\hat{\vec{\mathbf{y}}}^{\star} = b(\vec{\mathbf{x}}_1 + \vec{\mathbf{x}}_2 + \vec{\mathbf{x}}_3). \tag{4.11}$$

The latter equation is far easier to interpret and may fit only slightly less well than the original equation, particularly if there is considerable sampling variability. One can fit the simplified equation[3] and check whether the angle between $\hat{\vec{y}}^{\star}$ and $\vec{y}$ has declined substantially. Procedures for testing whether the difference in the quality of the fit is statistically reliable are discussed in Sections 6.4 and 7.2.

The values of the coefficients in a linear combination can only be compared when the variables are measured on the same scale. The numerical size of a coefficient is determined both by the relationship between the variables and by the scale on which the variables are measured. For example, the coefficient of a length measure is 2.54 times larger if that variable is measured in inches than if it is measured in centimeters (remember that one inch equals 2.54 centimeters). To make sense of the form of the linear combination, the scales for every variable must be the same. When all the variables measure quantities of the same type (all sizes, error scores, proportions of correct responses, etc.), then their coefficients can be compared directly. However, when the variables are of different types, their natural scales may be incompatible. They must be converted to a common basis before their values are interpreted.

When the members of a set of variables do not have the same natural scale, one way to put them on a comparable basis is to express them relative to their variability. In effect, each variable is treated as if it had the same variance. This standardization can be accomplished in two equivalent ways. One way is to convert all the scores to standard scores by subtracting their mean and dividing by their standard deviation before doing the analysis. Geometrically, this transformation replaces each vector by a new vector that has the same direction but a fixed length. Conceptually, it is easiest to think of these standardized vectors as having unit length. The transformation to standard length does not change any vector's direction, so the space $\mathcal{V}_{\mathsf{X}}$ spanned by the predictors is unchanged and has the same angular relationship to the outcome variable. However, because the lengths of the vectors are changed, their coefficients in $\hat{\vec{y}}$ are different. The unit-length interpretation makes these ***standardized regression coefficients*** b_j^{z} (sometimes called "beta weights") easy to calculate. For example, the two-predictor regression coefficients (Equations 4.2) have the standardized form

$$b_1^{\mathsf{z}} = \frac{(\vec{x}_1 \cdot \vec{y}) - (\vec{x}_2 \cdot \vec{y})(\vec{x}_1 \cdot \vec{x}_2)}{1 - (\vec{x}_1 \cdot \vec{x}_2)^2} \quad \text{and} \quad b_2^{\mathsf{z}} = \frac{(\vec{x}_2 \cdot \vec{y}) - (\vec{x}_1 \cdot \vec{y})(\vec{x}_1 \cdot \vec{x}_2)}{1 - (\vec{x}_1 \cdot \vec{x}_2)^2}, \tag{4.12}$$

where the dot products are equal to correlations. The multiple correlation coefficient, which is the same as that obtained from an unstandardized

[3]To estimate b in Equation 4.11, create the new variable $\mathsf{X_S} = \mathsf{X_1} + \mathsf{X_2} + \mathsf{X_3}$ and regress Y onto it.

regression, is calculated as

$$R^2 = b_1^{\mathbf{z}}(\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{y}}) + b_2^{\mathbf{z}}(\vec{\mathbf{x}}_2 \cdot \vec{\mathbf{y}}) + \cdots + b_p^{\mathbf{z}}(\vec{\mathbf{x}}_p \cdot \vec{\mathbf{y}}) \tag{4.13}$$

(cf. Equation 4.7). When the unstandardized regression coefficients have already been found, it is not necessary to calculate the standardized coefficients from scratch. The unstandardized regression coefficients can be rescaled to get the values that would have been obtained from a standardization of the scores by multiplying them by the ratio of the length of $\vec{\mathbf{x}}_j$ to the length of $\vec{\mathbf{y}}$:

$$b_j^{\mathbf{z}} = \frac{|\vec{\mathbf{x}}_j|}{|\vec{\mathbf{y}}|} b_j. \tag{4.14}$$

This approach is necessary when one is working with an existing regression analysis and no longer has easy access to the original data. However they are obtained, the standardized coefficients are a good place to start when proposing a mechanism for the formation of variables, unless one has a more rational way to scale the variables.

The second way to interpret the prediction $\hat{\vec{\mathbf{y}}}$ is to find its relationship to individual vectors that make it up. If one wishes to talk about the similarity of a predictor to the prediction, then this approach is more appropriate then looking at the regression coefficients. These relationships are measured by looking at the angles $\angle(\vec{\mathbf{x}}_j, \hat{\vec{\mathbf{y}}})$ or at the corresponding correlation coefficients, $\cos\angle(\vec{\mathbf{x}}_j, \hat{\vec{\mathbf{y}}})$. These correlations are known as the *loadings* of the component vectors. Variables with large loadings and small angles between vectors represent similar concepts, while those with small loadings and nearly orthogonal vectors denote different things.

The loadings are calculated from the usual angle formula:

$$\cos\angle(\vec{\mathbf{x}}_j, \hat{\vec{\mathbf{y}}}) = \frac{\vec{\mathbf{x}}_j \cdot \hat{\vec{\mathbf{y}}}}{|\vec{\mathbf{x}}_j||\hat{\vec{\mathbf{y}}}|}.$$

The vector $\hat{\vec{\mathbf{y}}}$ is a solution to the normal equations (Equation 4.3), so the numerator can be replaced by $\vec{\mathbf{x}}_j \cdot \vec{\mathbf{y}}$. Since $|\hat{\vec{\mathbf{y}}}|$ equals $R|\vec{\mathbf{y}}|$ (Equation 4.6), the angle is

$$\cos\angle(\vec{\mathbf{x}}_j, \hat{\vec{\mathbf{y}}}) = \frac{\vec{\mathbf{x}}_j \cdot \vec{\mathbf{y}}}{|\vec{\mathbf{x}}_j|R|\vec{\mathbf{y}}|} = \frac{\cos\angle(\vec{\mathbf{x}}_j, \vec{\mathbf{y}})}{R}. \tag{4.15}$$

Thus, the loadings are proportional to the angles that the vectors make with the original outcome variable $\vec{\mathbf{y}}$.

In some sets of data, the regression weights and the loadings differ substantially, a fact that points up the importance of looking at the geometry

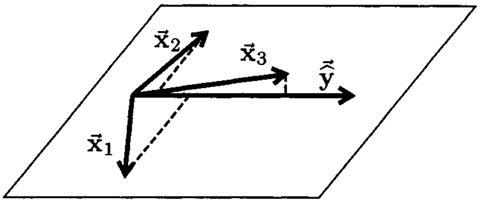


Figure 4.4: *A configuration that shows the importance of examining the loadings of a regression predictor. The vector $\hat{\vec{y}}$ lies in the plane of $\vec{x}_1$ and $\vec{x}_2$, but is angularly closer to $\vec{x}_3$, which is directly above it. The dashed lines project the vectors onto $\hat{\vec{y}}$.*

instead of relying only on the regression coefficients. For example, suppose that a three-predictor regression equation is found to be

$$\hat{\vec{y}} \approx b\vec{x}_1 + b\vec{x}_2 + 0\vec{x}_3.$$

Looking at this equation, one might be tempted to say that Y is substantially and equally related to X_1 and X_2 and that it has nothing to do with X_3. However, this conclusion could well be wrong. The configuration of vectors shown in Figure 4.4 illustrates the type of problem that can arise. This figure shows the three-dimensional regression space containing the predictors and the prediction—presumably the outcome vector $\vec{y}$ lies close to $\hat{\vec{y}}$ along an orthogonal dimension. The vector $\hat{\vec{y}}$ lies in the subspace spanned by $\vec{x}_1$ and $\vec{x}_2$ (shown by the plane in the figure), so it can be written as a linear combination of these two vectors. However, although the vector $\vec{x}_3$ lies outside this subspace, it is very close to it. Even though $\vec{x}_3$ is not required to construct $\hat{\vec{y}}$, it is angularly much more similar to it than are either $\vec{x}_1$ or $\vec{x}_2$. Thus, it makes more sense to say that $\hat{\vec{y}}$ is like $\vec{x}_3$ than that it is like either $\vec{x}_1$ or $\vec{x}_2$ or their difference. To assert that $\hat{\vec{y}}$ is unrelated to $\vec{x}_3$, as one might do by looking only at the regression coefficients, would be a serious error.

The ambiguity in this example occurs in good part because the three predictors fall almost into a plane. Their configuration is more nearly two dimensional then truly three dimensional. As discussed further in Section 5.2, this type of redundant configuration, known as near multicollinearity, creates manifold problems, both for parameter estimation and for interpretation.

Both the regression coefficients and the angular loadings give useful information about how to interpret a regression equation, but neither is invariably more correct. Unless the configuration of vectors is particularly clear—in which case the two approaches will agree with each other—both types of information should be examined in any serious attempt at interpretation.

Exercises

The best way to become comfortable with multivariate geometry is to apply it to describe variables whose meaning you understand. You should supplement the miniature problems in these exercises by looking at the geometry of multiple regression with some real data, preferably some you know well. What are the angular relationships among the original variables—here the predictors X_j and the outcome Y—and between them and any derived vectors such as $\widehat{\mathsf{Y}}$? What subspaces are present and what is their relationship to the old and new vectors? Are there any interesting or unusual relationships such as orthogonality or multicollinearity? Don't start by trying to look at too many variables at once, lest the dimension becomes too great. Describe the picture and what it means to somebody else—a good idea is to trade explanations with another reader of this book. The same thing should be done for the other procedures described in this book.

1. Each of the matrices below gives the covariances of three variables X_1, X_2 and Y (in this order). For each matrix, calculate the centered regression of Y onto the X_j, find R^2, and illustrate the regression with a diagram of $\mathcal{V}_\mathsf{X}$ and a sketch of the three-dimensional configuration.

$$\textbf{a.}\begin{bmatrix} 4.00 & 4.86 & -1.52 \\ 4.86 & 9.00 & -3.36 \\ -1.52 & -3.36 & 4.00 \end{bmatrix} \qquad \textbf{b.}\begin{bmatrix} 1.00 & 0.00 & 2.20 \\ 0.00 & 25.00 & 12.20 \\ 2.20 & 12.20 & 16.00 \end{bmatrix}$$

$$\textbf{c.}\begin{bmatrix} 1.00 & -0.65 & 0.05 \\ -0.65 & 1.00 & -0.52 \\ 0.05 & -0.52 & 1.00 \end{bmatrix} \qquad \textbf{d.}\begin{bmatrix} 1.00 & -0.92 & 0.00 \\ -0.92 & 2.00 & 0.00 \\ 0.00 & 0.00 & 3.00 \end{bmatrix}$$

2. Run an illustrate both a centered and an uncentered regression for the following data:

X_1	15	11	11	8	9	16	12	3	7
X_2	4	8	4	12	9	6	7	11	8
Y	11	18	12	23	21	15	15	22	17

3. The discussion of Figure 4.2 mentioned two special cases where a two-parameter equation fits no better than an equation that uses one of the two variables. Illustrate these configurations with vector diagrams.

4. Find the angles $\angle(\vec{\mathbf{x}}_1, \vec{\hat{\mathbf{y}}})$ and $\angle(\vec{\mathbf{x}}_2, \vec{\hat{\mathbf{y}}})$ for the variables in Problem 1. Do these angles give similar information to the regression coefficients?

5. Another formula for the multiple correlation coefficient uses the generalized volume measure:

$$1 - R_{\mathsf{y}\cdot\mathsf{x}_1\cdots\mathsf{x}_p} = \frac{\mathrm{vol}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_p, \vec{\mathbf{y}})}{\mathrm{vol}(\vec{\mathbf{x}}_1, \ldots, \vec{\mathbf{x}}_p)\mathrm{vol}(\vec{\mathbf{y}})}. \tag{4.16}$$

Draw a diagram showing the areas and volumes in this formula for the one-predictor and two-predictor cases, and try to explain why the formula works. It helps to consider a configuration where the multiple correlation is large and one where it is small. A more general version of this measure is discussed in Section 10.3.

Chapter 5

Configurations of multiple regression vectors

The final section of Chapter 4 introduced a few of the difficulties involved in interpreting a multiple regression equation. The present chapter continues this discussion by examining several situations in which the multivariate relationships among the variables give the analysis special properties or introduce particular complications.

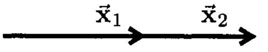
This chapter treats four such situations. The first special relationship occurs when the predictor variables are not a linearly independent set, a state known as multicollinearity. True multicollinearity is not that hard to treat. More insidious is the case when one variable can almost, but not quite, be written as a linear combination of the others. Near multicollinearity, probably the greatest bane of multivariate analysis, is the second topic discussed here. The final two sections describe configurations that are, in a sense, the opposite of multicollinearity, where certain variables are orthogonal.

5.1 Linearly dependent predictors

First consider what happens when some of the predictor variables are linear combinations of the others. The simplest such case occurs when one predictor variable is a linear function of another, say

$$\mathsf{X}_2 = c + d\mathsf{X}_1.$$

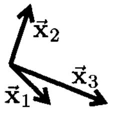
The centered vectors corresponding to these variables are scalar multiples of each other and lie along a single line:



Such vectors, and their corresponding variables, are said to be *collinear*. The multivariate extension of collinearity, known as *multicollinearity*, involves several vectors. Although no two vectors line up, one can write at least one of them as a linear combination of the others.

Consider the three vectors

$$\vec{x}_1 = [2,\ -2]', \qquad \vec{x}_2 = [1,\ 3]', \qquad \text{and} \qquad \vec{x}_3 = [5,\ -2]' :$$



No two of these vectors are collinear, but each is a linear combination of the others; for example,

$$\vec{x}_3 = \tfrac{17}{8}\vec{x}_1 + \tfrac{3}{4}\vec{x}_2.$$

As discussed in Section 2.4, a linear dependency such as this is often written without singling out one variable:

$$17\vec{x}_1 + 6\vec{x}_2 - 8\vec{x}_3 = 0.$$

In general, for p variables, a multicollinear relationship means that

$$a_1\vec{x}_1 + a_2\vec{x}_2 + \cdots + a_p\vec{x}_p = 0$$

with some of the a_j nonzero.

Perfect linear dependence among variables occurs in practice when some constraint restricts their values so that one is a function of the others. A researcher almost never chooses a linearly dependent set of variables intentionally, but can do so when unaware of the constraints. For example, consider a survey in which the respondents are asked to indicate their agreement or disagreement with each of 75 statements. The statements fall into three categories, and the results are scored by counting the number of times that the respondent agrees with a statement in each category. These variables, call them X_1, X_2, and X_3, are used in a regression. The researcher also knows that respondents differ in their overall tendency to make positive responses and believes that the amount of general agreement may help to predict the outcome variable. To allow for this possibility, the proportion X_a of ratings on which a respondent agreed with a statement is included

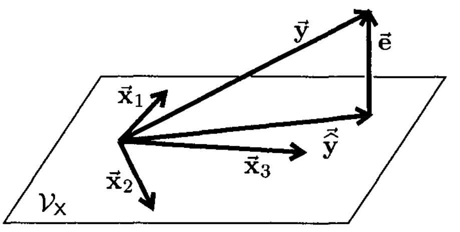


Figure 5.1: *Regression with three linearly dependent predictors. The three vectors $\vec{x}_j$ lie in the plane $\mathcal{V}_X$.*

as a regression variable. Although there are plausible reasons to use each of these four variables, together they form a linearly dependent set. The proportion of agreement depends on the amount of agreement with the individual categories, and the four variables obey the linear constraint

$$X_1 + X_2 + X_3 - 75X_a = 0.$$

Although there are four predictor variables here, only sets of three of them are linearly independent.

The presence of linear dependence among the predictors has several consequences. It complicates some parts of the regression problem, makes other parts unsolvable, and leaves still other parts unchanged. Geometrically, it makes the space $\mathcal{V}_X$ spanned by the predictors $\vec{x}_1$, $\vec{x}_2$, ..., $\vec{x}_p$ identical to the space spanned by a subset of these vectors. The dimension of $\mathcal{V}_X$ is less than would be expected from the number of predictors. In the example above, the $p = 4$ predictors span at most a three-dimensional space. Figure 5.1 shows a configuration of three linearly dependent predictors. The picture is like that of two-predictor regression in Figure 4.2, except that the regression plane contains three vectors.

A comparison of Figures 4.2 and 5.1 points out that certain aspects of the regression problem are unaffected by the introduction of multicollinear predictors. The new variable does not change the size and placement of the prediction space $\mathcal{V}_X$. The regression vector $\hat{\vec{y}}$ and the error vector $\vec{e}$ are determined by the relationship of the space $\mathcal{V}_X$ to the outcome vector $\vec{y}$, not by the particular vectors used to generate this space. This aspect of the problem remains well defined. The unique regression vector $\hat{\vec{y}} \in \mathcal{V}_X$ that is closest to $\vec{y}$ can be found and predictions of $\vec{y}$ can be made. The angle $\angle(\vec{y}, \hat{\vec{y}})$ is well defined, as is the multiple correlation coefficient $R = \cos\angle(\vec{y}, \hat{\vec{y}})$.

In contrast to the existence of a unique $\hat{\vec{y}}$, the coefficients of this vector in terms of the $\vec{x}_j$ are not well defined. The vector $\hat{\vec{y}}$ can be constructed from the $\vec{x}_j$ in infinitely many ways. Unique coefficients b_j in the equation for $\hat{\vec{y}}$ cannot be found, making it impossible to determine the role of an individual predictor in the combination. Because at least some of the $\vec{x}_j$ can be written using the other predictor vectors, an arbitrary proportion of any component along these vectors can be replaced by a combination of the other predictors to give a different set of coefficients.

From a computational point of view, multicollinearity makes the set of normal equations singular, so that any procedure for calculating the b_j fails. For example, with two collinear predictors $\vec{x}_1$ and $\vec{x}_2$, the denominators in both of Equations 4.2 are zero, and the ratios are undefined. Any well written computer program that attempts to calculate the coefficients should balk—above all, it should not transparently complete the analysis. The failure to complete the analysis is a warning to examine one's variables and identify the source of the dependency.

The best way to solve the problem of multicollinearity is to reduce the number of predictor variables. Geometrically, one's goal is to find a set of $\dim(\mathcal{V}_X)$ vectors that spans the space and that has a meaningful interpretation in the original situation. One can either eliminate variables from the predictor set until it becomes linearly independent or combine several predictors into a single variable. Often several alternatives are available, each with interpretative advantages and liabilities. For example, among the ways in which the three opinion measures and overall agreement proportion mentioned above could be brought to a linearly independent set are the following three:

- Eliminate the overall agreement proportion X_a from the regression and analyze X_1, X_2, and X_3 only. By doing so, one keeps the complete set of opinion measures, but loses overall agreement as an explanatory variable.

- Keep X_a, but drop one of the individual agreement measures, thereby keeping overall agreement as part of one's explanation. The cost is that one of the original agreement variables participates only implicitly in the analysis, spoiling its symmetry.

- Keep X_a and define two new variables that are combinations of the X_j and that span the original space when combined with X_a. Good candidates for the combination are the differences between pairs of scores, for example $X_2 - X_1$ and $X_3 - X_1$. By making this substitution, one divides the variables into two subsets: the variable X_a describes overall agreement and the two new variables describe how agreement

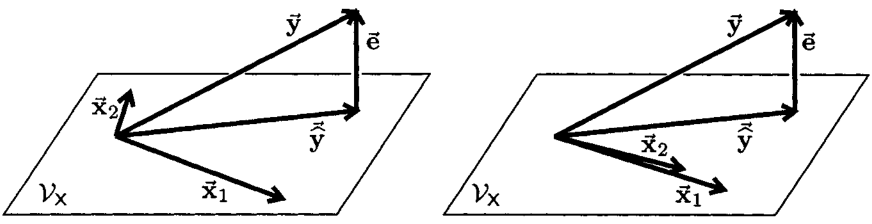


Figure 5.2: *Vector configuration for well conditioned (left) and poorly conditioned predictors in a bivariate regression.*

among the three questions differs. The new variables are farther from the original measures, however, and may be harder to explain.

Each of these approaches leads to a satisfactory result from a technical viewpoint; which one to use should depend on which is easiest to interpret.

5.2 Nearly multicollinear predictors

When the predictors in a regression problem are nearly multicollinear, but are not completely so, the difficulties that arise are more serious than with true multicollinearity. True multicollinearity stands out because the regression problem cannot be solved at all. With nearly multicollinear predictors, however, a regression line can be found and the regression coefficients estimated, but that the line is unstable and the coefficients are very inaccurately determined. The uncertainty can be large enough to make the solution useless either for the prediction of new values of Y or for a theoretical analysis of the relationships among the variables.

Figure 5.2 shows two pictures of regression vectors. The left panel reproduces the satisfactory configuration of distinct predictors from Figure 4.2. The right panel shows a pair of nearly collinear predictors. In both panels, the plane $\mathcal{V}_X$ is defined by $\vec{x}_1$ and $\vec{x}_2$, and $\hat{\vec{y}} \in \mathcal{V}_X$ in this plane estimates $\vec{y}$. The critical difference between the pictures is the placement of $\vec{x}_1$ and $\vec{x}_2$, which are angularly well separated in the left panel and very close to each other in the right panel. A regression problem such at the one on the left is said to be *well conditioned*, that on the right to be *poorly conditioned* or *ill conditioned*.

Everything would still be fine if the predictor vectors in the right panel of Figure 5.2 were known with perfect accuracy. The calculations would give the correct prediction for this configuration. However, in any real situation one does not have completely accurate measurement. Each observation has some uncertainty or sampling error associated with it. Without digressing here into sampling theory (the topic of Chapter 6), it is enough to realize that the placement of the vectors is a bit uncertain, and that in a replication of the study they would be in slightly different places and have slightly different lengths. The effects of this uncertainty are what make near multicollinearity so great a problem.

When a regression problem is poorly conditioned, with predictor variables that are nearly multicollinear, small variations in the variables have large effects on the regression. These effects can be seen by considering the two configurations in Figure 5.2 (which the reader should try to construct with pointers in three dimensions). Think of the sampling variability as producing vectors that are elastic and loosely jointed at $\mathbf{0}$. The angular uncertainty creates the most serious problems. In the well conditioned configuration on the left, small changes in the direction of $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ have little effect on the position of $\mathcal{V}_{\mathrm{X}}$. The wide stance of the vectors gives the plane stability. In contrast, variation in the position of the predictor vectors in the right panel makes the regression plane flap about. For example, suppose that $\vec{\mathbf{x}}_1$ drops slightly below its current position. On the left, little changes; on the right, the plane $\mathcal{V}_{\mathrm{X}}$ tips substantially upward, passing much closer to $\vec{\mathbf{y}}$. On the left, the multiple correlation coefficient $R = \cos\angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}})$ changes but little; on the right, the apparent fit of the regression improves substantially. The sensitivity to small changes in the poorly conditioned configuration amplifies the effects of sampling variability.

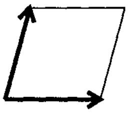
A similar amplification of sampling effects occurs with the estimates of the regression coefficients. As the vectors appear in the right panel of Figure 5.2, $\hat{\vec{\mathbf{y}}}$ is constructed by proceeding in the direction of $\vec{\mathbf{x}}_2$ for many times its length, then returning in the negative $\vec{\mathbf{x}}_1$ direction for a comparable distance. Variation in the angular placement of the vectors within $\mathcal{V}_{\mathrm{X}}$ substantially changes these coefficients. If $\vec{\mathbf{x}}_2$ moves slightly closer to $\vec{\mathbf{x}}_1$, then a considerable increase in the size of both coefficients is necessary to compensate. As with R^2, small sampling effects make large changes in the regression solution.

The difficulties here are caused by the configuration of the predictors, not by either of the predictors alone. A regression based on $\vec{\mathbf{x}}_1$ alone is completely satisfactory, as is one based on $\vec{\mathbf{x}}_2$ alone. Problems only arise when both predictors are used and their near collinearity can affect the analysis.

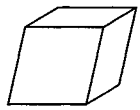
An ill-conditioned regression problem with more than two predictors is more difficult to visualize, although the essential problems are the same as in the two-predictor case. First, the regression space $\mathcal{V}_{\mathrm{X}}$ is poorly located and its relationship to the outcome $\vec{\mathbf{y}}$ is unstable. Second, within that space, accidental variation in the position of the vectors leads to large changes in the regression coefficients. Just as ill conditioning in the bivariate case does not implicate either predictor alone, ill conditioning when $p > 2$ does not require the near collinearity of any pair of vectors. It is enough that the full set almost falls into a subspace with dimension less than p. The regression space shown in Figure 4.4 exemplifies the problem in three dimensions. Here the three vectors are nearly planar, and do not comfortably span three-dimensional space. Small changes in the vectors can make large changes in the relationship of $\hat{\vec{y}}$ to the $\vec{x}_1$-$\vec{x}_2$ plane and can largely eliminate these variables in favor of $\vec{x}_3$. In this configuration, no single vector or pair of vectors can be identified as the culprit, for each pair of vectors is well conditioned. The conditioning problem does not emerge until all three predictors are used. In consequence, one cannot recognize ill conditioning by pairwise examination of the variables, such as by looking either at a table of bivariate correlation coefficients or at the angles between them. The difficulties only appear when the full set is examined.

To measure how well a problem is conditioned, one needs a way to determine how nearly the p-dimensional regression space $\mathcal{V}_X$ looks like a space with fewer than p dimensions. Evidence for ill conditioning can be obtained from several sources. One approach is based on regression. For each variable $\vec{x}_j$ construct its predictor $\hat{\vec{x}}_j$ based all the other predictor vectors. If the multiple correlation is very near one, then $\vec{x}_j$ almost falls into the space spanned by the other vectors, and can nearly be written as a linear combination of them. For example, in Figure 4.4 the vector $\vec{x}_3$ is well predicted by $\vec{x}_1$ and $\vec{x}_2$, indicating trouble. With a set of p predictor vectors, one runs p such tests and looks for suspiciously good fits.

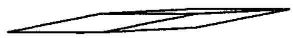
The difficulty with this approach is that it does not yield a single measure of ill conditioning. A more comprehensive approach uses the generalized volume measure described in Section 2.1. The predictor variables have a large volume when the problem is well conditioned and a small volume when it is poorly conditioned. Because the direction of the vectors determines their conditioning, not their lengths, one starts by standardizing the problem. Each predictor $\vec{x}_j$ is converted to a variable of unit length, $\vec{u}_j = \vec{x}_j/|\vec{x}_j|$. The generalized volume $\mathrm{vol}(\vec{u}_1, \vec{u}_2, \ldots, \vec{u}_p)$ defined by these vectors is near one when the problem is well conditioned and near zero when it is badly conditioned. The sensitivity of the generalized volume to ill conditioning is easily seen by comparing a well conditioned pair of vectors to a poorly conditioned pair in two dimensions:

 or 

In three dimensions well conditioned vectors have a cube-like structure:



Poorly conditioned vectors are wafer-like or stretched out like a stick:



The volume of both these configurations is near zero. In higher dimensions, the volumes of the multivariate parallelopiped have the same properties.

A third approach to locating near multicollinearity involves reexpressing $\mathcal{V}_{X}$ with a set of ordered orthogonal vectors of varying length. This technique is known as principal-component analysis and is discussed in Chapter 9.

Ill-conditioned regression problems of this type often result from an attempt to include too many predictors in the regression equation, and are particularly likely when the members of a constellation of variables have roughly the same meaning. A researcher faced with ill conditioning has several options. In principle, the problem can be solved by collecting more data, thereby decreasing the sampling variability to the point where the solutions are stable. Usually this approach is impractical, however, because even a moderately ill-conditioned problem may need millions of observations to provide adequate estimates. Even could such a large sample be obtained, it would be wasteful of resources. A more feasible approach is to simplify the problem so that it becomes well defined. Essentially, one wants to abandon the poorly defined dimensions of $\mathcal{V}_{X}$ and concentrate on its better-defined aspects. In the right-hand panel of Figure 5.2, one would concentrate on the common aspects of $\vec{x}_1$ and $\vec{x}_2$, and ignore their difference. Several courses of action are possible, the first two of which are also antidotes for complete multicollinearity.

- Eliminate variables from the problem until it is better conditioned. The selection of variables should be based on outside information, such as their interpretation, not on an analysis of the observed relationships among the variables.

- Agglomerate related variables into a single composite variable. If three variables measure roughly the same thing, a more stable analysis results when they are combined. Although the differences among the component variables are lost, the loss is not real, since the relationship of this aspect of the predictors to the outcome vector is too unstable to be measured. The principle-component techniques to be discussed in Chapter 9 are helpful here.

- Adopt selection criteria for $\vec{\hat{y}}$ that replace or are in addition to the minimization of $|\vec{e}|$. An appropriate choice of criteria can stabilize the position of the regression space and provide consistent regression coefficients. Procedures such as *ridge regression* do this, but are too involved to discuss here.

These solutions all require reducing the scope of the problem or adding information from outside the original collection of data.

5.3 Orthogonal predictors

In contrast to the unstable solutions that result from nearly multicollinear predictors, the configuration when the predictors are orthogonal is maximally stable. If $\vec{x}_1$ is orthogonal to $\vec{x}_2$, then the components of $\vec{y}$ in the $\vec{x}_1$ and $\vec{x}_2$ directions are unrelated to each other. Regression effects involving these vectors are separate and the presence or absence of one predictor in the regression equation does not affect the coefficients of the other predictor. More formally, the vectors $\vec{x}_1$ and $\vec{x}_2$ span two orthogonal subspaces, and the projection of $\vec{y}$ onto one space is unrelated to its projection onto the other space.

Figure 5.3 illustrates the configuration with two orthogonal predictors, $\vec{x}_1$ and $\vec{x}_2$, a situation that generalizes directly to any number of predictors. If $\vec{x}_1$ alone is used to predict $\vec{y}$, a regression vector $\vec{\hat{y}}_1$ is obtained with coefficient b_1. Likewise, predictors based only on $\vec{x}_2$ give the regression vector $\vec{\hat{y}}_2 = b_2\vec{x}_2$. The components of $\vec{y}$ along these orthogonal predictors are unrelated, and the magnitude of one does not depend on the presence or absence of the other. As the figure shows, when both predictors are used, the component of $\vec{\hat{y}}$ in the $\vec{x}_1$ direction is the same as it is when only $\vec{x}_1$ is used, and the two-dimensional prediction $\vec{\hat{y}}$ is the vector sum of the two one-dimensional predictions $\vec{\hat{y}}_1$ and $\vec{\hat{y}}_2$. The coefficients of the regression equation $\vec{\hat{y}} = b_1\vec{x}_1 + b_2\vec{x}_2$ are the same as the single-variable regression coefficients. The components along each predictor can be estimated just as though the other variable were not there. This simplicity complements the confusing interaction of ill-conditioned predictors described in the last section.

A more general way to view the situation uses subspaces. With orthogonal predictors, the subspace $\mathcal{V}_1$ spanned by $\vec{x}_1$ is orthogonal to the subspace $\mathcal{V}_2$ spanned by $\vec{x}_2$. In particular, the orthogonal complement of $\mathcal{V}_1$ in the regression space $\mathcal{V}_X$ is the space $\mathcal{V}_2$. In consequence, the portion of $\vec{y}$ that projects onto $\mathcal{V}_1$ is unrelated to the portion that projects onto $\mathcal{V}_2$ and can be found separately. This idea generalizes to sets of predictors. In

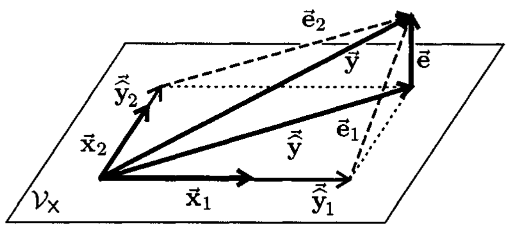


Figure 5.3: *Regression with two orthogonal predictors. The regression plane contains the two predictors $\vec{x}_1$ and $\vec{x}_2$ and the three regression vectors $\vec{\hat{y}}_1$, $\vec{\hat{y}}_2$, and $\vec{\hat{y}}$. The single-predictor errors $\vec{e}_1$ and $\vec{e}_2$ are drawn with dashed lines.*

a five-predictor regression, if $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ are orthogonal to $\vec{x}_4$ and $\vec{x}_5$, but not necessarily to each other, then the space $\mathcal{V}_{123}$ spanned by the first three vectors is orthogonal to the space $\mathcal{V}_{45}$ spanned by the latter two. The full five-variable regression problem splits into the three-variable regression that gives $\vec{\hat{y}}_{123} \in \mathcal{V}_{123}$ and the two-variable regression that gives $\vec{\hat{y}}_{45} \in \mathcal{V}_{45}$. The projection of $\vec{y}$ into $\mathcal{V}_X$ is the sum of the two components:

$$\vec{\hat{y}} = \vec{\hat{y}}_{123} + \vec{\hat{y}}_{45}.$$

This type of independence does not hold in general, and its lack creates the need for the conditional tests described in Chapter 7.

There are substantial advantages to working with orthogonal predictors. One advantage is computational. The estimates of the b_j are the same as if each variable were the only predictor in the equation. The separability of the estimates is easily shown analytically. The normal equations for two predictors (Equation 4.1) are

$$\begin{aligned} b_1(\vec{x}_1 \cdot \vec{x}_1) + b_2(\vec{x}_1 \cdot \vec{x}_2) &= \vec{x}_1 \cdot \vec{y}, \\ b_1(\vec{x}_1 \cdot \vec{x}_2) + b_2(\vec{x}_2 \cdot \vec{x}_2) &= \vec{x}_2 \cdot \vec{y}. \end{aligned}$$

With orthogonal predictors, $\vec{x}_1 \cdot \vec{x}_2 = 0$, so these equations reduce to

$$\begin{aligned} b_1(\vec{x}_1 \cdot \vec{x}_1) &= \vec{x}_1 \cdot \vec{y}, \\ b_2(\vec{x}_2 \cdot \vec{x}_2) &= \vec{x}_2 \cdot \vec{y}. \end{aligned}$$

These two equations are independent of each other and involve only a single predictor. Their solutions are the same as the single-predictor results

(Equation 3.4). This result also holds for any number of variables and for orthogonal sets of variables.

The second advantage is statistical. Sampling fluctuation associated with one coefficient is unrelated to that associated with the other. Just as the components of $\vec{y}$ along $\vec{x}_1$ and $\vec{x}_2$ are distinct, so also are their sampling errors. The proper representation of sampling effects is developed in Chapter 6; at this point it is only important to recognize that the projections of $\vec{y}$ onto $\vec{x}_1$ and onto $\vec{x}_2$ are onto mutually orthogonal subspaces. The statistical independence of the estimates minimizes their uncertainty and means that fewer subjects are needed to obtain good tests. With orthogonal predictors, almost the total sample is available to examine each variable. The number of subjects that is needed to obtain adequate accuracy and power is more like that of a single-variable regression than it is like that of a typical multiple-variable regression.

The orthogonality of the predictors gives an important additivity property to both the sums of squares and the variability that they explain. Consider the regression plane $\mathcal{V}_{\mathsf{X}}$ in Figure 5.3. The vectors $\vec{\hat{y}}_1$, $\vec{\hat{y}}_2$, and $\vec{\hat{y}}$ form a right triangle, so their squared lengths add:

$$|\vec{\hat{y}}|^2 = |\vec{\hat{y}}_1|^2 + |\vec{\hat{y}}_2|^2.$$

Squared distances correspond to sums of squares (Equation 2.15), making this equation equivalent to a sum-of-squares decomposition,

$$SS_{\mathsf{y}\cdot\mathsf{x}_1\mathsf{x}_2} = SS_{\mathsf{y}\cdot\mathsf{x}_1} + SS_{\mathsf{y}\cdot\mathsf{x}_2}.$$

These two relationships also hold when the variables $\vec{x}_1$ and $\vec{x}_2$ are replaced by an orthogonal pair of subspaces.

A multiple regression with several orthogonal predictors is not the same as a series of one-predictor regressions in all respects, however. In the material on statistical testing (Section 6.3) it will be seen that tests of the significance of a coefficient are performed by comparing the length of the regression vector $\vec{\hat{y}}$ to the length of the error vector $\vec{e}$. With one variable, $\vec{\hat{y}}_1$ is compared to $\vec{e}_1$ to test the hypothesis that the coefficient of $\vec{x}_1$ is zero. With two variables, the test of this hypothesis compares the same $\vec{\hat{y}}_1$ to the two-variable error $\vec{e}$ from which the systematic effects along both $\vec{x}_1$ and $\vec{x}_2$ have been removed. Again using the right-triangular relationships shown in Figure 5.3,

$$|\vec{e}_1|^2 = |\vec{e}|^2 + |b_2\vec{x}_2|^2$$

or

$$|\vec{e}|^2 = |\vec{e}_1|^2 - |b_2\vec{x}_2|^2.$$

With the component $b_2\vec{x}_2$ removed, the two-predictor error $\vec{e}$ is much shorter than the one-predictor error $\vec{e}_1$, greatly increasing the power of the test. As drawn in Figure 5.3, both predictors have large components along $\vec{y}$, and the reduction in the error vector when they are combined is substantial. In this way the introduction of an orthogonal predictor into a regression can increase the precision and reduce the sample size that one needs. This effect is exploited in the analysis of covariance in Section 8.4.

5.4 Suppressor variables

Sometimes a surprising relationship appears in multiple regression. A variable that is of little or no use as a predictor by itself becomes important when it is combined with another predictor. A variable with this property is known as a *suppressor variable.* At first, the effect of a suppressor variable is disconcerting: why should a variable that has nothing to do with the outcome variable help to predict it? The answer lies in the difference between the vectors and the subspaces that they span. The geometry makes the situation clear.

Figure 5.4 shows the regression of Y onto X_1 and X_2, where X_2 acts as a suppressor. The vectors $\vec{x}_1$ and $\vec{y}$ are not orthogonal, so that the regression $\hat{\vec{y}}_1$ of $\vec{y}$ on $\vec{x}_1$ in the space $\mathcal{V}_{X_1}$ is moderately effective. Nevertheless, it leaves a substantial portion $\vec{e}_1$ of $\vec{y}$ unexplained. The vector $\vec{x}_2$ is orthogonal to $\vec{y}$ and a single-variable regression using it as a predictor is valueless. However, $\vec{x}_2$ does point in a direction that is related to the direction in which $\vec{x}_1$ deviates from $\vec{y}$. The projection onto the regression space $\mathcal{V}_X$ spanned by both vectors shows this configuration (Figure 5.4, right panel). Including $\vec{x}_2$ in the regression allows it to eliminate (i.e., "suppress") the defects of $\vec{x}_1$ as a predictor of $\vec{y}$ and improve the prediction. Including $\vec{x}_2$ pulls $\hat{\vec{y}}$ around so that it, like $\vec{y}$, is orthogonal to $\vec{x}_2$—the normal equation for this variable implies that $\vec{x}_2 \cdot \hat{\vec{y}} = \vec{x}_2 \cdot \vec{y} = 0$ (Equation 4.3). Together $\vec{x}_1$ and $\vec{x}_2$ create a regression plane $\mathcal{V}_X$ that passes much closer to $\vec{y}$ than does the single-predictor space $\mathcal{V}_{X_1}$. Remember that the overall fit is determined by the position of $\mathcal{V}_X$, not by the orientations of the individual vectors to $\vec{y}$. In the illustration, almost all the error is eliminated so that the two-predictor $\hat{\vec{y}}$ is a good representation of $\vec{y}$.

The picture in Figure 5.4 has been drawn to show the suppressor structure particularly clearly, both by making $\vec{x}_2$ completely orthogonal to $\vec{y}$ and by giving it an appreciable association to $\vec{e}_1$. In practice, suppressor relationships are rarely so clean. Typically, the criterion variable and a suppressor-like variable have some correlation. However, what really characterizes suppressor activity is the greater effectiveness of a variable in the

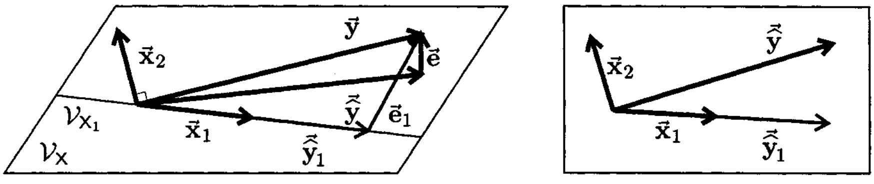


Figure 5.4: *Regression with a suppressor variable. The left panel shows the three-dimensional picture and the right panel shows the vectors that lie in the plane $\mathcal{V}_X$.*

presence of another predictor than in isolation. One often finds a variable for which the amount by which $SS_{\text{regression}}$ increases (i.e., $\hat{\vec{y}}$ becomes longer) when the variable is introduced in a regression that already includes several variables is greater than $SS_{\text{regression}}$ for the variable alone. Such variables have a suppressor quality.

The configuration of variables in a suppressor relationship is much easier to interpret using vector geometry than it is using some other common representations. Suppressor relationships are hard to see in the configuration of points in a variable-space scatterplot such as Figure 4.1, particularly with more than two predictors. Another frequently used representation is even worse. One cannot represent suppressor activity with the false metaphor of correlation as area, often presented using Venn diagrams to represent shared variability. Those diagrams provide no way to have the "overlap" of two variables increase when a third variable is added to the equation, as they do in a suppressor relationship.

The geometry of suppressor activity again points up a crucial aspect of multivariate interpretation. The regression effects depend on the spaces spanned by the sets of vectors, not on the individual variables acting alone. The space $\mathcal{V}_X$ determines the location of $\hat{\vec{y}}$ and the quality of the prediction, not the placement of any $\vec{x}_j$ by itself. One avoids paradoxes only by thinking of multivariate relationships in terms of the sets of variables and the spaces they span. Acquiring the ability to think in these terms is one of the most difficult parts of learning multivariate statistics. Here the geometry is essential.

Exercises

1. Suppose that the three predictors $\vec{x}_1$, $\vec{x}_2$, and $\vec{x}_3$ in multiple re-

gression are multicollinear, with $\dim(\mathcal{V}_X) = 2$. Can one still calculate the loadings of the variables on the regression prediction? Illustrate geometrically.

2. Section 5.1 described an example of collinear vectors involving three categories of agreement, X_1, X_2 and X_3, and an overall proportion of agreement X_a. Suppose that the researcher decided to measure overall agreement by counting the proportion of a different set of statements with which the respondent agrees—not the statements used to find X_1, X_2, and X_3. Would this change in the design solve the problem of multicollinearity? What new problems might it introduce into the analysis?

3. Consider two sets of data with the correlation matrices

$$\begin{bmatrix} 1.00 & -0.23 & 0.44 \\ -0.23 & 1.00 & 0.66 \\ 0.44 & 0.66 & 1.00 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1.00 & 0.96 & 0.44 \\ 0.96 & 1.00 & 0.66 \\ 0.44 & 0.66 & 1.00 \end{bmatrix}.$$

For each set, regress the third variable on the other two, and find R^2. Now suppose that the correlation between the second and third variable is changed from 0.66 to 0.60 in each analysis. How does R^2 change? What accounts for the difference between the two examples?

4. Calculate the generalized volume defined by the predictors in the two examples of Problem 3.

5. For which (if any) of the three configurations in Problem 4.1 does $R^2_{Y \cdot XZ} = R^2_{Y \cdot X} + R^2_{Y \cdot Z}$?

6. For which (if any) of the three configurations in Problem 4.1 does Z show suppressor activity? Explain both numerically and geometrically.

7. Section 5.3 described how a predictor orthogonal to $\vec{x}_1$ can improve the test of the X_1-Y relationship, and Section 5.4 described how a predictor orthogonal to $\vec{y}$ can also be useful. Can a predictor serve both functions at once—i.e., can a predictor that is orthogonal to both $\vec{x}_1$ and $\vec{y}$ have any value in a multiple regression analysis? Illustrate with pictures.

Chapter 6

Statistical tests

Fitting a model to a set of data does not ordinarily complete the analysis. One usually wants to establish which effects are likely to be real and which are accidents of sampling. Central to this endeavor are the hypothesis tests, which help one decide whether a particular effect stands out about the accidental fluctuation of the scores.[1] Most of the statistical testing procedures used in multivariate statistics can be formulated geometrically. This chapter describes this geometry.

6.1 The effect space and the error space

In its essentials, the statistical analysis of a multiple regression works by dividing subject space into two orthogonal subspaces, one of which contains the systematic effects of the regression and the other only the random effects of sampling error. Where the systematic subspace is multidimensional, it may be further divided into individual effects. An effect is deemed statistically significant when the component of $\hat{\mathbf{y}}$ falling into a systematic subspace is sufficiently much larger than the component falling into the error subspace. Developing these spaces requires a fuller consideration of the dimensional structure of the data than is necessary to estimate the regression predictor.

Begin by reviewing how a variable is represented by a vector in subject space. Each subject's scores are plotted on a separate axis, with the score from the first subject on the first axis, that from the second subject on the second axis, and so on. So the variable Y with N observations Y_1, Y_2, ...,

[1] Statistical techniques also are used to construct confidence intervals for parameters and scores, but those techniques are not discussed here.

Y_N is described by a vector from the origin to the point $(Y_1, Y_2, \ldots, Y_N)$. The space $\mathcal{V}_{\text{total}}$ containing this vector has N axes and is N-dimensional. All variation of both the centered and the uncentered vectors occurs in this space.

In the discussion of multiple regression to this point, it has only been necessary to consider that portion of subject space that actually contains vectors. When there are two variables X and Y, there are two vectors $\vec{\mathbf{x}}$ and $\vec{\mathbf{y}}$, and the regression problem can be solved in two-dimensional space. The remaining $N-2$ empty dimensions of subject space are irrelevant. For statistical testing, however, the entire space $\mathcal{V}_{\text{total}}$ must be considered. Essentially, one needs to know both where $\vec{\mathbf{y}}$ actually lies and where it might lie in another replication of the study.

The structure of a regression problem divides the N-dimensional space $\mathcal{V}_{\text{total}}$ into three orthogonal subspaces, $\mathcal{V}_1$, $\mathcal{V}_{\mathsf{X}}$, and $\mathcal{V}_{\mathsf{e}}$. The projection of the uncentered outcome vector $\vec{\mathbf{Y}}$ onto these three spaces breaks it into three components. Two distinctions determine this division, one between the centered and the centering components, the other between the parts of the space that can be reached by the regression vectors and the parts that cannot. First, as described in Section 3.3, the space can be separated into the components lying along the vector $\vec{\mathbf{1}}$ and those orthogonal to this vector. This decomposition divides $\mathcal{V}_{\text{total}}$ into the one-dimensional subspace $\mathcal{V}_1$ that contains the regression intercept and the $(N-1)$-dimensional subspace $\mathcal{V}_{\perp 1}$. Variation in the first of these subspaces concerns the regression constant, and variation in the second subspace concerns both the regression effects and the random variability of the outcome vector. Most regression questions are concerned with the centered effects occurring in $\mathcal{V}_{\perp 1}$, and to emphasize its importance that space is denoted simply by $\mathcal{V}$ below.

The other, and more important, part of the decomposition divides $\mathcal{V}$ into two orthogonal subspaces. One of these is the *regression space* or *effect space*, which is generated by the centered regression vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, $\ldots$, $\vec{\mathbf{x}}_p$. Denote this p-dimensional space by $\mathcal{V}_{\mathsf{X}}$. This space, which was discussed in the earlier chapters, contains both the predictors $\vec{\mathbf{x}}_j$ and the prediction $\hat{\vec{\mathbf{y}}}$. The orthogonal complement of $\mathcal{V}_{\mathsf{X}}$ in $\mathcal{V}$ is the *error space.* This multidimensional space contains the error vector $\vec{\mathbf{e}} = \vec{\mathbf{y}} - \hat{\vec{\mathbf{y}}}$. Since $\dim(\mathcal{V}) = N-1$ and $\dim(\mathcal{V}_{\mathsf{X}}) = p$, the orthogonal complement has dimension

$$\dim(\mathcal{V}_{\mathsf{e}}) = \dim(\mathcal{V}) - \dim(\mathcal{V}_{\mathsf{X}}) = N - p - 1.$$

The space $\mathcal{V}_{\mathsf{e}}$ contains everything that cannot be expressed by the regression effects in $\mathcal{V}_1$ and $\mathcal{V}_{\mathsf{X}}$. It represents the dimensions of variability available to the error.

Ordinary three-dimensional space is inadequate to picture the effect space and the error space simultaneously in as much generality as one would like. Figure 6.1 shows two representations of these spaces. In both illustrations the large box-shaped form lies in the overall space $\mathcal{V}$ and should be seen as three dimensional. In the top panel the effect space is two dimensional, drawn as a plane cutting through the box. Orthogonal to it is a one-dimensional error space. The familiar configuration of multiple regression vectors appears at the origin. The vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, and $\vec{\hat{\mathbf{y}}}$ lie in the plane $\mathcal{V}_{\mathsf{X}}$. The vector $\vec{\mathbf{e}}$ lies entirely in $\mathcal{V}_{\mathsf{e}}$, and $\vec{\mathbf{y}}$ has components in both spaces. Unlike in the diagrams used in the previous chapters, $\vec{\mathbf{e}}$ is drawn from the origin to show clearly that it is part of the error space. The dashed lines show the orthogonal decomposition of $\vec{\mathbf{y}}$ into the components $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$.

The top panel of Figure 6.1 is misleading in that it shows $\mathcal{V}_{\mathsf{e}}$ as unidimensional. In reality, $\dim(\mathcal{V}_{\mathsf{e}}) = N-p-1$, typically a large number. The lower panel of Figure 6.1 shows another approximation to the true picture, this time one in which $\mathcal{V}_{\mathsf{X}}$ is one dimensional and $\mathcal{V}_{\mathsf{e}}$ is two dimensional. Again the regression vectors are shown. In this picture, $\mathcal{V}_{\mathsf{e}}$ has an unused dimension that does not contain $\vec{\mathbf{e}}$. Of course in any real problem, $\dim(\mathcal{V}_{\mathsf{e}})$ is much greater than either of these pictures can show, and the unused portion of $\mathcal{V}_{\mathsf{e}}$ is very large. To properly visualize the testing situation one must keep in mind the multidimensional aspects of both parts of Figure 6.1 simultaneously, allowing $\mathcal{V}_{\mathsf{e}}$ to be multidimensional, as in the lower picture, but recognizing that $\mathcal{V}_{\mathsf{X}}$ is p-dimensional, as in the upper picture.

The decomposition into three spaces divides the uncentered outcome vector $\vec{\mathbf{Y}}$ into three pieces:

$$\vec{\mathbf{Y}} = \overline{Y}\vec{\mathbf{1}} + \vec{\hat{\mathbf{y}}} + \vec{\mathbf{e}}, \tag{6.1}$$

the three terms being in $\mathcal{V}_1$, $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{e}}$, respectively. The three terms can be grouped in two sensible ways. In one grouping, the first two terms are the systematic components of the regression and the last term is the random portion; in the other, the first term expresses the constant and the last two terms are the centered portion of the regression.

The dimension of each space determines how freely a vector can vary within it. In a low-dimensional space a vector is confined, while in a high-dimensional space it has many more possible positions. Vector $\vec{\mathbf{e}}$ is more confined by the one-dimensional error space in the upper panel of Figure 6.1 than it is by the two-dimensional space in the lower panel, and, of course, it is still less confined by the high-dimensional space in a realistic problem. In more familiar terms, the dimension of the space that contains a statistical vector is known as the *degrees of freedom* (abbreviated *df*) of the corresponding variable. The overall space of uncentered observations is N-dimensional, so

$$df_{\text{total}} = \dim(\mathcal{V}_{\text{total}}) = N.$$

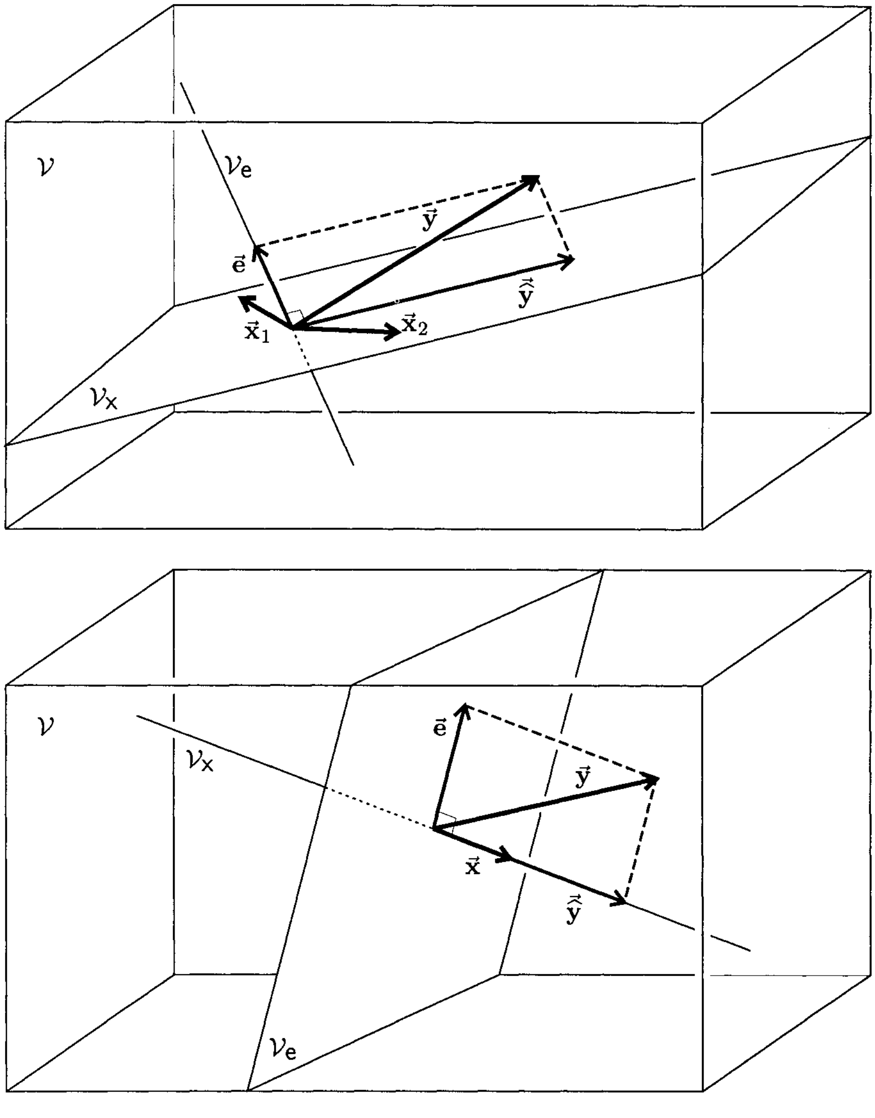


Figure 6.1: *Two illustrations of the effect space and the error space. At the top the effect space $\mathcal{V}_X$ is two dimensional, at the bottom it is one dimensional. In both diagrams the space $\mathcal{V}$ is drawn as three dimensional.*

The dimensions of the three subspaces containing the constant, the centered regression predictors, and the error vector, expressed as degrees of freedom, are:

$$\begin{aligned} df_{\text{const}} &= \dim(\mathcal{V}_1) = 1, \\ df_{\text{regression}} &= \dim(\mathcal{V}_\mathsf{X}) = p, \\ df_{\text{error}} &= \dim(\mathcal{V}_\mathsf{e}) = N-p-1. \end{aligned} \tag{6.2}$$

Because the decomposition into these spaces is orthogonal, the dimensions add:

$$df_{\text{const}} + df_{\text{regression}} + df_{\text{error}} = 1 + p + (N-p-1) = N = df_{\text{total}}.$$

The discussion above treated the predictor vectors as linearly independent, so that they span a full space and $\dim(\mathcal{V}_\mathsf{X}) = p$. If dependencies exist among the scores, then this dimension is reduced. With a single linear dependence among the predictors, $\dim(\mathcal{V}_\mathsf{X})$ drops by 1 to $p - 1$, with a consequent reduction in the effect degrees of freedom. The dimensions of $\mathcal{V}_{\text{total}}$ and $\mathcal{V}_1$ do not change, so there is a concomitant increase in $\dim(\mathcal{V}_\mathsf{e})$ to $N-p$. Where ambiguities arise, the degrees of freedom are determined by the dimensions of the spanned spaces, not by the number of vectors that are used to construct them.

6.2 The population regression model

Statistical tests in multiple regression and other multivariate procedures can be completely developed from the geometry. Even the form of the F distribution used to find critical values can be found from geometric considerations, although this book only hints at how this is done.

A statistical testing procedure lets one make decisions about whether an effect is present in the hypothetical population from which the observed data have been sampled. To construct these decision procedures, one needs a description of the underlying population, both of the systematic effects common to every observation and of the random error that makes the observations differ. This section describes the population model for regression, and the following section applies that model to the statistical tests.

The conventional model for multiple regression combines a fixed regression effect with a random error component. The underlying true relationship between the predictors and the outcome is assumed to be linear, with population regression coefficients α and β_1, β_2, ..., β_p. Under this model, the theoretical expected value of the score Y_i (i.e., its mean in the population) for a particular combination of predictors is a linear function, either

$$\mathsf{E}(Y_i) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip},$$

or, using the centered parameters and the population mean μ_{Y} of Y,

$$\mathsf{E}(Y_i) = \mu_{\mathsf{Y}} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}.$$

To this perfect linear relationship is added a random component, representing the uncertainties of sampling error. In the usual regression sampling model, the individual scores are distributed about $\mathsf{E}(Y_i)$ according to a normal distribution with a mean of zero and a variance of σ^2, with the errors for different subjects being independent. Let ε_i be a random variable with this normal distribution. The value of a particular score is a random variable composed of the population regression and this random variability,

$$Y_i = \mu_{\mathsf{Y}} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i. \tag{6.3}$$

This model for the individual observations translates directly into a vector model:

$$\vec{\mathbf{Y}} = \mu_{\mathsf{Y}}\vec{\mathbf{1}} + \beta_1\vec{\mathbf{x}}_1 + \beta_2\vec{\mathbf{x}}_2 + \cdots + \beta_p\vec{\mathbf{x}}_p + \vec{\varepsilon}. \tag{6.4}$$

This equation has both a systematic part and a random part. The systematic part is the population mean vector:

$$\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}} = \mu_{\mathsf{Y}}\vec{\mathbf{1}} + \beta_1\vec{\mathbf{x}}_1 + \beta_2\vec{\mathbf{x}}_2 + \cdots + \beta_p\vec{\mathbf{x}}_p.$$

The random components are consolidated in the error vector $\vec{\varepsilon}$. Figure 6.2 shows these vectors, with $\vec{\mathbf{Y}}$ as the sum of the vectors $\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}}$ and $\vec{\varepsilon}$. The picture is like that of regression in the sample (e.g., Figure 4.2), but differs in that $\vec{\varepsilon}$ need not be orthogonal to $\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}}$. Some components of the error may fortuitously mimic part of the regression effects.

In the regression model, the vector $\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}}$, being formed from $\vec{\mathbf{1}}$ and the predictor vectors, is fixed in position, while the vector $\vec{\varepsilon}$ is random, both in direction and length. The random vector $\vec{\varepsilon}$ is formed from N independent random components, $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$, lying along the N axes of subject space. Each of these components has an independent normal distribution with a mean of zero. This random structure gives $\vec{\varepsilon}$ a distribution with no preferred direction. It is equally likely to point in any direction within the N-dimensional space $\mathcal{V}_{\text{total}}$. In Figure 6.2 this spherically-symmetric distribution is indicated by drawing a circle about the tip of $\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}}$. When looking at this picture, one should remember that the length of $\vec{\varepsilon}$ is not fixed. Reflecting the distribution of the components ε_i that make it up, the distribution of the tip of $\vec{\varepsilon}$ is densest near the end of $\vec{\mu}_{\mathsf{Y}\cdot\mathsf{X}}$ and falls off away from the center according to a normal distribution.

The normal distribution is essential for $\vec{\varepsilon}$ to have this direction-free character. The standard normal density function is

$$f(\varepsilon) = \frac{1}{\sqrt{2\pi}} e^{-\varepsilon^2/2}.$$

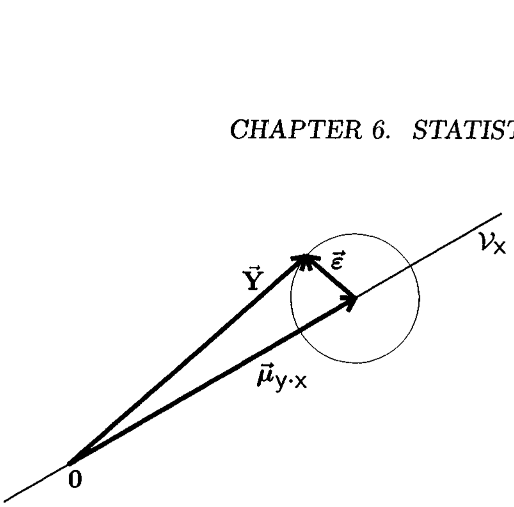


Figure 6.2: *The population model for multiple regression. The random vector $\vec{\varepsilon}$ points in any direction from the tip of the systematic component $\vec{\mu}_{Y\cdot X}$. The circle symbolizes the symmetrical distribution of positions potentially taken by $\vec{\varepsilon}$.*

When N independent normal distributions are combined to form $\vec{\varepsilon}$, the multivariate density is the product

$$\begin{aligned} f(\vec{\varepsilon}) &= \left[\frac{1}{\sqrt{2\pi}}e^{-\varepsilon_1^2/2}\right]\left[\frac{1}{\sqrt{2\pi}}e^{-\varepsilon_2^2/2}\right]\cdots\left[\frac{1}{\sqrt{2\pi}}e^{-\varepsilon_N^2/2}\right] \\ &= \frac{1}{(2\pi)^{N/2}}\exp\left[\frac{-(\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_N^2)}{2}\right]. \end{aligned}$$

The sum of squares in the exponential is the squared length of $\vec{\varepsilon}$, so that

$$f(\vec{\varepsilon}) = \frac{1}{(2\pi)^{N/2}}\exp(-|\vec{\varepsilon}|^2/2). \tag{6.5}$$

This derivation shows that the density of $\vec{\varepsilon}$ is a function of its length, but is independent of its angle. Its distribution is uniformly distributed over the angle, and the vector $\vec{\varepsilon}$ is equally likely to point in any direction. Distributions of the individual errors ε_j other than the normal do not give a spherically symmetrical distribution for $\vec{\varepsilon}$.

6.3 Testing the regression effects

The hypothesis tests for regression are based on the background developed in the last sections. Consider the question of whether the predictor variables X_1, X_2, ..., X_p help to predict Y. The population regression model

(Equation 6.4) is

$$\vec{\mathbf{Y}} = \mu_{\mathsf{Y}}\vec{\mathbf{1}} + \beta_1\vec{\mathbf{x}}_1 + \beta_2\vec{\mathbf{x}}_2 + \cdots + \beta_p\vec{\mathbf{x}}_p + \vec{\boldsymbol{\varepsilon}} = \mu_{\mathsf{Y}}\vec{\mathbf{1}} + \vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}} + \vec{\boldsymbol{\varepsilon}}.$$

To find out whether the regression is helpful, one tests the null hypothesis that $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}} = \vec{\mathbf{0}}$ or, more specifically, that the population regression coefficients are zero:

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0.$$

Rejection of this hypothesis in favor of the alternative that one or more of the β_j are nonzero means that the X_j and Y are related.

Figure 6.3 shows the configuration when H_0 is true on the left and when it is false on the right. For simplicity the component along $\vec{\mathbf{1}}$ is ignored, and only the centered spaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{e}}$ are shown. When H_0 is true, the population model is simply $\vec{\mathbf{y}} = \vec{\boldsymbol{\varepsilon}}$. The variability is centered about the origin $\mathbf{0}$. When H_0 is false, the center of the variability is displaced away from $\mathbf{0}$ along the vector $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}} \in \mathcal{V}_{\mathsf{X}}$. The same error vector $\vec{\boldsymbol{\varepsilon}}$ appears in both panels, but the systematic part $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}}$ differs. If one could measure the component $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}}$ directly, then the testing problem could be solved by checking whether its length was nonzero. However, these vectors are part of the theoretical model and cannot be resolved empirically. Instead, the observed vector $\vec{\mathbf{y}}$ is decomposed into $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$, as shown in both panels of the figure. Because $\vec{\boldsymbol{\varepsilon}}$ almost certainly has some component in $\mathcal{V}_{\mathsf{X}}$, this breakdown differs from the theoretical one. The vector $\vec{\hat{\mathbf{y}}}$ is unlikely to equal $\vec{\mathbf{0}}$, even when $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}} = \vec{\mathbf{0}}$.

A comparison of the two panels of Figure 6.3 shows two effects of a false null hypothesis. First, it changes the relative lengths of the vectors $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$ that make up the empirical decomposition of $\vec{\mathbf{y}}$. Although the length of the component $\vec{\mathbf{e}}$ in the error space is approximately the same size in both pictures, the component $\vec{\hat{\mathbf{y}}}$ of $\vec{\mathbf{y}}$ in the effect space is much longer when the null hypothesis is false. Second, as the error variability is displaced away from $\mathbf{0}$ by a nonzero $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}}$, the angle between $\vec{\mathbf{y}}$ and $\vec{\hat{\mathbf{y}}}$ becomes smaller. The idea behind the statistical tests for regression now is straightforward. First one picks a measure of the effect, either the angle between $\vec{\mathbf{y}}$ and $\vec{\hat{\mathbf{y}}}$ or the relative lengths of the vectors $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$. The random orientation of $\vec{\boldsymbol{\varepsilon}}$ under the null hypothesis lets one derive the distribution of these statistics. The null hypothesis is rejected whenever a configuration occurs that would be improbable if $\vec{\mathbf{y}}$ were randomly oriented with respect to $\mathcal{V}_{\mathsf{X}}$.

Of the two representations of effect, length and angle, length is the more often used. When the hypothesis that $\vec{\boldsymbol{\mu}}_{\mathsf{Y}\cdot\mathsf{X}} = \vec{\mathbf{0}}$ is true, the vectors $\vec{\hat{\mathbf{y}}}$ and $\vec{\mathbf{e}}$ are projections of the randomly oriented vector $\vec{\boldsymbol{\varepsilon}}$ into the spaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{e}}$. One would expect their lengths to be in some sense comparable.

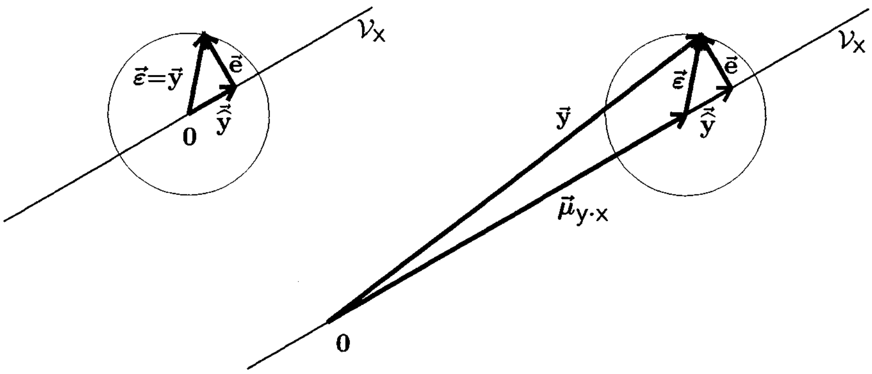
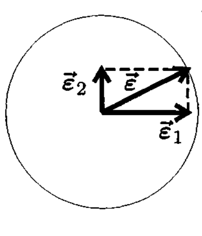


Figure 6.3: *The population model for multiple regression when the null hypothesis of no regression effect is true (left) and false (right). Only one dimension of the regression space $\mathcal{V}_X$ is shown, and the subspace $\mathcal{V}_1$ is omitted.*

In contrast, when the hypothesis is false, the vector $\hat{\vec{y}}$ contains a nonrandom systematic component and should be long relative to $\vec{e}$. The lengths cannot be compared directly, however, because $\hat{\vec{y}}$ and $\vec{e}$ reside in spaces with different dimensions. To find the proper adjustment for the differing dimensions, a short digression on the length of a random vector is needed.

Consider a random vector $\vec{\varepsilon}$ in a d-dimensional vector space $\mathcal{V}$ and let $\vec{\varepsilon}_j$ be the component of $\vec{\varepsilon}$ projected onto the jth axis of $\mathcal{V}$; for example, in two dimensions the projections are



By the Pythagorean theorem, the squared length of $\vec{\varepsilon}$ is the sum of the squares of the lengths parallel to each axis:

$$|\vec{\varepsilon}|^2 = |\vec{\varepsilon}_1|^2 + |\vec{\varepsilon}_2|^2 + \cdots + |\vec{\varepsilon}_d|^2.$$

The expected value of a sum equals the sum of the expected values, so that the average squared length of $\vec{\varepsilon}$ is a sum of components from each dimension:

$$\mathsf{E}(|\vec{\varepsilon}|^2) = \mathsf{E}(|\vec{\varepsilon}_1|^2) + \mathsf{E}(|\vec{\varepsilon}_2|^2) + \cdots + \mathsf{E}(|\vec{\varepsilon}_d|^2).$$

For a random vector with no preferred direction, the projection of $\vec{\varepsilon}$ along each axis has the same average length. As a result, the expected squared length of $\vec{\varepsilon}$ is proportional to the dimension d of the space—if the average squared length of a component is M, then

$$\mathsf{E}(|\vec{\varepsilon}|^2) = dM.$$

The same principle applies to subspaces. In a subspace of dimension d_1, the component of $\vec{\varepsilon}$ lying along each orthogonal dimension has the same average length. The expected squared length of the projection of $\vec{\varepsilon}$ onto the subspace is d_1M. It is proportional to the dimension of the subspace and is unrelated to other characteristics of that subspace such as how it is generated.

The squared lengths of random vectors from different subspaces can be compared after each quantity is divided by the dimension of the subspace in which the vector lies. Comparisons are made using the per-dimension squared length, which, for the vector $\vec{\varepsilon} \in \mathcal{V}$, is denoted here by

$$M(\vec{\varepsilon}) = \frac{|\vec{\varepsilon}|^2}{\dim(\mathcal{V})}. \tag{6.6}$$

In the conventional analysis-of-variance terminology, the length of a vector corresponds to a sum of squares and the dimension of the space corresponds to the degrees of freedom, so $M(\vec{\varepsilon})$ is the *mean square.*

Returning to the regression problem, the two mean squares needed for a test of the hypothesis that $\vec{\mu}_{\mathsf{y}\cdot\mathsf{x}} = \vec{0}$ are those of the regression vector and the error vector:

$$M(\vec{\hat{\mathbf{y}}}) = \frac{|\vec{\hat{\mathbf{y}}}|^2}{\dim(\mathcal{V}_{\mathsf{X}})} \qquad \text{and} \qquad M(\vec{\mathbf{e}}) = \frac{|\vec{\mathbf{e}}|^2}{\dim(\mathcal{V}_{\mathsf{e}})}.$$

A comparison of these lengths tests the null hypothesis of no regression effect. When $\vec{\mu}_{\mathsf{y}\cdot\mathsf{x}} = \vec{0}$, neither the effect space nor the error space has any special status and the projection of $\vec{\mathbf{y}}$ into them are the projections of a random vector. Their mean squares are approximately the same:

$$M(\vec{\hat{\mathbf{y}}}) \approx M(\vec{\mathbf{e}}).$$

When $\vec{\mu}_{\mathsf{y}\cdot\mathsf{x}} \neq \vec{0}$, the systematic component $\vec{\hat{\mathbf{y}}}$ tends to be longer than the purely random part $\vec{\mathbf{e}}$. Except for occasional accidents of sampling,

$$M(\vec{\hat{\mathbf{y}}}) > M(\vec{\mathbf{e}}).$$

The conventional way to compare these mean-square lengths is to look at their ratio:

$$F = \frac{M(\vec{\hat{y}})}{M(\vec{e})} = \frac{\dim(\mathcal{V}_{\mathrm{e}})|\vec{\hat{y}}|^2}{\dim(\mathcal{V}_{\mathrm{X}})|\vec{e}|^2}. \tag{6.7}$$

When the null hypothesis is true, $F \approx 1$, while when it is false, F generally exceeds 1. Thus, for a particular type I error probability[2] α, a critical value F_α is found and the null hypothesis is rejected whenever $F > F_\alpha$. Because the presence of a systematic effect (i.e., a nonnull $\vec{\mu}_{\mathrm{Y}\cdot\mathrm{X}}$) increases $M(\vec{\hat{y}})$ but not $M(\vec{e})$, a one-sided critical region is appropriate. The statistic that results is the usual F statistic familiar from tests in the analysis of variance and multiple regression. Critical values F_α are obtained from tables that appear in most statistic books. Computer packages often calculate the probability that a random choice from this distribution exceeds the value observed in a set of data—the *descriptive level* or "p value" of the result.

It is easy to rewrite Equation 6.7 using the multiple correlation coefficient. The vector lengths from Equations 4.9 and 4.10 are

$$|\vec{\hat{y}}|^2 = R^2|\vec{y}|^2 \qquad \text{and} \qquad |\vec{e}|^2 = (1 - R^2)|\vec{y}|^2.$$

Using these lengths and the degrees of freedom from Equations 6.2, the mean squares are

$$M(\vec{\hat{y}}) = \frac{R^2|\vec{y}|^2}{p} \qquad \text{and} \qquad M(\vec{e}) = \frac{(1 - R^2)|\vec{y}|^2}{N - p - 1}. \tag{6.8}$$

Their ratio is the familiar F test for multiple regression:

$$F = \frac{(N-p-1)R^2}{p(1 - R^2)}. \tag{6.9}$$

Other multiple regression tests are run in the same way in other subspaces, a fact that underlies the restricted and conditional tests described in Sections 6.4 and 7.3. The F statistic is the ratio of the mean squares in the effect and error spaces, as in Equation 6.7.

From this analysis, one might suppose that since the intercept α expresses variation in the subspace $\mathcal{V}_1$, the null hypothesis $\alpha = 0$ could be tested by comparing the observed component $a\vec{1}$ in this space to the random variability measured by $\vec{e}$. As it happens, the situation is somewhat more complicated. Recall that the intercept a is found by subtracting $b_1\overline{X}_1 + b_2\overline{X}_2 + \cdots + b_p\overline{X}_p$ from outcome mean $\overline{Y}$ (Equation 4.5). This calculation can only be made after the regression coefficients have been fitted.

[2]Do not confuse this probability with the intercept parameter α of the population regression equation.

It is influenced by the accuracy with which the b_j are estimated, which in turn depends on variability in the space $\mathcal{V}_X$ orthogonal to $\mathcal{V}_1$. The presence of the b_j increases the variability of a and adds a term to the denominator of the F ratio. For a single predictor, the test, in both geometric and algebraic terms, is

$$F_{\text{const}} = \frac{M(a\vec{\mathbf{1}})}{\left[1 + \dfrac{|\vec{\bar{\mathbf{X}}}|^2}{|\vec{\mathbf{x}}|^2}\right] M(\vec{\mathbf{e}})} = \frac{Na^2}{\left[1 + \dfrac{N\overline{X}^2}{\sum x_i^2}\right] M(\vec{\mathbf{e}})}. \tag{6.10}$$

The ratio of $|\vec{\bar{\mathbf{X}}}|$ to $|\vec{\mathbf{x}}|$ measures how far the regression line is projected away from the center of the data to reach $\mathbf{0}$, this distance being measured relative to the variability of the scores used to estimate b. Equation 6.10 can also be used to test the null hypothesis that α equals a value α_0 other than 0 by replacing a by $a - \alpha_0$.

The second way to construct a test statistic uses the angle between $\vec{\mathbf{y}}$ and $\hat{\vec{\mathbf{y}}}$, in effect, the multiple correlation coefficient $R = \cos\angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}})$. As Figure 6.3 suggests, this angle tends to be much larger when the null hypothesis is true than when that hypothesis is false. The angle itself can be used directly as a test statistic, with the null hypothesis being rejected at the α level whenever $\angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}})$ is less than some critical value θ_α. A test can also be based on the correlation by rejecting the null hypothesis whenever $R = \cos\angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}})$ is greater than a critical value R_α. The monotonic relationship between R^2 and F (Equation 6.9) means that tests using the angle and the vector-length criteria are equivalent.

The final thing that one needs to run a statistical test is the distribution of the test statistic under the null hypothesis. From this distribution are obtained the critical values F_α, θ_α, or R_α used to make a decision. Figure 6.4 shows how these critical values can be obtained from the geometry, a development worth following more for the insight it gives into the geometry of the tests than for any practical calculation. Both panels of Figure 6.4 show the vector space $\mathcal{V}$ when the null hypothesis is true, with a one-dimensional regression space $\mathcal{V}_X$. In the left panel, the error space $\mathcal{V}_e$ is one dimensional to best show the vectors in the plane of $\vec{\mathbf{x}}$ and $\vec{\mathbf{e}}$. The right panel illustrates the effect of the empty dimensions of $\mathcal{V}_e$ with a two dimensional error space—the circle should be seen as a sphere. One extends the picture to higher-dimensional error spaces by thinking of the sphere as its multidimensional generalization.

First consider the two-dimensional case on the left. Under the null hypothesis, $\vec{\mathbf{y}}$ is random and, because of the normal sampling distribution, is equally likely to point in any direction. It can be directed toward any

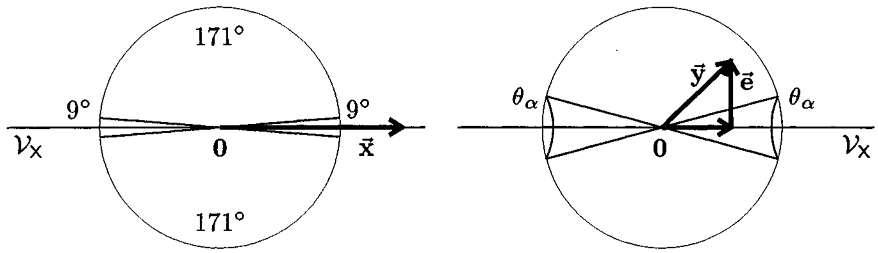


Figure 6.4: *Angular critical values for a test in multiple regression, drawn in two dimensions on the left and in more dimensions on the right. For the vector $\vec{y}$ in the right panel, the null hypothesis would be retained.*

point on the circumference the circle. In contrast, when the null hypothesis is false, $\vec{y}$ tends to point in the direction of $\mathcal{V}_X$. A decision rule for a hypothesis test is to reject the null hypothesis whenever $\vec{y}$ points at the portion of the circle that is nearest to $\mathcal{V}_X$. Suppose that the test is to be conducted at the $\alpha = 0.05$ level. Five percent of the 360° of a circle is 18°. This 18° is divided into two 9° wedges, one pointed in each direction along $\mathcal{V}_X$. Any time that $\vec{y}$ falls within either of these wedges, that is, within $\theta_\alpha = 4.5°$ of $\mathcal{V}_X$, one rejects the null hypothesis.

In the more general picture on the right, the two wedges have become cones, still pointing in opposite directions along the effect space. Whenever $\vec{y}$ is outside these cones, the null hypothesis is rejected. The critical angle θ_α that defines the cones is chosen so that each cone cuts off an area of the surface of sphere that is a proportion $\alpha/2$ of the whole surface area. When the null hypothesis is true, the probability that the random vector $\vec{y}$ falls within one of these cones is α and the probability that it falls outside the cones is $1-\alpha$. When the null-hypothesis is false, so that $\vec{y}$ lies closer to $\mathcal{V}_X$, the probability that $\angle(\vec{y}, \vec{\hat{y}})$ is less than θ_α exceeds α. In considering this picture, one needs to remember that the "sphere" resides in the $(N-1)$-dimensional space $\mathcal{V}$ and the "line" of $\mathcal{V}_X$ is a p-dimensional subspace. The principles illustrated by Figure 6.4 still hold, however.

The critical-cone argument provides (in principle) critical values for the statistic $\angle(\vec{y}, \vec{\hat{y}})$. Because this angle is monotonically related to both the multiple correlation (by the cosine) and the F statistic (by Equation 6.9), the values for one statistic can be converted to those of the others. It is possible to combine facts about the angles subtended by multidimensional cones and the surface areas of multidimensional spheres with the trigonometric relationships between the distances and angles to derive the sampling

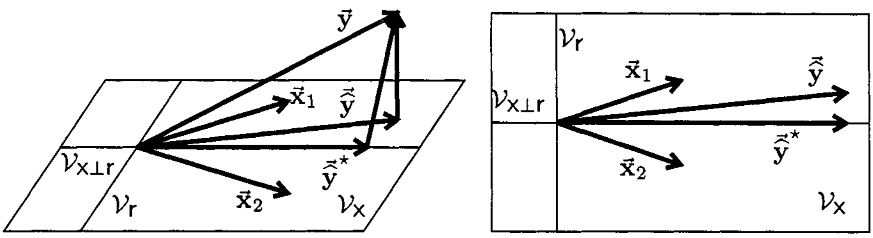


Figure 6.5: *The restriction of a bivariate regression to a one-dimensional space by the restriction $b_1 = b_2$. The left panel shows the three-dimensional configuration and the right panel shows the space $\mathcal{V}_X$.*

distributions of these statistics. However, these calculations add little to one's intuition for the procedure and are not pursued here.

6.4 Parameter restrictions

One sometimes wishes to investigate a model in which the parameters are fitted under some constraint. For example, Section 4.3 described a comparison between an unconstrained three-parameter regression model and one in which each predictor has the same slope:

$$\hat{\vec{y}} = b_1\vec{x}_1 + b_2\vec{x}_2 + b_3\vec{x}_3 \qquad \text{and} \qquad \hat{\vec{y}}^* = b\vec{x}_1 + b\vec{x}_2 + b\vec{x}_3$$

(Equation 4.11). A slightly more complicated comparison involves a decision as to whether the two regression coefficients b_2 and b_3 are the same, which suggests the model

$$\hat{\vec{y}}^* = b_1\vec{x}_1 + b_2\vec{x}_2 + b_2\vec{x}_3 \tag{6.11}$$

(note that b_2 is substituted for b_3 here). In these situations, one usually wants to know whether the restrictions are tenable. One also may want to check whether the restricted regression equation is significantly related to the outcome variable. Both questions are answered by considering how the constraints restrict the regression space.

When a set of linear restrictions is applied to the regression coefficients, it confines the regression vector to a subspace of the full regression space $\mathcal{V}_X$. Figure 6.5 shows the effect of the restriction $b_1 = b_2$ on a two-predictor regression. The panel on the left shows the familiar three-dimensional picture and the panel on the right shows the regression plane $\mathcal{V}_X$. The equal-coefficient restriction gives the regression equation the form

$$\vec{\hat{y}}^{\star} = b\vec{x}_1 + b\vec{x}_2 = b(\vec{x}_1 + \vec{x}_2).$$

If this restriction holds, then the population regression vector $\vec{\mu}_{y \cdot x}$ is collinear with $\vec{x}_1 + \vec{x}_2$ and the unconstrained regression vector $\vec{\hat{y}}$ fitted to the data has roughly the same direction. The component of $\vec{\hat{y}}$ orthogonal to $\vec{x}_1 + \vec{x}_2$ should be negligible. The vector $\vec{\hat{y}}$ in the figure appears to have both these properties.

A more complicated restriction is illustrated in Figure 6.6. The constraint $b_2 = b_3$ is applied to a three-variable regression, leading to Equation 6.11. For simplicity here, the three predictors are drawn as orthogonal, although they need not be so in general. When the parameters are unconstrained, the vector $\vec{\hat{y}}$ can fall anywhere in the three-dimensional region implied by the cubic structure. When the restriction is imposed, the vector falls in a two-dimensional plane or subspace of $\mathcal{V}_X$. Any value of b_1 is possible, so the restricted $\vec{\hat{y}}^{\star}$ can end at any position along the $\vec{x}_1$ dimension, but only points with identical components along $\vec{x}_2$ and $\vec{x}_3$ are consistent with the restriction. The diagonal plane $\mathcal{V}_{X \perp r}$ in the left panel of Figure 6.6 is this restricted regression space.

More generally, any set of linear restrictions on the parameters breaks $\mathcal{V}_X$ into two complementary subspaces, a space $\mathcal{V}_r$ that should be avoided and a space $\mathcal{V}_{X \perp r}$ in which the combined regression vector should lie. These subspaces are orthogonal complements within the regression space $\vec{x}$. In Figure 6.6 the acceptable space $\mathcal{V}_{X \perp r}$ is the plane striking through $\mathcal{V}$ in the left panel. Any vector in which the $\vec{x}_2$ and $\vec{x}_3$ components are the same is in $\mathcal{V}_{X \perp r}$. The restricted space $\mathcal{V}_r$ is orthogonal to this vector and is shown in the plane of $\vec{x}_2$ and $\vec{x}_3$ in the right-hand panel. It is the space generated by the difference $\vec{x}_2 - \vec{x}_3$.

The two subspaces $\mathcal{V}_r$ and $\mathcal{V}_{X \perp r}$ are are orthogonal complements, so that their dimensions sum to that of the whole space:

$$\dim(\mathcal{V}_X) = \dim(\mathcal{V}_r) + \dim(\mathcal{V}_{X \perp r}).$$

Each independent restriction on the parameters increases the dimension of $\mathcal{V}_r$ by one and reduces that of $\mathcal{V}_{X \perp r}$ by one. In Figure 6.6, $\dim(\mathcal{V}_X) = 3$ and there is one restriction, so $\dim(\mathcal{V}_r) = 1$ and $\dim(\mathcal{V}_{X \perp r}) = 2$. Written as degrees of freedom and with the terms rearranged,

$$df_{\text{restricted effect}} = df_{\text{unrestricted effect}} - df_{\text{restrictions}}. \tag{6.12}$$

The regression vector $\vec{\hat{y}}$ has orthogonal components in the two subspaces: $\vec{\hat{y}}_r \in \mathcal{V}_r$ and $\vec{\hat{y}}_{X \perp r} \in \mathcal{V}_{X \perp r}$. Each of these components represents

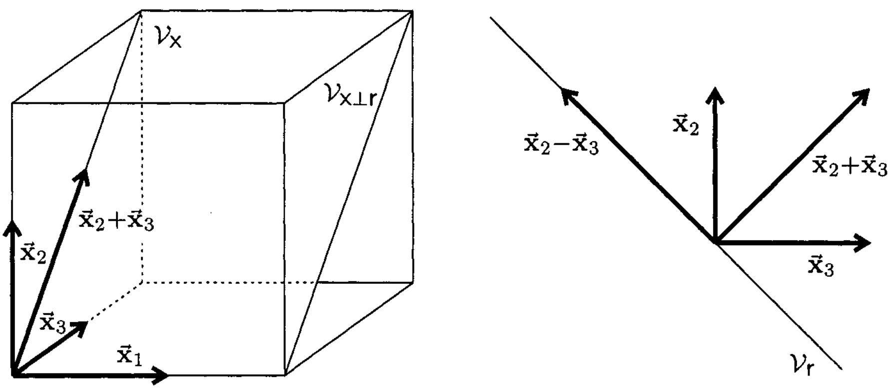


Figure 6.6: *The effect space of a three-variable regression with the constraint $b_2 = b_3$. The cube-like space in the left panel is the unconstrained space of three predictors, and the diagonal plane is the constrained space. The right panel shows the plane of $\vec{\mathbf{x}}_2$ and $\vec{\mathbf{x}}_3$ and the one-dimensional restriction space $\mathcal{V}_{\mathsf{r}}$.*

a different aspect of the regression problem, and a comparison of each to the error vector tests a different hypothesis. The component in $\mathcal{V}_{\mathsf{r}}$ gives information about the restriction. If the null hypothesis asserted by the restriction is true, then $\vec{\mu}_{\mathsf{y}\cdot\mathsf{x}}$ falls entirely in the space $\mathcal{V}_{\mathsf{x}\perp\mathsf{r}}$ and $\vec{\hat{\mathbf{y}}}_{\mathsf{r}}$ is nothing but sampling error. When the restriction fails, then $\vec{\hat{\mathbf{y}}}_{\mathsf{r}}$ contains a systematic component. The component in $\mathcal{V}_{\mathsf{x}\perp\mathsf{r}}$ tells about the usefulness of the restricted equation. If this equation is not a predictor of $\vec{\mathbf{y}}$, then $\vec{\hat{\mathbf{y}}}_{\mathsf{x}\perp\mathsf{r}}$ is comparable to the error vector. When this hypothesis is rejected, one concludes that some of the coefficients of the restricted equation are nonzero. Mean squares associated with the two effect vectors are calculated within their respective subspaces:

$$M(\vec{\hat{\mathbf{y}}}_{\mathsf{r}}) = \frac{|\vec{\hat{\mathbf{y}}}_{\mathsf{r}}|^2}{\dim(\mathcal{V}_{\mathsf{r}})} \qquad \text{and} \qquad M(\vec{\hat{\mathbf{y}}}_{\mathsf{x}\perp\mathsf{r}}) = \frac{|\vec{\hat{\mathbf{y}}}_{\mathsf{x}\perp\mathsf{r}}|^2}{\dim(\mathcal{V}_{\mathsf{x}\perp\mathsf{r}})}.$$

The spherical distribution of $\vec{\varepsilon}$ lets these mean squares be compared to the overall estimate of error obtained from $\vec{\mathbf{e}}$, in the same way that the omnibus test was run in the regression space as a whole. The two F ratios are

$$F_{\mathsf{r}} = \frac{M(\vec{\hat{\mathbf{y}}}_{\mathsf{r}})}{M(\vec{\mathbf{e}})} \qquad \text{and} \qquad F_{\mathsf{x}\perp\mathsf{r}} = \frac{M(\vec{\hat{\mathbf{y}}}_{\mathsf{x}\perp\mathsf{r}})}{M(\vec{\mathbf{e}})}.$$

A large value of the first of these statistics implies that the restrictions fail, and a large value of the second statistic implies that the restricted equation accounts for an nonaccidental portion of the variability.

Although it is far more common to test $\hat{\vec{y}}_{\mathsf{r}}$ than $\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}}$, it is usually easiest to find the length of the former vector indirectly. First, the restrictions are used to find a basis for $\mathcal{V}_{\mathsf{X}\perp\mathsf{r}}$; next, $\vec{y}$ is regressed on these basis vectors to find $\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}}$; finally, the squared length of the component in the restriction space is determined by subtraction:

$$|\hat{\vec{y}}_{\mathsf{r}}|^2 = |\hat{\vec{y}}|^2 - |\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}}|^2. \tag{6.13}$$

The squared multiple correlation coefficients obey the same rule. For example, consider the restriction $b_2 = b_3$ in a three-variable regression. Substituting b_2 for b_3 in the full equation gives

$$\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}} = b_1\vec{x}_1 + b_2(\vec{x}_2 + \vec{x}_3)$$

(Equation 6.11). This equation shows that $\vec{x}_1$ and the new predictor vector $\vec{x}_4 = \vec{x}_2 + \vec{x}_3$ are a basis for $\mathcal{V}_{\mathsf{X}\perp\mathsf{r}}$. The restricted regression equation is fitted as the unconstrained two-variable regression

$$\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}} = b_1\vec{x}_1 + b_2\vec{x}_4.$$

With $\hat{\vec{y}}_{\mathsf{X}\perp\mathsf{r}}$ in hand, Equation 6.13 is applied to get the length of $\vec{y}_{\mathsf{r}}$. The algebraic procedures for conditional tests have the effect of subdividing the spaces in this way.

Exercises

1. Suppose that you have data from four variables measured on 45 subjects. You regress one variable on the other three.

a. What are the dimensions of the subspaces $\mathcal{V}_1$, $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{e}}$?

b. How do these results change if the three predictors are multicollinear with one redundancy?

2. For the regression in Problem 3.1 (using data from Problem 2.5), test the following hypotheses:

a. $\beta = 0$

b. $\alpha = 0$

c. $\alpha = 5$

3. Suppose that the true mean of variable Y is $\mu_{\mathsf{Y}} = 3$ and that observations are subject to non-normal error ε that is uniformly distributed between $-\frac{1}{2}$ and $+\frac{1}{2}$ (i.e., Y is equally likely to take any value between $2\frac{1}{2}$ and $3\frac{1}{2}$). Consider two observations. Draw the vector $\vec{\boldsymbol{\mu}}$ in subject space, and around its tip shade the region where $\vec{\mathbf{y}}$ can fall. Note that this region is not circularly symmetrical. What angles are most likely to occur?

4. Suppose that the restriction $b_1 = -2b_2$ is made to a two-variable regression.

a. Draw a picture of the effect space showing the subspaces $\mathcal{V}_{\mathsf{r}}$ and $\mathcal{V}_{\perp\mathsf{r}}$.

b. Test this hypothesis for the data in part **a** of Problem 4.1.

5. Suppose that you wish to test the joint hypothesis

$$b_1 = 2b_3 \qquad \text{and} \qquad b_4 = 0$$

in a four-predictor regression. How would you find $M(\vec{\hat{\mathbf{y}}}_{\mathsf{r}})$?

Chapter 7

Conditional relationships

The causal influences among the variables in an large set of multivariate data can be very complicated. One variable may influence several others and by doing so induce associations among them. Often it is conceptually important to be able to exclude the influences of certain variables from consideration when one looks at the relationships among other variables. For example, a common problem is to look at the *conditional association* between two variables (or sets of variables) while eliminating, in some sense, any effects related to a third variable (or set). Geometrically, removing the effects of a set of variables is accomplished by projection—this is the particular sense in which the word "removed" is used in multivariate analysis. To eliminate the influence of a variable or set of variables, the entire analysis is projected into the orthogonal complement of the space generated by the removed variables. This chapter examines this problem in two important situations, partial correlation and multiple regression.

7.1 Partial correlation

Suppose that one has three intercorrelated variables, X, Y, and Z. Moreover, suppose that the variable X has an effect that acts in some sense prior to Y and Z so that it influences both these variables but is not influenced by them. The correlation of Y and Z with the precursor variable X produces an association. For example, if high scores in X tend to lead to high scores in both Y and Z, then the latter two variables will have a positive correlation, even without any causal connection between them. Consequently, the simple correlation of Y and Z is a mixture of effects, combining the association that can be attributed to their correlation to X with any unique association of Y and Z. One wants a way to extract the latter component and to measure it separately from the part that depends on X. The partial correlation coefficient is the desired index.

Figure 7.1 shows the geometry of the partial correlation coefficient. Three vectors $\vec{x}$, $\vec{y}$, and $\vec{z}$ correspond to the three variables. The left panel shows the three-dimensional configuration of the variables, along with three subspaces drawn as planes. The plane $\mathcal{V}_{XZ}$ contains the vectors $\vec{x}$ and $\vec{z}$ (indeed, it is generated by them), and the plane $\mathcal{V}_{XY}$ contains the vectors $\vec{x}$ and $\vec{y}$. As drawn, there is an acute angle between each pair of vectors, indicating positive pairwise correlations. In particular, when $\vec{y}$ and $\vec{z}$ are considered in isolation, the angle between them is less than 90°, as shown in the subspace $\mathcal{V}_{YZ}$ illustrated at the upper right. The partial correlation coefficient is based on the assertion that any variability of Y and Z that is in common with X can be attributed entirely to X. The component $\vec{y}_X$ of $\vec{y}$ parallel to $\vec{x}$ and the comparable component $\vec{z}_X$ of $\vec{z}$ are deemed to be due to the effects of X. The unique parts of $\vec{y}$ and $\vec{z}$ are carried by components orthogonal to these. The projections $\vec{y}_{\perp X}$ and $\vec{z}_{\perp X}$ of $\vec{y}$ and $\vec{z}$ onto the orthogonal complement $\mathcal{V}_{\perp X}$ of $\vec{x}$ describe the variation unrelated to X. The left panel of Figure 7.1 shows these projections. The residual relationship between Y and Z after excluding X is defined by the angle between these projections, as shown in the small panel at lower right.

The projections $\vec{y}_{\perp X}$ and $\vec{z}_{\perp X}$ describe the components of Y and Z from which the effects of X have been extracted. Thus, the relationship between these vectors is a measure of the conditional association of the variables. As always in traditional multivariate statistics, the association is measured by the correlation of the variables, which equals the cosine of the angle between the corresponding vectors. With this justification, the *partial correlation coefficient* is defined to be $r_{YZ \cdot X} = \cos \angle(\vec{y}_{\perp X}, \vec{z}_{\perp X})$. This coefficient measures the portion of the relationship between Y and Z that has no linear relationship to X. The subscripts to the left of the dot in $r_{YZ \cdot X}$ indicate the variables whose association is being measured, and those to the right of the dot indicate the variable(s) whose variation is excluded.

Figure 7.1 illustrates an important fact: projecting two vectors into a subspace can change the angle between them. The two panels at the right of the figure show the vectors representing Y and Z, before and after conditioning on X. The angle between them changes from acute to obtuse. Correspondingly, the simple correlation $r_{YZ} = \cos \angle(\vec{y}, \vec{z})$ is positive, and the partial correlation $r_{YZ \cdot X} = \cos \angle(\vec{y}_{\perp X}, \vec{z}_{\perp X})$ is negative. This change is not as paradoxical as it may seem at first. Both $\vec{y}$ and $\vec{z}$ have an appreciable component along $\vec{x}$. This common component gives the vectors their initial positive relationship. When this portion of their variability is removed, the residuals are quite different from the original vectors and have a negative relationship. Superficial paradoxes such as this can arise whenever a large

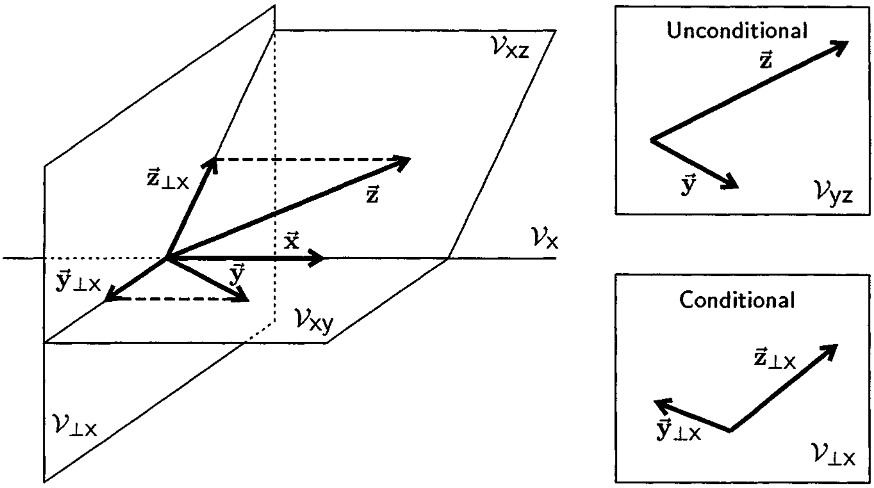


Figure 7.1: *The partial correlation of* Y *and* Z *given* X. *The left panel shows the three-dimensional picture with the plane orthogonal to* $\vec{x}$ *and the two planes spanned by* $\vec{x}$ *and* $\vec{z}$ *and by* $\vec{x}$ *and* $\vec{y}$. *The two panels at right show the vector relationship between the variables* Y *and* Z *in the unconditional and the conditional spaces.*

component of two variables is eliminated. Because their greater part is gone, the residual vectors may point in a very different direction from their unadjusted forms.

A user of partial correlation should keep two other points in mind. First, removing the component of two variables related to a third is sensible only when the decomposition into parts is meaningful. The causal relationship among the variables is crucial here. If X, in some sense, precedes both Y and Z, so that their common part arises from the same source, then the elimination of this part is interpretable. However, if one or both of Y and Z precede X, so that their association with X comes from different sources, then the space $\mathcal{V}_{\perp X}$ has no good interpretation and projection into that space is inappropriate. The geometry does not distinguish between these possibilities—it is the same for both. Instead, one must consider the logical structure of the problem.

The second point concerns the nature of the adjustment. Projection is only one way to accommodate an associated variable. Removing $\vec{x}$ acts in an essentially linear way on $\vec{y}$ and $\vec{z}$. This form of adjustment is the simplest and the most probable approach a priori, but it is still a statistical fiction. The space $\mathcal{V}_{\perp X}$ is constructed by mathematical or geometrical projection, not measured from real variables. Again, the geometry itself cannot tell one whether the influences of X on Y and Z really work that way. If the true relationship is substantially nonlinear (as a scatterplot would show), then the projective adjustments are inappropriate and the partial correlation is meaningless. One should particularly remember that geometrical projection does not substitute for forms of control such as those produced by an experimental design. For example, the adjusted correlation is not necessarily the same as the value that would have been obtained had all instances been chosen with the same value of variable X.

A formula to calculate the value of the partial correlation coefficient can be derived from the geometry. The partial correlation coefficient $r_{YZ\cdot X}$ is the angle between the residual vectors $\vec{\mathbf{y}}_{\perp X}$ and $\vec{\mathbf{z}}_{\perp X}$. The regressions of $\vec{\mathbf{y}}$ and $\vec{\mathbf{z}}$ on $\vec{\mathbf{x}}$ are

$$\vec{\mathbf{y}}_X = b_{Y\cdot X}\vec{\mathbf{x}} \qquad \text{and} \qquad \vec{\mathbf{z}}_X = b_{Z\cdot X}\vec{\mathbf{x}},$$

with the regression coefficients (Equation 3.4)

$$b_{Y\cdot X} = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{y}}}{|\vec{\mathbf{x}}|^2} \qquad \text{and} \qquad b_{Z\cdot X} = \frac{\vec{\mathbf{x}}\cdot\vec{\mathbf{z}}}{|\vec{\mathbf{x}}|^2}.$$

The residual vectors are the differences

$$\vec{\mathbf{y}}_{\perp X} = \vec{\mathbf{y}} - \vec{\mathbf{y}}_X = \vec{\mathbf{y}} - b_{Y\cdot X}\vec{\mathbf{x}}$$

and

$$\vec{\mathbf{z}}_{\perp X} = \vec{\mathbf{z}} - \vec{\mathbf{z}}_X = \vec{\mathbf{y}} - b_{Z\cdot X}\vec{\mathbf{x}}.$$

The cosine of the angle between these two vectors is the partial correlation:

$$r_{YZ\cdot X} = \cos\angle(\vec{\mathbf{y}}_{\perp X}, \vec{\mathbf{z}}_{\perp X}) = \frac{\vec{\mathbf{y}}_{\perp X}\cdot\vec{\mathbf{z}}_{\perp X}}{|\vec{\mathbf{y}}_{\perp X}||\vec{\mathbf{z}}_{\perp X}|}.$$

Replacing the projected vectors by the differences and applying the distributive law gives

$$\begin{aligned} r_{YZ\cdot X} &= \frac{(\vec{\mathbf{y}} - b_{Y\cdot X}\vec{\mathbf{x}})\cdot(\vec{\mathbf{z}} - b_{Z\cdot X}\vec{\mathbf{x}})}{|\vec{\mathbf{y}}_{\perp X}||\vec{\mathbf{z}}_{\perp X}|} \\ &= \frac{\vec{\mathbf{y}}\cdot\vec{\mathbf{z}} - b_{Z\cdot X}(\vec{\mathbf{y}}\cdot\vec{\mathbf{x}}) - b_{Y\cdot X}(\vec{\mathbf{z}}\cdot\vec{\mathbf{x}}) + b_{Y\cdot X}b_{Z\cdot X}(\vec{\mathbf{x}}\cdot\vec{\mathbf{x}})}{|\vec{\mathbf{y}}_{\perp X}||\vec{\mathbf{z}}_{\perp X}|}. \end{aligned}$$

Substitution of the regression coefficients gives, after some algebra,

$$r_{YZ\cdot X} = \frac{\vec{\mathbf{y}}\cdot\vec{\mathbf{z}} - \dfrac{(\vec{\mathbf{y}}\cdot\vec{\mathbf{x}})(\vec{\mathbf{z}}\cdot\vec{\mathbf{x}})}{|\vec{\mathbf{x}}|^2}}{|\vec{\mathbf{y}}_{\perp X}||\vec{\mathbf{z}}_{\perp X}|}.$$

It remains only to rewrite this expression using correlations. Substituting such facts as $\vec{\mathsf{y}}\cdot\vec{\mathsf{z}} = |\vec{\mathsf{y}}||\vec{\mathsf{z}}|r_{\mathsf{yz}}$ and $|\vec{\mathsf{y}}_{\perp\mathsf{x}}|^2 = (1 - r^2_{\mathsf{xy}})|\vec{\mathsf{y}}|^2$ (cf. Equation 3.9) gives

$$r_{\mathsf{yz}\cdot\mathsf{x}} = \frac{r_{\mathsf{yz}} - r_{\mathsf{xy}}r_{\mathsf{xz}}}{\sqrt{1 - r^2_{\mathsf{xy}}}\sqrt{1 - r^2_{\mathsf{xz}}}}. \tag{7.1}$$

This is the conventional way to compute the partial correlation coefficient.

Similar principles let one condition a relationship on several variables simultaneously. The essential geometry is the same as in the single-variable case. Instead of a single conditioning vector $\vec{\mathsf{x}}$ whose effects are to be eliminated (as in Figure 7.1), one has a set of vectors $\vec{\mathsf{x}}_1, \vec{\mathsf{x}}_2, \ldots, \vec{\mathsf{x}}_p$ spanning a p-dimensional subspace $\mathcal{V}_{\mathsf{X}}$. Together with the two vectors whose relationship is to be measured, these vectors span a (p+2)-dimensional space composed of $\mathcal{V}_{\mathsf{X}}$ and its orthogonal complement $\mathcal{V}_{\perp\mathsf{X}}$. The vectors to be conditioned are projected onto $\mathcal{V}_{\perp\mathsf{X}}$, and the components in this space are examined. The partial correlation coefficient is the cosine of the angle between the projections.

Although the geometry is similar, the algebraic formula for the partial correlation coefficient conditioned on several variables is more complicated than Equation 7.1, particularly when the conditioning variables are not mutually orthogonal. As in multiple regression, a set of simultaneous equations must be solved. The procedure in matrix form is given in algebraically-oriented texts.

7.2 Conditional effects in multiple regression

A common problem in multiple regression analysis is to evaluate the effects of one or more predictors in the context of another predictor or set of predictors. These tests are important for several reasons. One reason is practical. Sometimes one has an equation based on one set of predictors and wishes to see whether a second predictor (or set of predictors) improves one's ability to predict the outcome variable. If so, then one can profitably use the larger equation; if not, then the simpler equation is better. Another reason is theoretical. Suppose that one believes that the relationship between several predictor variables and an outcome variable is obscured by their association with another set of variables. A better picture of the obscured relationship is obtained if one removes the effects of the obscuring set. The intent here is like that underlying partial correlation, but set in a predictor-outcome context. Both types of questions are answered by a conditional test of the multiple regression effects.

Denote the two sets of predictors by $\mathsf{X}_1, \mathsf{X}_2, \ldots, \mathsf{X}_p$ and $\mathsf{Z}_1, \mathsf{Z}_2, \ldots, \mathsf{Z}_q$ and the outcome variable by Y. For simplicity, assume that there are no linear dependencies among these variables—if there are, then eliminate or redefine variables until the sets are linearly independent. The goal of the analysis is to create a test for the effect of the Z_k on Y that excludes the variability explicable by the X_j. The sets of vectors $\vec{\mathbf{x}}_1, \vec{\mathbf{x}}_2, \ldots, \vec{\mathbf{x}}_p$ and $\vec{\mathbf{z}}_1, \vec{\mathbf{z}}_2, \ldots, \vec{\mathbf{z}}_q$ span spaces $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Z}$ of dimension p and q, respectively. They have only the null vector $\vec{\mathbf{0}}$ in common. Considered separately, the two sets of predictors lead to the regression vectors $\hat{\vec{\mathbf{y}}}_\mathsf{X} \in \mathcal{V}_\mathsf{X}$ and $\hat{\vec{\mathbf{y}}}_\mathsf{Z} \in \mathcal{V}_\mathsf{Z}$. When the sets of vectors are taken together, a space $\mathcal{V}_\mathsf{XZ}$ results that includes both $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Z}$. The best combined predictor $\hat{\vec{\mathbf{y}}}_\mathsf{XZ}$ lies in this space. The tests of Section 6.3 can be applied to each of the spaces, comparing the mean-squared length of an effect vector to that of an error vector in the unexplained space.

If the two spaces $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Z}$ were orthogonal to each other, then each could be examined separately as described in Section 5.3. If they are not orthogonal, then they share some variability, and the analysis must take account of this shared component. In a conditional analysis, the common portion of the variability is attributed to one of the sets of variables and excluded from the other set. Since the X_j are deemed to be the obscuring variables, everything in $\mathcal{V}_\mathsf{X}$ is attributed to them, and only the residual effects of the Z_k are examined. Instead of examining prediction vectors in the space $\mathcal{V}_\mathsf{Z}$, one looks at vectors in the space $\mathcal{V}_{\mathsf{Z}\perp\mathsf{X}}$ that is orthogonal to the conditioning space $\mathcal{V}_\mathsf{X}$.

Geometrically, the variability is separated by constructing a series of orthogonal subspaces. First, there are the overall regression space $\mathcal{V}_\mathsf{XZ}$ and the error space $\mathcal{V}_\mathbf{e}$. Then, the regression space $\mathcal{V}_\mathsf{XZ}$ is broken into the two orthogonal subspaces $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_{\mathsf{Z}\perp\mathsf{X}}$. The first of these spaces is spanned by the $\vec{\mathbf{x}}_j$, and the second space is the orthogonal complement of $\mathcal{V}_\mathsf{X}$ in $\mathcal{V}_\mathsf{XZ}$. The outcome vector splits into components in these three spaces:

$$\vec{\mathbf{y}} = \hat{\vec{\mathbf{y}}}_\mathsf{XZ} + \vec{\mathbf{e}}_\mathsf{XZ} = \hat{\vec{\mathbf{y}}}_\mathsf{X} + \hat{\vec{\mathbf{y}}}_{\mathsf{Z}\perp\mathsf{X}} + \vec{\mathbf{e}}_\mathsf{XZ}. \tag{7.2}$$

The conditional prediction of Y by the Z_k given the X_j is determined by the vector $\hat{\vec{\mathbf{y}}}_{\mathsf{Z}\perp\mathsf{X}} \in \mathcal{V}_{\mathsf{Z}\perp\mathsf{X}}$, not by the vector $\hat{\vec{\mathbf{y}}}_\mathsf{Z}$ in the unconditional space. If the conditional vector is comparatively short, then the Z_k add little to the X_j, while if it is comparatively long, then the Z_k have something to contribute.

When visualizing these operations, it helps to restrict the picture to three dimensions. The geometrical examples below use one variable in each set and drop the subscripts on X_j and Z_k. However, the procedures apply for any p and q (at least, if the sample size is large enough), and their

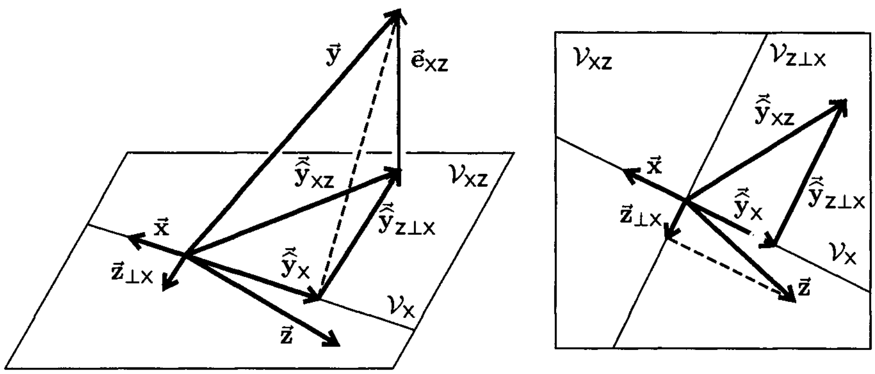


Figure 7.2: *The geometry of conditional testing in multiple regression. The left panel shows the three-dimensional picture, the right panel shows the regression space $\mathcal{V}_{XZ}$ with its subspaces.*

multivariate nature should not be forgotten. Throughout the examples, the single variables X and Z should be seen as standing in for the sets $\{X_1, X_2, \ldots, X_p\}$ and $\{Z_1, Z_2, \ldots, Z_q\}$ and the vectors $\vec{x}$ and $\vec{z}$ as proxies for the subspaces $\mathcal{V}_X$ and $\mathcal{V}_Z$. Figure 7.2 illustrates the partitioning. On the left is the picture in three dimensions, and on the right is the regression plane $\mathcal{V}_{XZ}$ spanned by $\vec{x}$ and $\vec{z}$. The basic representation of multiple regression is familiar, with two predictors and one outcome variable. As drawn, the predictors $\vec{x}$ and $\vec{z}$ are not orthogonal. The right-hand panel shows the full regression space $\mathcal{V}_{XZ}$ containing both the space $\mathcal{V}_X$ spanned by $\vec{x}$ and its orthogonal complement $\mathcal{V}_{Z\perp X}$. Note that $\mathcal{V}_{Z\perp X}$ is not the same subspace as $\mathcal{V}_Z$. Although $V_{Z\perp X}$ contains a projection of $\vec{z}$, it does not contain $\vec{z}$ itself. Two regression vectors are shown, the vector $\hat{\vec{y}}_X$ determined by $\vec{x}$ alone and the vector $\hat{\vec{y}}_{XZ}$ based on both $\vec{x}$ and $\vec{z}$. The vector $\vec{z}$ is not used directly as a predictor. Instead, $\hat{\vec{y}}_{XZ}$ is viewed as an orthogonal combination of $\hat{\vec{y}}_X \in \mathcal{V}_X$ and $\hat{\vec{y}}_{Z\perp X} \in \mathcal{V}_{Z\perp X}$. The latter vector is drawn displaced away from $\mathbf{0}$ to emphasize the addition.

Another way to think of this analysis is to imagine that the to-be-conditioned predictors are replaced by their projections into $\mathcal{V}_{Z\perp X}$ before $\vec{y}$ is regressed onto them. The vectors $\vec{z}_1$, $\vec{z}_2$, ..., $\vec{z}_q$ become the vectors $\vec{z}_{1\perp X}$, $\vec{z}_{2\perp X}$, ..., $\vec{z}_{q\perp X}$, each of which is orthogonal to all the $\vec{x}_j$. These vectors represent the portion of the Z_k that is unrelated to the X_j and that span the space $\mathcal{V}_{Z\perp X}$. Figure 7.2 also shows the vector $\vec{z}_{\perp X}$. The sets $\{\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_p\}$ and $\{\vec{z}_{1\perp X}, \vec{z}_{2\perp X}, \ldots, \vec{z}_{q\perp X}\}$ are orthogonal and can be analyzed by separate regressions in $\mathcal{V}_X$ and $\mathcal{V}_{Z\perp X}$, respectively.

The second approach points up the similarity between conditional regression and the partial correlation coefficient discussed in the last section. In both procedures, the effects of a set of variables are projected into a subspace before an analysis of association is made. However, the two procedures differ in their treatment of the variable Y. In partial correlation, both the vectors $\vec{y}$ and $\vec{z}$ are projected into the space orthogonal to $\mathcal{V}_X$. In multiple regression, one remains interested in predicting Y, not a transformation of it, so only the predictor variables are adjusted. The same outcome vector $\vec{y}$ is used in both the conditional and the unconditional analyses.

The regression decomposition of Equation 7.2 shows $\vec{y}$ written as the sum of three mutually orthogonal vectors, $\vec{\hat{y}}_X \in \mathcal{V}_X$, $\vec{\hat{y}}_{Z\perp X} \in \mathcal{V}_{Z\perp X}$, and $\vec{e}_{XZ} \in \mathcal{V}_e$. Applying the Pythagorean Theorem to this configuration gives

$$|\vec{y}|^2 = |\vec{\hat{y}}_X|^2 + |\vec{\hat{y}}_{Z\perp X}|^2 + |\vec{e}_{XZ}|^2. \tag{7.3}$$

These lengths measure the importance of the three components of $\vec{y}$. In more conventional notation, writing sums of squares for the squared lengths,

$$SS_Y = SS_{Y\cdot X} + SS_{Y\cdot Z|X} + SS_{\text{error}}. \tag{7.4}$$

They represent, respectively, the amount of the variability of Y that can be attributed to a linear predictor based on the X_j, the additional amount of variation that can be attributed to the introduction of the Z_k, and the error that cannot be explained by a linear predictor based on both variables.

The squared lengths in Equation 7.3 can be expressed by multiple correlation coefficients. Let $R_{Y\cdot X}$ denote the multiple correlation coefficient for the regression of Y on the variables X_j, and $R_{Y\cdot XZ}$ denote the multiple correlation coefficients for the regression of Y on both the X_j and the Z_k. From the decomposition of the sums of squares in basic multiple regression analysis (as in Equations 4.8, 4.9, and 4.10), the components of $\vec{y}$ are related to its length by

$$SS_{Y\cdot X} = |\vec{\hat{y}}_X|^2 = R^2_{Y\cdot X}|\vec{y}|^2,$$

$$SS_{Y\cdot XZ} = |\vec{\hat{y}}_{XZ}|^2 = R^2_{Y\cdot XZ}|\vec{y}|^2,$$

and

$$SS_{\text{error}} = |\vec{e}_{XZ}|^2 = (1 - R^2_{Y\cdot XZ})|\vec{y}|^2.$$

The orthogonality of $\vec{\hat{y}}_X$ and $\vec{\hat{y}}_{Z\perp X}$ lets the portion attributable to $\vec{z}$ be found by subtraction:

$$SS_{Y\cdot Z|X} = |\vec{\hat{y}}_{Z\perp X}|^2 = |\vec{\hat{y}}_{XZ}|^2 - |\vec{\hat{y}}_X|^2 = (R^2_{Y\cdot XZ} - R^2_{Y\cdot X})|\vec{y}|^2. \tag{7.5}$$

As with the partial correlation coefficient, it is important to recognize that the effects of conditioning can be substantial. When the space $\mathcal{V}_{\mathsf{X}}$ of the conditioning variables is nearly orthogonal to that of the conditioned variable $\mathcal{V}_{\mathsf{Z}}$, then the conditioning operation has little effect on $\vec{\hat{\mathbf{y}}}_{\mathsf{Z}}$, and the conditional and unconditional effects are similar. The discussion of orthogonal effects in Section 5.3 illustrates this happy case. However, when $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Z}}$ are angularly close to each other, projection of the $\vec{\mathbf{z}}_k$ into $\mathcal{V}_{\perp\mathsf{X}}$ removes a substantial component and can considerably change the vectors' directions. The same shift illustrated for partial correlation in Figure 7.1 can occur. This change of direction will usually change the relationship of the vectors and the space they span to $\vec{\mathbf{y}}$. Depending on the particular geometry, the effect attributable to the Z_j can either increase or decrease when conditioned. In Figure 7.2, the component associated with $\vec{\mathbf{z}}$ is longer in the conditional measure than when $\vec{\mathbf{x}}$ is not part of the equation (i.e., it exhibits the suppressor activity discussed in Section 5.4). More commonly, X and Z have much variability in common, both with each other and with Y, and one finds that the component along $\vec{\mathbf{z}}_{\perp\mathsf{X}}$ is smaller than the component along $\vec{\mathbf{z}}$ alone.

As in all such analyses, the important relationships are those between the subspaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Z}}$. It is not necessary for the individual variables to be closely related in order for the subspaces to have a similar component—recall the example of multicollinearity in Figure 5.1 where no pair of vectors is mutually closely related, yet the space spanned by any pair of vectors includes the third. The analysis must be considered at the level of the spaces, not of the individual vectors.

7.3 Statistical tests of conditional effects

Tests of conditional multiple regression effects are essentially the same as those for unconditional effects described in Section 6.3. Under the null hypothesis that the X_j have no relationship to Y, $\mathcal{V}_{\mathsf{X}}$ is orthogonal to $\vec{\mathbf{y}}$, and the vector $\vec{\hat{\mathbf{y}}}_{\mathsf{X}}$ is a random vector in that p-dimensional space. Likewise, if the Z_j have no relationship to Y other than that already captured by X, then $\vec{\hat{\mathbf{y}}}_{\mathsf{Z}\perp\mathsf{X}}$ is a random vector in the q-dimensional space $\mathcal{V}_{\mathsf{Z}\perp\mathsf{X}}$. Because the spaces $\mathcal{V}_{\mathsf{X}}$ and $V_{\mathsf{Z}\perp\mathsf{X}}$ are orthogonal, they can be examined separately.

The tests are made by comparing the average per-dimension squared lengths of the hypothesis vectors to that of the error vector. For the X_j the mean-square length is

$$M(\vec{\hat{\mathbf{y}}}_{\mathsf{X}}) = \frac{|\vec{\hat{\mathbf{y}}}_{\mathsf{X}}|^2}{\dim(\mathcal{V}_{\mathsf{X}})} = \frac{R^2_{\mathsf{Y}\cdot\mathsf{X}}|\vec{\mathbf{y}}|^2}{p}.$$

Source	SS	df	MS	F
X	$\lvert\vec{\mathsf{y}}\rvert^2 R^2_{\mathsf{y}\cdot\mathsf{x}}$	p	$M(\vec{\hat{\mathsf{y}}}_{\mathsf{X}})$	$MS_{\mathsf{X}}/MS_{\mathsf{e}}$
$\mathsf{Z} \perp \mathsf{X}$	$\lvert\vec{\mathsf{y}}\rvert^2 (R^2_{\mathsf{y}\cdot\mathsf{xz}} - R^2_{\mathsf{y}\cdot\mathsf{x}})$	q	$M(\vec{\hat{\mathsf{y}}}_{\mathsf{Z}\perp\mathsf{X}})$	$MS_{\mathsf{Z}\perp\mathsf{X}}/MS_{\mathsf{e}}$
X, Z	$\lvert\vec{\mathsf{y}}\rvert^2 R^2_{\mathsf{y}\cdot\mathsf{xz}}$	$p+q$		
Error	$\lvert\vec{\mathsf{y}}\rvert^2 (1 - R^2_{\mathsf{y}\cdot\mathsf{xz}})$	$N-p-q-1$	$M(\vec{\mathsf{e}})$	
Total	$\lvert\vec{\mathsf{y}}\rvert^2$	$N-1$		

Table 7.1: *Unconditional and conditional effects in multiple regression, expressed as an analysis-of-variance table.*

For the conditional test of the Z_k, the mean-square length is

$$M(\vec{\hat{\mathsf{y}}}_{\mathsf{Z}\perp\mathsf{X}}) = \frac{\lvert\vec{\hat{\mathsf{y}}}_{\mathsf{Z}\perp\mathsf{X}}\rvert^2}{\dim(\mathcal{V}_{\mathsf{Z}\perp\mathsf{X}})} = \frac{(R^2_{\mathsf{y}\cdot\mathsf{xz}} - R^2_{\mathsf{y}\cdot\mathsf{x}})\lvert\vec{\mathsf{y}}\rvert^2}{q}.$$

The mean-square length of the error vector is

$$M(\vec{\mathsf{e}}) = \frac{\lvert\vec{\mathsf{e}}\rvert^2}{\dim(\mathcal{V}_{\mathsf{e}})} = \frac{(1 - R^2_{\mathsf{y}\cdot\mathsf{xz}})\lvert\vec{\mathsf{y}}\rvert^2}{N - p - q - 1}.$$

Combining these length measures gives two F ratios. One ratio tests the unconditional hypothesis that the X_j are unrelated to Y:

$$F_{\mathsf{X}} = \frac{(N-p-q-1)R^2_{\mathsf{y}\cdot\mathsf{x}}}{p(1 - R^2_{\mathsf{y}\cdot\mathsf{xz}})}. \tag{7.6}$$

The other ratio tests the hypothesis that Z_j have no incremental relationship to Y:

$$F_{\mathsf{Z}|\mathsf{X}} = \frac{(N-p-q-1)(R^2_{\mathsf{y}\cdot\mathsf{xz}} - R^2_{\mathsf{y}\cdot\mathsf{x}})}{q(1 - R^2_{\mathsf{y}\cdot\mathsf{xz}})}. \tag{7.7}$$

Both tests use the error term $M(\vec{\mathsf{e}})$ derived from the full regression. It is common to write the tests of Equations 7.6 and 7.7 in an analysis-of-variance table, such as Table 7.1, here filled in with the vector quantities. A table of this sort organizes and systematizes the computation and presents it in a readily understood form.

Both of these tests differ from those that would be run were the other set of variables absent. The unconditional test of Equation 7.6 differs from a simple multiple regression test of X alone (e.g., Equation 6.9) in that it uses an error term that derives from the error vector for a regression with all variables present. In this way, one is assured that no systematic components related to either the X_j or the Z_j contaminate the test. If the Z_k were conditionally related to Y but were ignored in the analysis, then the systematic component $\vec{\hat{y}}_{\mathsf{Z}\perp\mathsf{X}}$ would be thrown into the error term, increasing its size. In Figure 7.2 the two candidate error terms are shown by the dashed lines. The line with $\vec{\hat{y}}_{\mathsf{Z}\perp\mathsf{X}}$ removed is much shorter than the line obtained from $\vec{\hat{y}}_{\mathsf{X}}$ alone. If $\vec{\hat{y}}_{\mathsf{Z}\perp\mathsf{X}}$ is not removed, then the apparent mean-square length $M(\vec{e})$ is a biased estimate of its true value, and the F ratio is biased downward from its correct value, reducing the power of the test. By keeping systematic effects out of the error, $M(\vec{e})$ is made smaller, giving tests that are both more accurate and more powerful. The conditional test such as Equation 7.7 is doubly different from a direct test of Z based on the projection of $\vec{y}$ onto $\mathcal{V}_{\mathsf{Z}}$. First, the conditional effect is tested instead of the direct effect; second, the error term benefits from the removal of any systematic variability associated with $\mathcal{V}_{\mathsf{X}}$.

The difference between the simple tests and the conditional tests when the sets of predictor variables are related is illustrated by the following numerical example. Suppose that the variables X, Y, and Z have been measured on $N = 50$ subjects and have the covariance and correlation matrices

$$\mathbf{S} = \begin{bmatrix} 4.0 & 4.8 & 12.0 \\ 4.8 & 9.0 & 18.0 \\ 12.0 & 18.0 & 100.0 \end{bmatrix} \qquad \text{and} \qquad \mathbf{R} = \begin{bmatrix} 1.0 & 0.8 & 0.6 \\ 0.8 & 1.0 & 0.6 \\ 0.6 & 0.6 & 1.0 \end{bmatrix}. \qquad (7.8)$$

Geometrically, the vectors $\vec{x}$ and $\vec{y}$ have lengths of 2 and 3 units, respectively, and they are set at an angle of arccos(0.8) = 36.9° to each other. The vector $\vec{y}$ is 10 units long and makes an angle of arccos(0.6) = 53.1° with both $\vec{x}$ and $\vec{y}$. These relationships are shown in Figure 7.3, on the left in the plane of $\vec{x}$ and $\vec{y}$, and on the right in the regression plane of $\vec{x}$ and $\vec{z}$—the reader should construct the configuration in three dimensions. Simple regression using either X or Z alone gives the equations

$$\vec{\hat{y}}_{\mathsf{X}} = 3\vec{x} \qquad \text{and} \qquad \vec{\hat{y}}_{\mathsf{Z}} = 2\vec{z}.$$

Both X and Z have the same correlation to Y, so both fit equally well, with $R^2_{\mathsf{y}\cdot\mathsf{X}} = R^2_{\mathsf{y}\cdot\mathsf{Z}} = 0.36$. These single-predictor regression vectors are both significantly different from chance. The test statistic for either variable treated alone is

$$F = \frac{(N-p-1)R^2}{p(1-R^2)} = \frac{(50-1-1)(0.36)}{(1)(1-0.36)} = 27.00$$

(Equation 6.9). With $p = 1$ and $N-p-1 = 48$ degrees of freedom, the hypothesis of no effect is rejected.

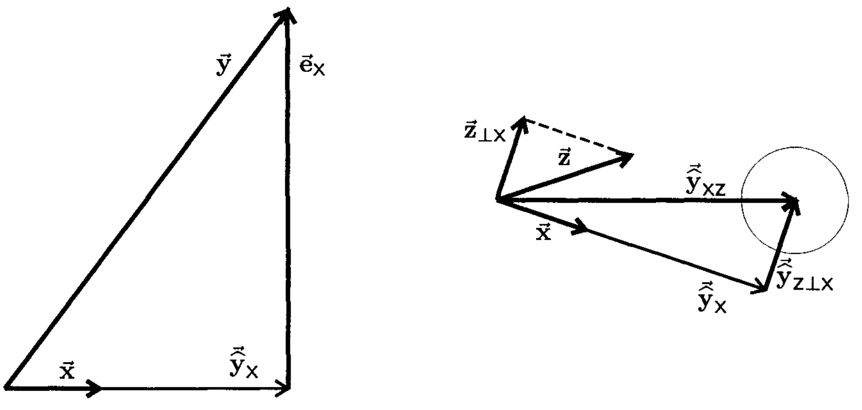


Figure 7.3: *Vectors representing a two-predictor regression with the covariance structure in Equation 7.8. The left panel shows single-variable regression of $\vec{y}$ onto $\vec{x}$. The right panel shows the regression space $\mathcal{V}_{XZ}$ with two regression vectors and a circle indicating the uncertainty of $\vec{y}$.*

When both predictors are used, the resulting regression equation (shown in the right panel of Figure 7.3) is

$$\hat{\vec{y}}_{XZ} = 1.667\vec{x} + 1.111\vec{z}.$$

The multiple correlation increases somewhat (it cannot get smaller) to $R^2_{y \cdot XZ} = 0.40$, but the improvement in fit attributable to adding Z to X is small, the difference being only 0.04. With both variables in the equation, the test of $\vec{x}$ using Equation 7.6 is

$$F_X = \frac{(50 - 1 - 1 - 1)(0.36)}{(1)(1 - 0.40)} = 28.2.$$

This statistic has $p = 1$ and $N - p - q - 1 = 47$ degrees of freedom and remains significant at well better than the 5% level. A test of the increment provided by Z (Equation 7.7) gives only

$$F_{Z \perp X} = \frac{(50 - 1 - 1 - 1)(0.40 - 0.36)}{(1)(1 - 0.40)} = 3.13.$$

With 1 and 47 degrees of freedom, this statistic is not significant at the 5% level. The contribution of Z is insufficient to be distinguished from sampling variability. Table 7.2 summarizes these results.

Source	*SS*	*df*	*MS*	*F*
X	1800	1	1800.00	28.20
$Z \perp X$	200	1	200.00	3.13
X, Z	2000	2		
Error	3000	47	63.83	
Total	5000	49		

Table 7.2: *Analysis-of-variance table summarizing the regression analysis of the covariance matrix in Equation 7.8.*

The ineffectiveness of the second predictor is apparent from the geometry. Think of erecting the right panel of Figure 7.3 on the left panel—the same vector $\hat{\vec{y}}_X$ appears in both panels—then tipping it slightly so that $\vec{y}$ lies directly above $\vec{y}_{XZ}$. In contrast to Figure 7.2, the component $\hat{\vec{y}}_{Z\perp X}$ adds little to the quality of the fit, either by lengthening $\hat{\vec{y}}$ or by reducing $\vec{e}$. When visualizing the statistical tests in this example, it helps to think of the sampling variability as adding a cloud of uncertainty, as in Figure 6.3. The center of this distribution lies at the tip of the population vector $\vec{\mu}_{Y\cdot XZ}$, which is not known, but the magnitude of the uncertainty can be indicated conveniently by a circle drawn about the tip of $\hat{\vec{y}}_{XZ}$. The circle drawn in the right-hand panel of Figure 7.3 has a radius of $\sqrt{M(\vec{e})/N} = 1.13$. The variability represented by this circle is small enough that both the vectors $\hat{\vec{y}}_X$ and $\hat{\vec{y}}_{XZ}$ stand out from $\mathbf{0}$, but is of the same order as the shorter vector $\hat{\vec{y}}_{Z\perp X}$. Differences in the length of $\hat{\vec{y}}$ are obscured by this variability.

This example also illustrates an annoying pattern of effects that often turns up in the analysis of real data with correlated predictors. Either simple regression or the unconditional portion of the bivariate regression indicates that X and Z are both significantly related to Y. However, one cannot conclude from these results that both variables are required for an explanation of Y. When either variable is included, the other is unnecessary. All one can tell from this configuration is that either X or Z helps to predict Y, but that they are too similar for both to be necessary. The data cannot help one decide which predictor to use here, since either one gives a roughly comparable fit. If extrastatistical considerations, such as the theoretical priority of one variable over the other, do not help select a predictor, then one should simply recognize the ambiguity as a target for future research.

The angular relationship between $\vec{y}$ and the projected predictor $\hat{\vec{y}}_{Z\perp X}$ can be expressed as a correlation coefficient. This coefficient measures the predictive relation between $\vec{y}$ and the part of $\mathcal{V}_Z$ unrelated to $\mathcal{V}_X$. It is known either as the *part correlation* or as the *semipartial correlation* and is denoted by $r_{\mathsf{y(z\cdot x)}}$. The part correlation differs from the partial correlation in that no correction to $\vec{\mathbf{y}}$ is made—compare the angle between $\vec{\mathbf{z}}_{\perp\mathsf{x}}$ and $\vec{\mathbf{y}}$ in Figure 7.2 to the angle between $\vec{\mathbf{z}}_{\perp\mathsf{x}}$ and $\vec{\mathbf{y}}_{\perp\mathsf{x}}$ in Figure 7.1. The right-triangular relationships between $\hat{\vec{\mathbf{y}}}_{\mathsf{x}}$, $\hat{\vec{\mathbf{y}}}_{\mathsf{z\perp x}}$, and $\hat{\vec{\mathbf{y}}}_{\mathsf{xz}}$ let one calculate $r_{\mathsf{y(z\cdot x)}}$ without directly evaluating the angle. As usual in regression, the length of the projection of the outcome vector on a subspace is given by

$$|\vec{\mathbf{y}}_{\mathsf{z\perp x}}| = |\vec{\mathbf{y}}| \cos \angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}}_{\mathsf{z\perp x}}),$$

so that

$$r^2_{\mathsf{y(z\cdot x)}} = \cos^2 \angle(\vec{\mathbf{y}}, \hat{\vec{\mathbf{y}}}_{\mathsf{z\perp x}}) = \frac{|\hat{\vec{\mathbf{y}}}_{\mathsf{z\perp x}}|^2}{|\vec{\mathbf{y}}|^2}.$$

Using the orthogonal decomposition of $\hat{\vec{\mathbf{y}}}_{\mathsf{xz}}$ to write $|\hat{\vec{\mathbf{y}}}_{\mathsf{z\perp x}}|^2$ as a difference,

$$r^2_{\mathsf{y(z\cdot x)}} = \frac{|\hat{\vec{\mathbf{y}}}_{\mathsf{xz}}|^2 - |\hat{\vec{\mathbf{y}}}_{\mathsf{x}}|^2}{|\vec{\mathbf{y}}|^2} = R^2_{\mathsf{y\cdot xz}} - R^2_{\mathsf{y\cdot x}}. \tag{7.9}$$

Thus, the squared semipartial correlation is the increment in the squared multiple correlation between the one-set and the two-set regressions. In practice, the semipartial correlation is rarely interpreted as a correlation. Its major importance is as the hierarchical contribution in a series of conditional tests.

Exercises

1. Find the partial correlation of $\mathsf{X_1}$ and $\mathsf{X_2}$ given Y for the correlation matrices in Problem 4.1. Comment on any differences between these values and the simple correlations.

2. Devise sets of vectors for which the following conditions hold, and sketch a vector diagram:

a. Both r_{xy} and $r_{\mathsf{xy\cdot z}}$ are positive.
b. r_{xy} is positive and $r_{\mathsf{xy\cdot z}}$ is negative.
c. r_{xy} is negative and $r_{\mathsf{xy\cdot z}}$ is positive.

3. Suppose that the data in Problem 4.1 were based on 50 observations. Calculate the unconditional test statistic for Z as a predictor of Y and the conditional statistic for Z in the context of X. Which are significant?

4. For the data in Problem 4.2 calculate and test the unconditional effect of $\mathsf{X_1}$ as a predictor of Y and its conditional effect in the context of $\mathsf{X_2}$. Use a diagram to explain why these results differ.

5. Figure 7.2 shows a configuration of vectors for which both $R^2_{\mathsf{y}\cdot\mathsf{z}}$ and $R^2_{\mathsf{y}\cdot\mathsf{xz}} - R^2_{\mathsf{y}\cdot\mathsf{x}}$ are substantial. Draw a picture of a different configuration, one in which the Z effect is still large but the conditional $\mathsf{Z} \perp \mathsf{X}$ effect is negligible.

6. Suppose that the orthogonal variables X and Z are used as predictors of Y. Draw a picture to explain why $M(\vec{\mathsf{y}}_{\mathsf{z}}) = M(\vec{\mathsf{y}}_{\mathsf{z}\perp\mathsf{x}})$ for this configuration.

7. *Stepwise regression* is a procedure often used to choose a small set of predictors from among many candidates. Starting with p predictors, one begins by selecting the variable that has the closest association to the outcome variable, then conditions the remaining $p-1$ predictors on that effect. One then obtains a second predictor for the regression equation by selecting the variable with the largest conditional association. The $p-2$ unused predictors are conditioned on the two variables chosen so far, and a third variable is selected. This procedure is repeated until none of the remaining predictors accounts for a substantial portion of R^2, conditional on the ones that have already been selected.

a. Suppose that in a two-predictor regression variables X_1 and X_2 have a correlation of 0.75 with each other and correlations of 0.85 and 0.80, respectively, with variable Y. How would a stepwise regression analyze this configuration?

b. Suppose that in a replication of the experiment the correlation between X_1 and X_2 remains the same but their correlations to Y are now 0.75 and 0.81. How does the analysis change? Draw a picture to explain what has happened.

Chapter 8

The analysis of variance

In one respect, a chapter on the analysis of variance seems out of place in a book on multivariate techniques. It is essentially a univariate procedure, in which the means of one dependent variable are recorded in several groups of subjects and compared. Nevertheless, there are several reasons to discuss the analysis of variance here. First, it can be developed as an extension of multiple regression, in a form known as the *general linear model.* This approach is very general and is necessary to analyze complex designs. The geometry of this approach gives considerable insight into the analysis of multifactor designs, particularly those in which the groups have unequal sizes. Second, integration of the analysis of variance into the regression framework lets one examine a mixture of categorical and graded variables within a single analysis. Finally, the analysis of variance provides the framework for two techniques that are truly multivariate, the analysis of covariance and the multivariate analysis of variance, discussed in Sections 8.4 and 10.4, respectively.

8.1 Representing group differences

Section 3.4 described how a single-predictor regression is used to investigate the difference between two groups. A dummy vector $\vec{x}$ is created that takes one value for all observations in one group and a different value for all observations in the other group. The difference between the group means is measured by regressing this vector onto the vector of scores $\vec{y}$. The vectors $\vec{x}$ and $\vec{y}$ are angularly close when the group means differ and orthogonal when the means are identical.

A similar logic applies to the differences among $g > 2$ groups. The observations from the g groups are treated as a single collection of scores, expressed by the outcome vector $\vec{\mathsf{y}}$. To represent the group structure, $g-1$ dummy variables, $\vec{\mathsf{x}}_1$, $\vec{\mathsf{x}}_2$, ..., $\vec{\mathsf{x}}_{g-1}$, are defined. The space $\mathcal{V}_{\mathsf{X}}$ created by these dummy variables—in this context more often called an effect space than a regression space—contains all variation among the groups. The measured vector $\vec{\mathsf{y}}$ is projected using multiple regression onto the vector $\hat{\vec{\mathsf{y}}} \in \mathcal{V}_{\mathsf{X}}$, and the angle between $\vec{\mathsf{y}}$ and $\hat{\vec{\mathsf{y}}}$ is used to determine how closely the observations match the group structure. The projection operation, the measurement of agreement by such statistics as $R^2_{\mathsf{y}\cdot\mathsf{X}}$, and the test of significance are identical to those in other multiple regression problems. The new step is the construction of the dummy variables or vectors. Accordingly, this section primarily discusses how these dummy vectors are formed and used.

Dummy variables are characterized by two properties: first, a dummy variable has the same value for every member of a group; second, this value is not the same in every group. Any linearly-independent set of $g-1$ such variables can discriminate among g groups. For example, the differences among three groups can be represented by two variables with the values

$$\mathsf{X}_1 = \begin{cases} 1, & \text{for every subject in group 1,} \\ 0, & \text{for every subject in group 2,} \\ -1, & \text{for every subject in group 3,} \end{cases}$$

and

$$\mathsf{X}_2 = \begin{cases} 0, & \text{for every subject in group 1,} \\ 1, & \text{for every subject in group 2,} \\ -1, & \text{for every subject in group 3.} \end{cases}$$

Many other pairs of dummy vectors could represent the differences equally well.

The systematic construction of the vector space in which the differences lie, like that of the two-group space discussed in Section 3.4, starts with the a set of g dummy variables U_1, U_2, ..., U_g, each of which tags the members of one group:

$$\mathsf{U}_j = \begin{cases} 1, & \text{for every subject in group } j, \\ 0, & \text{for all other subjects.} \end{cases}$$

Whenever one of these variables equals 1, all the others are 0, so their dot product equals 0, and they are mutually orthogonal. The space $\mathcal{V}_{\mathsf{U}}$ that they span has dimension g. Any other dummy variable X can be written as a linear combination of the U_j; in vector terms,

$$\vec{\mathbf{X}} = a_1\vec{\mathbf{U}}_1 + a_2\vec{\mathbf{U}}_2 + \cdots + a_g\vec{\mathbf{U}}_g. \tag{8.1}$$

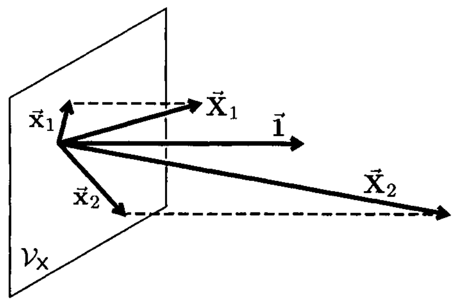


Figure 8.1: *Centering the two vectors $\vec{\mathbf{X}}_1$ and $\vec{\mathbf{X}}_2$ by projection onto the orthogonal complement of $\vec{\mathbf{1}}$.*

In particular, the two variables just mentioned are

$$\begin{aligned} \vec{\mathbf{X}}_1 &= \vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_3, \\ \vec{\mathbf{X}}_2 &= \vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_3. \end{aligned} \tag{8.2}$$

The $\vec{\mathbf{U}}_j$ are uncentered and the space $\mathcal{V}_{\mathsf{U}}$ contains the vector

$$\vec{\mathbf{1}} = \vec{\mathbf{U}}_1 + \vec{\mathbf{U}}_2 + \cdots + \vec{\mathbf{U}}_g,$$

as illustrated in Figure 3.2 for the two-group case. Between-group effects take place in the orthogonal complement of $\mathcal{V}_1$. Projecting the vectors $\vec{\mathbf{X}}_1$, $\vec{\mathbf{X}}_2$, ..., $\vec{\mathbf{X}}_{g-1}$ into $\mathcal{V}_{\perp 1}$ gives a space of dimension $g - 1$ that expresses only differences in means. This space, which will hereafter be denoted as a centered regression predictor space $\mathcal{V}_{\mathsf{X}}$, contains the centered predictors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_{g-1}$. Figure 8.1 shows how two uncentered dummy vectors $\vec{\mathbf{X}}_1$ and $\vec{\mathbf{X}}_2$ are centered. The component of these vectors along $\vec{\mathbf{1}}$ is removed, projecting them onto the centered space $\mathcal{V}_{\mathsf{X}}$ as the vectors $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$, respectively. The centered vectors express only differences among the groups and are sometimes known as *contrasts* or *comparisons.* Of course in practice, the projection into $\mathcal{V}_{\perp 1}$ occurs automatically when the regression intercept is fitted, but conceptually it is important to understand how it is done.

The observed data vector $\vec{\mathbf{Y}}$ is also partitioned into a part in the subspace $\mathcal{V}_1$ and a part orthogonal to it. One component lies in $\mathcal{V}_1$ and expresses the grand mean of all scores. The other component lies in $\mathcal{V}_{\perp 1}$ and embodies the between-group and between-subject differences. It is the centered data vector $\vec{\mathbf{y}}$ that is regressed onto $\mathcal{V}_{\mathsf{X}}$ to give $\hat{\vec{\mathbf{y}}}$.

Any linearly independent set of $g-1$ vectors that span the space $\mathcal{V}_X$ can be used to construct $\vec{\hat{\mathbf{y}}}$, to measure the quality of the fit by the angle $\angle(\vec{\mathbf{y}}, \vec{\hat{\mathbf{y}}})$, and to detect differences among the groups. Each such set of dummy vectors leads to same omnibus test of the null hypothesis that every group has the same mean. In this sense all definitions of the dummy variables are equivalent. However, it is easier to understand whatever differences among groups are found if each dummy variable is sensitive to a particular characteristic of the groups. When considered one variable at a time, all definitions are not equivalent, and one tries to define the dummy variables so that each one has its own distinct meaning.

The best way to define interpretable dummy variables is to think of patterns of means that one wants to detect and to assign the coefficients in Equation 8.1 to match these patterns. For example, consider the problem of detecting trend in a numerically classified set of groups. Suppose that a four-group study has been run in which the groups differ on a quantitative variable, the first group has one unit, the second group two units, and so forth. Assume that n subjects are observed in each group. Because the group classification is quantitative, one can examine the functional relationship between the numerical classification and the means, a procedure sometimes known as *trend analysis.* With four groups, there are three degrees of freedom among the means, so three dummy variables are needed to span $\mathcal{V}_X$. In a quantitative functions, the patterns of linear, quadratic, and cubic change are natural functions to look for. These forms are detected by three dummy variables, one with coefficients that are directly equal to the group numbers,

$$\vec{\mathbf{X}}_{\text{lin}} = \vec{\mathbf{U}}_1 + 2\vec{\mathbf{U}}_2 + 3\vec{\mathbf{U}}_3 + 4\vec{\mathbf{U}}_4,$$

one with coefficients that are squares of the group numbers,

$$\vec{\mathbf{X}}_{\text{quad}} = \vec{\mathbf{U}}_1 + 4\vec{\mathbf{U}}_2 + 9\vec{\mathbf{U}}_3 + 16\vec{\mathbf{U}}_4,$$

and one with cubic coefficients,

$$\vec{\mathbf{X}}_{\text{cube}} = \vec{\mathbf{U}}_1 + 8\vec{\mathbf{U}}_2 + 27\vec{\mathbf{U}}_3 + 64\vec{\mathbf{U}}_4.$$

The coefficients of these vectors, scaled to equal height, are plotted in the first row of Figure 8.2.

These three vectors are not orthogonal to $\mathcal{V}_1$; for example, the projection of $\vec{\mathbf{X}}_{\text{lin}}$ into $\mathcal{V}_1$ is

$$\left(\frac{\vec{\mathbf{1}}\cdot\vec{\mathbf{X}}_{\text{lin}}}{|\vec{\mathbf{1}}|^2}\right)\vec{\mathbf{1}} = \left(\frac{10n}{4n}\right)\vec{\mathbf{1}} = \tfrac{5}{2}\vec{\mathbf{1}}$$

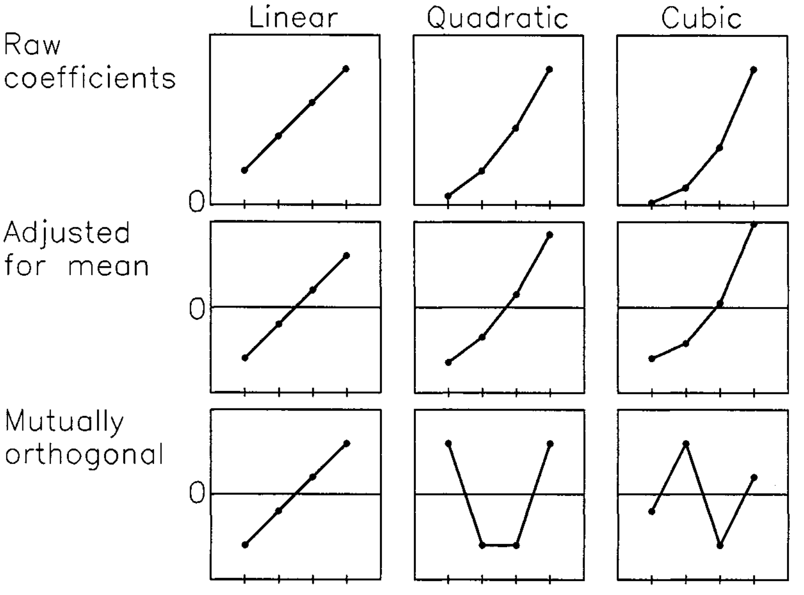


Figure 8.2: *Contrast coefficients for measuring trend in equally spaced groups. To emphasize the shapes, the three vertical axes in a row are scaled to give the plots the same height.*

(Equation 2.20). Subtracting the components along $\vec{\mathbf{1}}$ projects the vectors into the space $\mathcal{V}_X$ of centered vectors:

$$\begin{aligned}
\vec{\mathbf{x}}_{\text{lin}} &= \vec{\mathbf{X}}_{\text{lin}} - \tfrac{5}{2}\vec{\mathbf{1}} &&= -\tfrac{3}{2}\vec{\mathbf{U}}_1 - \tfrac{1}{2}\vec{\mathbf{U}}_2 + \tfrac{1}{2}\vec{\mathbf{U}}_3 + \tfrac{3}{2}\vec{\mathbf{U}}_4, \\
\vec{\mathbf{x}}_{\text{quad}} &= \vec{\mathbf{X}}_{\text{quad}} - \tfrac{15}{2}\vec{\mathbf{1}} &&= -\tfrac{13}{2}\vec{\mathbf{U}}_1 - \tfrac{7}{2}\vec{\mathbf{U}}_2 + \tfrac{3}{2}\vec{\mathbf{U}}_3 + \tfrac{17}{2}\vec{\mathbf{U}}_4, \\
\vec{\mathbf{x}}_{\text{cube}} &= \vec{\mathbf{X}}_{\text{cube}} - 25\vec{\mathbf{1}} &&= -24\vec{\mathbf{U}}_1 - 17\vec{\mathbf{U}}_2 + 2\vec{\mathbf{U}}_3 + 39\vec{\mathbf{U}}_4.
\end{aligned} \tag{8.3}$$

The coefficients of these vectors are plotted in the second row of Figure 8.2. The shapes of these functions are unchanged, but they are now centered on zero. Figure 8.3 shows these vectors in the three-dimensional space $\mathcal{V}_X$.

Although these vectors are safely located in $\mathcal{V}_X$, their interpretation is complicated by the fact that they are not mutually orthogonal—for example, $\angle(\vec{\mathbf{x}}_{\text{lin}}, \vec{\mathbf{x}}_{\text{quad}}) \approx 10°$. In Figure 8.2 the three functions in the middle row ascend from left to right, and in Figure 8.3 the three vectors point to the right. To distinguish among the effects, they are projected in separate subspaces, as described in Section 7.2. First one orders the vectors

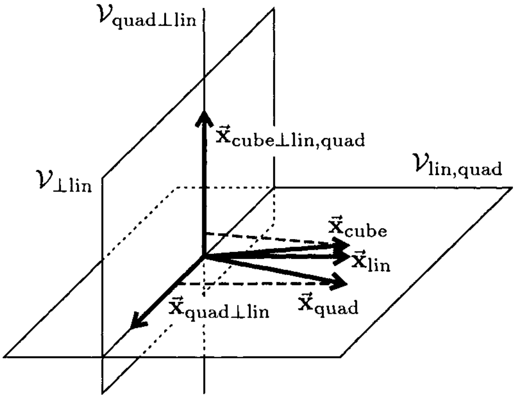


Figure 8.3: *The orthogonalization of the first three trend vectors.*

from most important to least important. Then the less important vectors are projected into subspaces orthogonal to the more important ones. With trend components, an explanation with small exponents is preferable to one with big exponents, so one orthogonalizes $\mathcal{V}_X$ by removing components in the order linear, quadratic, cubic. The vector $\vec{\mathbf{x}}_{\text{lin}}$ is fundamental and is used as it is. Next, the vector $\vec{\mathbf{x}}_{\text{quad}}$ is projected onto $\mathcal{V}_{\perp\text{lin}}$. Since only the direction of this projection is critical for the analysis, it can be scaled either to give the vector $\vec{\mathbf{x}}_{\text{quad}\perp\text{lin}}$ a length equal to that of $\vec{\mathbf{x}}_{\text{lin}}$ or to turn the coefficients into simple integers. Finally, the vector $\vec{\mathbf{x}}_{\text{cube}}$ is projected orthogonally to both $\vec{\mathbf{x}}_{\text{lin}}$ and $\vec{\mathbf{x}}_{\text{quad}}$. These projections are made by subtracting the components in the lower-order spaces from the higher-order vectors, and, after adjustment of the vector length to clear fractions from the coefficients, gives the *orthogonal polynomial contrast vectors*:

$$\begin{aligned}
\vec{\mathbf{x}}_{\text{lin}} &= -3\vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_2 + \vec{\mathbf{U}}_3 + 3\vec{\mathbf{U}}_4, \\
\vec{\mathbf{x}}_{\text{quad}\perp\text{lin}} &= \vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_3 + \vec{\mathbf{U}}_4, \qquad (8.4) \\
\vec{\mathbf{x}}_{\text{cube}\perp\text{lin,quad}} &= -\vec{\mathbf{U}}_1 + 3\vec{\mathbf{U}}_2 - 3\vec{\mathbf{U}}_3 + \vec{\mathbf{U}}_4.
\end{aligned}$$

These functions are shown in the final row of Figure 8.2. The three contrasts express systematic changes upward (or downward), in the curvature, and in the tendency to reverse direction. Although the greater portion of the original vectors $\vec{\mathbf{x}}_{\text{lin}}$, $\vec{\mathbf{x}}_{\text{quad}}$, and $\vec{\mathbf{x}}_{\text{cube}}$ has been eliminated, the residual components are considerably easier to interpret than are the nonorthogonalized vectors.

An orthogonal set of contrast vectors is convenient to work with because they can be fitted and tested individually. However, one need not actually orthogonalize a sequence of contrast vectors such as these to use them. Because the spaces are orthogonal, one can find the lengths of the components in the conditional subspaces by subtracting squared vector lengths or squared multiple regression coefficients (for example, as in Equation 7.5). For the polynomial trend analysis, one could fit three regression projections, one onto $\mathcal{V}_{\text{lin}}$, one onto $\mathcal{V}_{\text{lin,quad}}$, and one onto the full $\mathcal{V}_{\mathbf{X}}$. The squared lengths of the projections of $\vec{\mathbf{y}}$ into the orthogonal subspaces are given by

$$\begin{aligned}
|\hat{\vec{\mathbf{y}}}_{\text{lin}}|^2 &= R^2_{\mathbf{y}\cdot\text{lin}}|\vec{\mathbf{y}}|^2, \\
|\hat{\vec{\mathbf{y}}}_{\text{quad}\perp\text{lin}}|^2 &= |\hat{\vec{\mathbf{y}}}_{\text{lin,quad}} - \hat{\vec{\mathbf{y}}}_{\text{lin}}|^2 \\
&= (R^2_{\mathbf{y}\cdot\text{line,quad}} - R^2_{\mathbf{y}\cdot\text{lin}})|\vec{\mathbf{y}}|^2, \\
|\hat{\vec{\mathbf{y}}}_{\text{cube}\perp\text{lin,quad}}|^2 &= |\hat{\vec{\mathbf{y}}}_{\text{lin,quad,cube}} - \hat{\vec{\mathbf{y}}}_{\text{lin,quad}}|^2 \\
&= (R^2_{\mathbf{y}\cdot\text{lin,quad,cube}} - R^2_{\mathbf{y}\cdot\text{lin,quad}})|\vec{\mathbf{y}}|^2.
\end{aligned} \tag{8.5}$$

Subtraction of R^2 is particularly helpful when the functions are complicated or when the groups have unequal sizes.

8.2 Unequal sample sizes

The flexibility of the regression-based analysis of variance allows it to treat data from unequally sized groups in a way that the conventional formulation of the analysis of variance cannot do. The presence of groups of different sizes complicates the analysis because it changes the angular relationships among the effect subspaces. With unequal groups, a set of "orthogonal" contrasts such as Equations 8.4 may not be orthogonal with respect to the actual data. The analysis must resolve the ambiguities created by the nonorthogonality. Reflecting this need, these procedures are sometimes known as the *nonorthogonal analysis of variance.*

How one resolves the association among effects induced by unequal samples depends in part on how one plans to interpret the results. From the outset, it is necessary to distinguish between two approaches to the differences among the conditions. In the *equally weighted groups* approach, one treats the groups as distinct objects to be compared. The sizes of the samples used to measure them are unrelated to the hypotheses that one wants to test. In the *equally weighted subjects* approach, one views the number of subjects in each group as a reflection of an intrinsic characteristic of the population and wants to make inferences about a population with this distribution of sizes. In this analysis, each score has equal importance, although more weight is given to large groups than to small groups thereby. The terms *unweighted groups* and *unweighted subjects* also describe these approaches.[1] The goal of many psychological studies is to make general statements about the treatments or subpopulations that define the groups, and so equally weighting the groups is the more common approach. Which way to treat the data in a particular study depends on how they are to be interpreted and cannot be determined from the data alone.

If one only wants to know whether the groups differ, then unequal sample sizes pose no problem. The effect space spanned by a complete set of dummy vectors is the same no matter which individual vectors are used to span it. However, when one is interested in specific contrasts or effect subspaces defined by groups of dummy variables, the weighting of the groups and subjects affects the orthogonality relationships among the spaces. Consider a three-group experiment with the data shown in Table 8.1. Suppose that the researcher wishes to know whether conditions 1 and 2 differ and whether the average score in these conditions differs from that in condition 3. These questions are expressed by a pair of comparison vectors whose coefficients mirror the effects to be detected:

$$\begin{aligned} \vec{\mathbf{X}}_1 &= \vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_2, \\ \vec{\mathbf{X}}_2 &= \tfrac{1}{2}\vec{\mathbf{U}}_1 + \tfrac{1}{2}\vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_3. \end{aligned} \tag{8.6}$$

Had the three groups been the same size, these vectors would be orthogonal to each other and to the vector $\vec{\mathbf{1}}$. For the data in Table 8.1, however, neither orthogonality holds. For example, remembering that their definition makes $\vec{\mathbf{U}}_j \cdot \vec{\mathbf{U}}_j = n_j$ and $\vec{\mathbf{U}}_j \cdot \vec{\mathbf{U}}_k = 0$ for $j \neq k$,

$$\begin{aligned} \vec{\mathbf{X}}_1 \cdot \vec{\mathbf{X}}_2 &= (\vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_2) \cdot (\tfrac{1}{2}\vec{\mathbf{U}}_1 + \tfrac{1}{2}\vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_3) \\ &= \tfrac{1}{2}\vec{\mathbf{U}}_1 \cdot \vec{\mathbf{U}}_1 - \tfrac{1}{2}\vec{\mathbf{U}}_2 \cdot \vec{\mathbf{U}}_2 = \tfrac{1}{2}(n_1 - n_2), \end{aligned}$$

which is nonzero unless $n_1 = n_2$.

Consider first an ordinary regression analysis, which gives equal weight to each observation. First center the dummy variables. With the unequal group sizes as given, the component of $\vec{\mathbf{X}}_1$ along $\vec{\mathbf{1}}$ is removed by subtracting

[1] The equally weighted subjects method is sometimes called the *method of weighted means* or the *method of least squares*, although both approaches use least-squares estimation. The equally weighted means method is sometimes called the *method of unweighted means*, although this term also refers to an older approximation to this method. There are other ways to treat unequally sized groups that are not described here, some of which are obsolete.

Group	Scores						Sum	Mean
1	3	5					8	4
2	7	7	8	8	9	9	48	8
3	5	6	6	7			24	6

Table 8.1: *A small set of data with unequal group sizes.*

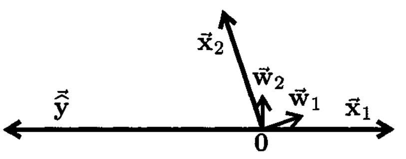


Figure 8.4: *The configuration of contrast vectors in $\mathcal{V}_X$ (from Equations 8.7 and 8.8) representing the groups in Table 8.1 .*

$-\frac{1}{3}\vec{\mathbf{1}}$, and the vector $\vec{\mathbf{X}}_2$ is already orthogonal to $\vec{\mathbf{1}}$. The centered vectors are

$$\begin{aligned} \vec{\mathbf{x}}_1 &= \tfrac{4}{3}\vec{\mathbf{U}}_1 - \tfrac{2}{3}\vec{\mathbf{U}}_2 + \tfrac{1}{3}\vec{\mathbf{U}}_3, \\ \vec{\mathbf{x}}_2 &= \tfrac{1}{2}\vec{\mathbf{U}}_1 + \tfrac{1}{2}\vec{\mathbf{U}}_2 - \vec{\mathbf{U}}_3. \end{aligned} \tag{8.7}$$

The configuration of these vectors is shown in Figure 8.4. These two vectors are not orthogonal. Their dot product is $\vec{\mathbf{x}}_1 \cdot \vec{\mathbf{x}}_2 = -2$, and the angle between them is 108.4°. A regression of $\vec{\mathbf{y}}$ onto $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ (using Equations 4.2) gives $b_1 = -2$ and $b_2 = 0$—in calculating these coefficients it helps to remember that $\vec{\mathbf{U}}_j \cdot \vec{\mathbf{Y}}$ equals the sum of the scores in group j and that because the $\vec{\mathbf{x}}_j$ are orthogonal to $\vec{\mathbf{1}}$, the projections of $\vec{\mathbf{Y}}$ onto them are the same whether $\vec{\mathbf{Y}}$ is centered or not. Thus, $\vec{\hat{\mathbf{y}}} = -2\vec{\mathbf{x}}_1$, and the vector $\vec{\mathbf{y}}$ lies directly over the space generated by $\vec{\mathbf{x}}_1$ (it points in the opposite direction). The angle between $\vec{\mathbf{y}}$ and $\vec{\hat{\mathbf{y}}}$ is 28.7°. Although the best predictor here does not require $\vec{\mathbf{x}}_2$, an examination of the angles in Figure 8.4 shows that $\vec{\mathbf{x}}_2$ and $\vec{\hat{\mathbf{y}}}$ are not orthogonal—the angle between them is 71.6°. This angle reflects the fact that when groups 1 and 2 are pooled, the mean of the eight scores they contain is 7, which does not equal the mean of 6 in group 3. Thus, in this subject-weighted view, there is evidence for the difference among means expressed by the second contrast.

This result is inconsistent with the equally weighted groups perspective. Both the average of the means of the first two groups (4 and 8) and the third group's mean are exactly 6. On these grounds, one would expect $\vec{\mathbf{x}}_2$ and

$\vec{\mathbf{y}}$ to be orthogonal, although the calculation in the last paragraph shows that they are not so. An analysis in which they are orthogonal is created by adjusting the length of the vectors $\vec{\mathbf{U}}_j$ so that the groups have equal importance and the projections of these vectors onto other vectors do not involve the sample size. Instead of unit components for each observation in the group vectors $\vec{\mathbf{U}}_j$, one uses components of length $1/n_j$, creating the new vectors

$$\vec{\mathbf{U}}_j^{\mathsf{W}} = \frac{1}{n_j}\vec{\mathbf{U}}_j$$

(the superscript w here and on other quantities below indicates that this weighting has been applied). Contrast vectors such as Equation 8.6 are rewritten with the new basis vectors as

$$\begin{aligned}\vec{\mathbf{w}}_1 &= \vec{\mathbf{U}}_1^{\mathsf{W}} - \vec{\mathbf{U}}_2^{\mathsf{W}},\\ \vec{\mathbf{w}}_2 &= \tfrac{1}{2}\vec{\mathbf{U}}_1^{\mathsf{W}} + \tfrac{1}{2}\vec{\mathbf{U}}_2^{\mathsf{W}} - \vec{\mathbf{U}}_3^{\mathsf{W}},\end{aligned} \tag{8.8}$$

or for the sample sizes in Table 8.1,

$$\begin{aligned}\vec{\mathbf{w}}_1 &= \tfrac{1}{2}\vec{\mathbf{U}}_1 - \tfrac{1}{6}\vec{\mathbf{U}}_2,\\ \vec{\mathbf{w}}_2 &= \tfrac{1}{4}\vec{\mathbf{U}}_1 + \tfrac{1}{12}\vec{\mathbf{U}}_2 - \tfrac{1}{4}\vec{\mathbf{U}}_3.\end{aligned}$$

Because the original coefficients of these contrasts were chosen to sum to zero, the $\vec{\mathbf{w}}_j$ are orthogonal to $\vec{\mathbf{1}}$, and thus they span $\mathcal{V}_{\mathsf{X}}$ without centering. They are also shown in Figure 8.4.

A regression using the $\vec{\mathbf{w}}_j$ gives the best-fitting vector in the unweighted groups interpretation:

$$\vec{\hat{\mathbf{y}}} = -\tfrac{20}{3}\vec{\mathbf{w}}_1 + \tfrac{8}{3}\vec{\mathbf{w}}_2.$$

Since the vectors $\vec{\mathbf{w}}_1$ and $\vec{\mathbf{w}}_2$ span the same space as $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$, this vector is identical to $\vec{\hat{\mathbf{y}}} = -2\vec{\mathbf{x}}_1$ fitted using the $\vec{\mathbf{x}}_j$. The fit in $\mathcal{V}_{\mathsf{X}}$, as measured by $\angle(\vec{\mathbf{y}}, \vec{\hat{\mathbf{y}}})$, is equally good for either set. However, the interpretation at the contrast level differs. As Figure 8.4 shows, $\vec{\mathbf{w}}_2 \perp \vec{\hat{\mathbf{y}}}$, indicating that there is no difference between $\overline{X}_3$ and $(\overline{X}_1 + \overline{X}_2)/2$ at the group level.

Neither the unweighted subjects vectors $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ nor the unweighted means vectors $\vec{\mathbf{w}}_1$ and $\vec{\mathbf{w}}_2$ is an orthogonal pair. Although the original questions expressed by Equation 8.6 seem to be independent, the unequal sample sizes induce dependencies in either approach. This nonorthogonality raises all the problems of ambiguous attribution discussed in Chapters 5 and 7. Often the best way to treat this association is simply to acknowledge it when describing the analysis. However, when one has a good a priori reason to order the explanatory importance of the questions, they can be examined conditionally. If the first question is primary, then $\vec{\mathbf{w}}_1$ or $\vec{\mathbf{x}}_1$ would be examined first, followed by a conditional look at the second question using the projection of $\vec{w}_2$ or $\vec{x}_2$ orthogonal to the first vector. In more complex designs, this type of orthogonalization may be applied to the subspaces spanned by sets of vectors.

Although the unweighted groups analysis may give substantially different results from the unweighted subjects analysis, one approach is not more correct than the other. Both sets of contrasts express legitimate questions about the scores. The analysis with equally weighted individuals examines hypotheses that respect the variations in group size. The analysis with equally weighted groups examines hypotheses about the groups as abstract types, unrelated to their sample frequencies. Either approach may be appropriate for a particular research problem. Of course, one is not free to choose between them based on their fit to the data. Usually an investigator has one type of question or the other in mind from the start.

8.3 Factorial designs

Most analysis-of-variance designs have a richer organization of the groups than that discussed in the previous sections. In a *factorial design*, the groups are classified along two or more dimensions or *factors*, with every level of one factor appearing with every combination of the levels of the other factors. These designs allow a researcher to investigate both the separate *main effect* of each factor on the outcome variable and the way that the effects of one factor are modulated by the levels of the other factors. The latter effects, known as *interactions*, are of central interest in many studies.

The factorial analysis of variance partitions the differences among the means into a series of effects, a main effect for each factor in the design and an interaction for each combination of factors. A three-factor design with factors labeled A, B, and C has seven effects, three main effects named after the factors and four interactions, AB, AC, BC, and ABC. From a geometric point of view, each of these effects corresponds to a subspace of the effect space, and the effects are measured by projecting the score vector $\vec{y}$ into these spaces.

The first step in the analysis is to identify the subspaces associated with the various effects. Consider a two-factor design with two levels of factor A and three levels of factor B. Denote this structure by a pair of subscripts that give the level of factor A and the level of factor B, respectively; for example, the population mean μ_{12} refers to the group with the first level of A and the second level of B. This design has two main effects, A and B, and one interaction AB. The effect space $\mathcal{V}_X$ contains three subspaces associated with these effects, $\mathcal{V}_{\mathsf{a}}$, $\mathcal{V}_{\mathsf{b}}$, and $\mathcal{V}_{\mathsf{ab}}$. The component of $\vec{\mathbf{y}}$ in each of these spaces is used to measure the magnitude of the corresponding effect.

The main effects are spanned by sets of vectors that are identical to those that span a one-way set of groups, with the coefficients held constant across the other factor. For example, a main effect of A in the 2×3 design is present when there is a difference between the means of the two levels over B, that is, when

$$\tfrac{1}{3}(\mu_{11}+\mu_{12}+\mu_{13}) \qquad \text{differs from} \qquad \tfrac{1}{3}(\mu_{21}+\mu_{22}+\mu_{23}).$$

Expressed as a dummy vector, this contrast is

$$\tfrac{1}{3}\vec{\mathbf{U}}_{11}+\tfrac{1}{3}\vec{\mathbf{U}}_{12}+\tfrac{1}{3}\vec{\mathbf{U}}_{13}-\tfrac{1}{3}\vec{\mathbf{U}}_{21}-\tfrac{1}{3}\vec{\mathbf{U}}_{22}-\tfrac{1}{3}\vec{\mathbf{U}}_{23}.$$

The component of $\vec{\mathbf{y}}$ parallel to this vector is long when an A effect is present and short when one is not. It is convenient to clear the fractions here by tripling this vector's length without altering its direction, giving the dummy vector

$$\vec{\mathbf{X}}_{\mathsf{a}} = \vec{\mathbf{U}}_{11}+\vec{\mathbf{U}}_{12}+\vec{\mathbf{U}}_{13}-\vec{\mathbf{U}}_{21}-\vec{\mathbf{U}}_{22}-\vec{\mathbf{U}}_{23}. \tag{8.9}$$

An association between this dummy vector and $\vec{\mathbf{y}}$ indicates an A effect.

Factor B has three levels, so that the space $\mathcal{V}_{\mathsf{b}}$ is two dimensional. It is spanned by any pair of noncollinear vectors constructed so that all subjects at a give level of B have the same value. A common choice is the difference between pairs of means mentioned above in Equations 8.2 and here expressed by

$$\begin{aligned}\vec{\mathbf{X}}_{\mathsf{b}_1} &= \vec{\mathbf{U}}_{11}+\vec{\mathbf{U}}_{21}-\vec{\mathbf{U}}_{13}-\vec{\mathbf{U}}_{23},\\ \vec{\mathbf{X}}_{\mathsf{b}_2} &= \vec{\mathbf{U}}_{12}+\vec{\mathbf{U}}_{22}-\vec{\mathbf{U}}_{13}-\vec{\mathbf{U}}_{23}.\end{aligned} \tag{8.10}$$

The interaction space $\mathcal{V}_{\mathsf{ab}}$ is also two dimensional. The 2×3 design has six groups, so $\dim(\mathcal{V}_{\mathsf{X}}) = 5$, and three of these are occupied with main effects, leaving two for the interaction. The interaction space could be constructed from any pair of noncollinear vectors with components outside $\mathcal{V}_{\mathsf{a}}$ and $\mathcal{V}_{\mathsf{b}}$. A convenient way to create these interaction vectors is to take the coefficient-wise product of each spanning vector from $\mathcal{V}_{\mathsf{a}}$ with each spanning vector from $\mathcal{V}_{\mathsf{b}}$. Matching Equation 8.9 with each member of Equations 8.10 and multiplying the coefficients gives the pair of vectors

$$\begin{aligned}\vec{\mathbf{X}}_{\mathsf{ab}_1} &= \vec{\mathbf{U}}_{11}-\vec{\mathbf{U}}_{21}-\vec{\mathbf{U}}_{13}+\vec{\mathbf{U}}_{23},\\ \vec{\mathbf{X}}_{\mathsf{ab}_2} &= \vec{\mathbf{U}}_{12}-\vec{\mathbf{U}}_{22}-\vec{\mathbf{U}}_{13}+\vec{\mathbf{U}}_{23}.\end{aligned} \tag{8.11}$$

The full space of effects in the 2×3 design is spanned by the five vectors in Equations 8.9, 8.10, and 8.11. When the sample sizes are equal, these five vectors are mutually orthogonal and orthogonal to $\vec{1}$. Each effect can be examined independently of the others. The component of $\vec{y}$ in the space $\mathcal{V}_{\mathsf{a}}$ is orthogonal to the components in $\mathcal{V}_{\mathsf{b}}$ and $\mathcal{V}_{\mathsf{ab}}$. The orthogonality simplifies the calculation by allowing separate regressions in each of the effect subspaces. It also makes the analysis easier to interpret. The formulae presented in traditional analysis-of-variance texts reflect these simplifications. However, when the group sizes are unequal, the three subspaces may not be orthogonal. One must then decide whether to weight the groups or the subjects equally and, whichever choice is made, must deal with any ambiguities arising from the nonorthogonalities.

The choice between equally weighted individuals analysis and equally weighted groups corresponds to the use of the vectors $\vec{\mathbf{U}}_{jk}$ and $\vec{\mathbf{U}}^{\mathsf{W}}_{jk}$ in Equations 8.9, 8.10, and 8.11. Whichever choice is made, the three subspaces $\mathcal{V}_{\mathsf{a}}$, $\mathcal{V}_{\mathsf{b}}$, and $\mathcal{V}_{\mathsf{ab}}$ are unlikely to be orthogonal. To the extent that the spaces are angular close, it may be impossible to unambiguously attribute a difference in the means to one factor or the other. The ambiguity is most serious when the group classifications are appreciably correlated. For example, suppose that in a 2×2 design most subjects in the first category of factor A also are in the first category of factor B, and similarly for the second categories, so that the group sizes are

	b_1	b_2
a_1	10	4
a_2	4	10

This pattern of frequencies creates an association between the two main-effect spaces, with $\angle(\vec{\mathbf{x}}_{\mathsf{a}}, \vec{\mathbf{x}}_{\mathsf{b}}) = 64.6°$ and $\angle(\vec{\mathbf{w}}_{\mathsf{a}}, \vec{\mathbf{w}}_{\mathsf{b}}) = 115.4°$, both about 25° off orthogonal. With this degree of association one may find that a main effect is clearly present, but that it is impossible to find out whether factor A, factor B, or some combination of them is responsible.

One way to render the effects orthogonal is to use conditional tests. For example, if factor A has priority over factor B, then one can examine effects in the orthogonal subspaces $\mathcal{V}_{\mathsf{a}}$ and $\mathcal{V}_{\mathsf{b}\perp\mathsf{a}}$. Such tests are often run, but should be backed by a strong justification for the order of orthogonalization. The ordering cannot be determined from the data and must derive from outside information. Giving priority to the space that has the smallest angle to $\vec{y}$ (i.e., the largest simple F statistic), as is sometimes done, makes no sense here. As in the comparable situation with continuous predictors, the best that one may be able to do is to recognize the existence of the ambiguity and plan to resolve it in a later investigation.

At first glance the equal weighting of the groups appears harder to do,

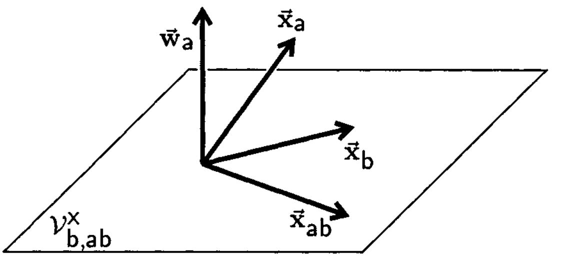


Figure 8.5: *The effect space for a 2×2 design showing the unweighted groups dummy vector $\vec{w}_a$ and the dummy vectors $\vec{x}_a$, $\vec{x}_b$, and $\vec{x}_{ab}$ from an unweighted subjects analysis.*

since it requires weights that are particular to the experiment. However, it is possible to capitalize on some orthogonality relationships among the two types of dummy vectors to run the unweighted groups analysis using the dummy vectors $\vec{x}_j$ constructed with equal subject weighting. The key observation here is that when the coefficients of the dummy vectors are such as to make them orthogonal in an analysis with equal sample sizes, then each effect subspace for the unweighted groups is orthogonal to the subspaces for every other effect in the unweighted subjects analysis. In a two-factor analysis, the subspace $\mathcal{V}^{w}_{a}$ for the unweighted groups A effect is orthogonal to the subspaces $\mathcal{V}^{x}_{b}$ and $\mathcal{V}^{x}_{ab}$ for the unweighted subjects B and AB effects. Figure 8.5 shows the relationship among the vectors in a 2×2 design with unidimensional subspaces. The vector $\vec{w}_a$ is orthogonal to both $\vec{x}_b$ and $\vec{x}_{ab}$, and thus to the space spanned by these vectors. Equivalent relationships, which are not shown, put $\vec{w}_b$ orthogonal to $\vec{x}_a$ and $\vec{x}_{ab}$ and $\vec{w}_{ab}$ orthogonal to $\vec{x}_a$ and $\vec{x}_b$. Similar orthogonality relationships can be observed in Figure 8.4, where $\vec{w}_1 \perp \vec{x}_2$ and $\vec{w}_2 \perp \vec{x}_1$.

In more detail, the equally weighted groups vector for the A effect in the 2×3 design is

$$\vec{w}_a = \frac{\vec{U}_{11}}{n_{11}} + \frac{\vec{U}_{12}}{n_{12}} + \frac{\vec{U}_{13}}{n_{13}} - \frac{\vec{U}_{21}}{n_{21}} - \frac{\vec{U}_{22}}{n_{22}} - \frac{\vec{U}_{23}}{n_{23}}.$$

As can be seen by computing the dot product with Equations 8.10 and 8.11, this vector is orthogonal to $\mathcal{V}^{x}_{b}$ and $\mathcal{V}^{x}_{ab}$. Then, using the conditional tests described in Section 7.2, the projection of $\vec{y}$ in $\mathcal{V}^{w}_{a}$ is found by subtracting the projection of $\vec{y}$ into $\mathcal{V}^{x}_{b,ab}$ from the projection into the full effect space:

$$\vec{\hat{y}}^{w}_{a} = \vec{\hat{y}}^{x} - \vec{\hat{y}}^{x}_{b,ab}.$$

Expressing the squared vector lengths with the multiple correlation coefficient, the sum of squares needed to test the effect is

$$SS_{\mathsf{A}}^{\mathsf{W}} = |\vec{\mathsf{y}}_{\mathsf{a}}^{\mathsf{W}}|^2 = |\vec{\mathsf{y}}|^2 (R^2_{\mathsf{y}\cdot\mathsf{a,b,ab}} - R^2_{\mathsf{y}\cdot\mathsf{b,ab}}). \tag{8.12}$$

Comparable expressions exist for the other effects, both in the two-way design and in higher designs. Many computer programs operate this way.

When interpreting an analysis of variance done in this way, one should remember that although Equation 8.12 has the form of a conditional test, it is equivalent to the unconditional test of a direct effect within the context of the equally weighted groups analysis. The subtraction in Equation 8.12 does not mean that the B and AB effects have been "removed" from the A effect. The conditional appearance is simply a convenient computational trick.

8.4 The analysis of covariance

The analysis of covariance is similar to the analysis of variance in the hypotheses that it investigates. It extends that technique by allowing the effects of an irrelevant source of variability to be extracted from the data before the difference among groups is tested. If successful, this adjustment improves both the stability and the statistical reliability of the results. The procedure depends on the presence of an ancillary variable that is correlated with the variable whose means are being examined. The fundamental idea is to use the projection operations to eliminate the systematic component associated with this variable from the error vector.

Consider a test for the difference between the mean of the variable Y in two groups. Represent the two groups by the dummy variable X. Now suppose that within each group Y is substantially correlated with a second variable Z, known as the *covariate*. The upper panel of Figure 8.6 shows an example of such data as a scatterplot in variable space. The two groups of scores overlap considerably on the outcome variable Y, which is plotted on the ordinate, although for any given value of the covariate Z on the abscissa the groups are clearly separated. The variability of Y at any particular value of Z is much less than the unconditional variability. The analysis of covariance exploits the differences between the conditional and unconditional variability by comparing the difference between groups to the variability of Y at a fixed value of Z instead of to the overall variability. In effect, the analysis looks at the differences of the scores relative to the pooled regression line from the two groups, placed between the groups in the figure.

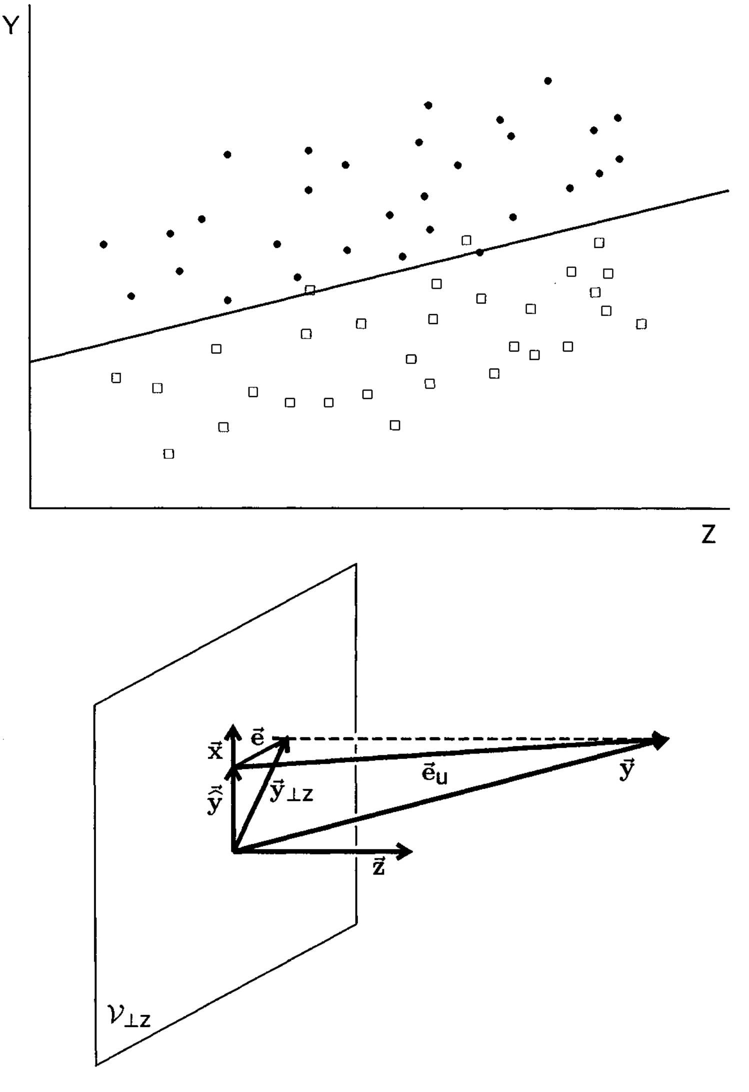


Figure 8.6: *The analysis of covariance of variable* Y *with covariate* Z *in variable space (upper panel) and subject space (lower panel).*

The lower panel of Figure 8.6 shows the subject-space version of this picture. Three vectors define the problem: the dummy vector $\vec{x}$ that represents the groups, the outcome vector $\vec{y}$, and the covariate vector $\vec{z}$. The two measured variables Y and Z are associated, as shown by the acute angle between $\vec{y}$ and $\vec{z}$. The dummy vector $\vec{x}$ is not orthogonal to the dependent-variable vector $\vec{y}$, giving evidence of a difference between the group means, although that angle is large and the uncorrected error of prediction $\vec{e}_U$ is substantial. The dummy vector $\vec{x}$ is essentially unrelated to the covariate $\vec{z}$. In part, the reason that the relationship between $\vec{x}$ and $\vec{y}$ is poor is because an appreciable component of $\vec{y}$ lies in the direction of $\vec{z}$. This dimension has nothing to do with $\vec{x}$, but it contributes to $\vec{e}_U$. To improve the relationship, the analysis of covariance removes the component of $\vec{e}_U$ along $\vec{z}$ by projecting $\vec{y}$ into the subspace $\mathcal{V}_{\perp Z}$ orthogonal to the covariate. When the analysis is effective, as it is in Figure 8.6, the projection $\vec{y}_{\perp Z}$ is angularly closer to $\vec{x}$ then is $\vec{y}$. The closer association does not arise because the projection on $\vec{x}$ has changed—note that both $\vec{y}$ and $\vec{y}_{\perp Z}$ lead to the same $\hat{\vec{y}}$. Instead, the shift of $\vec{y}$ along a dimensional orthogonal to $\vec{x}$ increases the association between the grouping and the scores.

The projection of the score vector in the analysis of covariance is the complement of the projections that are made in conditional tests with multiple regression or the analysis of variance. In those techniques the projection operations remove effects on the predictor side of the equation. In the analysis of covariance, variation is removed from the outcome side. One of these operations does not preclude the other. The same gamut of conditional tests discussed in the first portion of this chapter is available on the dummy-variable side of the analysis of covariance. In particular, a two-factor analysis of covariance with unequal sample sizes would both remove the covariate's effect from the outcome variable by projecting $\vec{y}$ into $\mathcal{V}_{\perp Z}$ and adjust the dummy vectors within $\mathcal{V}_X$ to define appropriately weighted main effect and interaction variables. The latter steps follow the procedures and interpretation discussed in the earlier parts of this chapter and can be considered without reference to the adjustment of $\vec{y}$.

The illustration in Figure 8.6 is something of an idealization, in that the dummy vector $\vec{x}$ is shown as perfectly orthogonal to the covariate $\vec{z}$. In real data the presence of sampling variability makes perfect orthogonality unlikely—the group means for the covariate almost certainly differ somewhat. This nonorthogonality makes the projection $\hat{\vec{y}}_{\perp Z}$ of the vector $\vec{y}_{\perp Z}$ onto $\vec{x}$ slightly different from the projection $\hat{\vec{y}}$ of the original $\vec{y}$. The difference between these vectors is small when the covariate is essentially unrelated to the groups. It corresponds to adjusting the group means in variable space by projecting them parallel to the regression line to a common abscissa, usually the grand mean of all groups. Whereas the regression projection $\vec{\vec{y}}$ of $\vec{y}$ onto $\vec{x}$ represents the difference between the original means, the regression projection $\vec{\vec{y}}_{\perp z}$ represents the difference between these *adjusted means.* It is usual to report the adjusted means as part of an analysis of covariance, since these are the quantities whose differences the procedure tests.

The configuration in Figure 8.6 illustrates the analysis of covariance in its least ambiguous form. The projection operation eliminates the effect of an extraneous variable from the residual, but the component of $\vec{y}$ along the grouping vector $\vec{x}$ is similar in the conditional and the unconditional analyses—as noted, both $\vec{y}$ and $\vec{y}_{\perp z}$ project to the same vector $\vec{\vec{y}}$. Application of the analysis of covariance is more problematic when the projection operation changes the relationship of the scores to the predictor. Figure 8.7 shows a particularly salient example of this situation. The primary change from Figure 8.6 is that the covariate is no longer unrelated to the grouping. In the variable-space portion of the figure, note that although the values of Y are somewhat larger for the squares than for the dots, the dots generally lie above the common within-group regression line while the squares generally lie below it. The adjusted means, shown by the projections to the grand mean of Z in subject space, have the opposite order from the simple means. In the unconditional analysis, the mean of squares is greater than that of dots, while in the conditional analysis, which uses the position of the scores relative to the regression line, this relationship is reversed. The analysis of covariance tests the latter effect. The meaning of the adjusted difference in such a case is usually obscure, for it contradicts the simple relationship of the scores. Not least, it depends critically on the linearity of the adjustment line and its extrapolation from the centers of the distributions.

The subject-space picture in the lower part of Figure 8.7 shows this geometry. The association between the grouping and the covariate makes $\vec{x}$ and $\vec{z}$ nonorthogonal. Because of this association, when $\vec{y}$ is projected into $\mathcal{V}_{\perp z}$, both the component orthogonal to $\vec{x}$ and the component parallel to it change. Indeed, in the extreme case illustrated here, the association to $\vec{x}$ reverses. In the two small diagrams at the bottom right, which show the plane containing $\vec{x}$ and $\vec{y}$ and the plane containing $\vec{x}$ and $\vec{y}_{\perp z}$, the angle changes from acute to obtuse.

The ambiguity introduced by the conditional analysis pictured in Figure 8.7 is particularly egregious. Usually the adjustment does not reverse the outcome. However, as the example points out, whenever $\vec{x}$ and $\vec{z}$ are not orthogonal the component of $\vec{y}$ along $\vec{x}$ differs from the component of $\vec{y}_{\perp z}$ along $\vec{x}$. If the difference is appreciable, then the adjusted means are difficult to interpret. These difficulties are worse when more than two groups

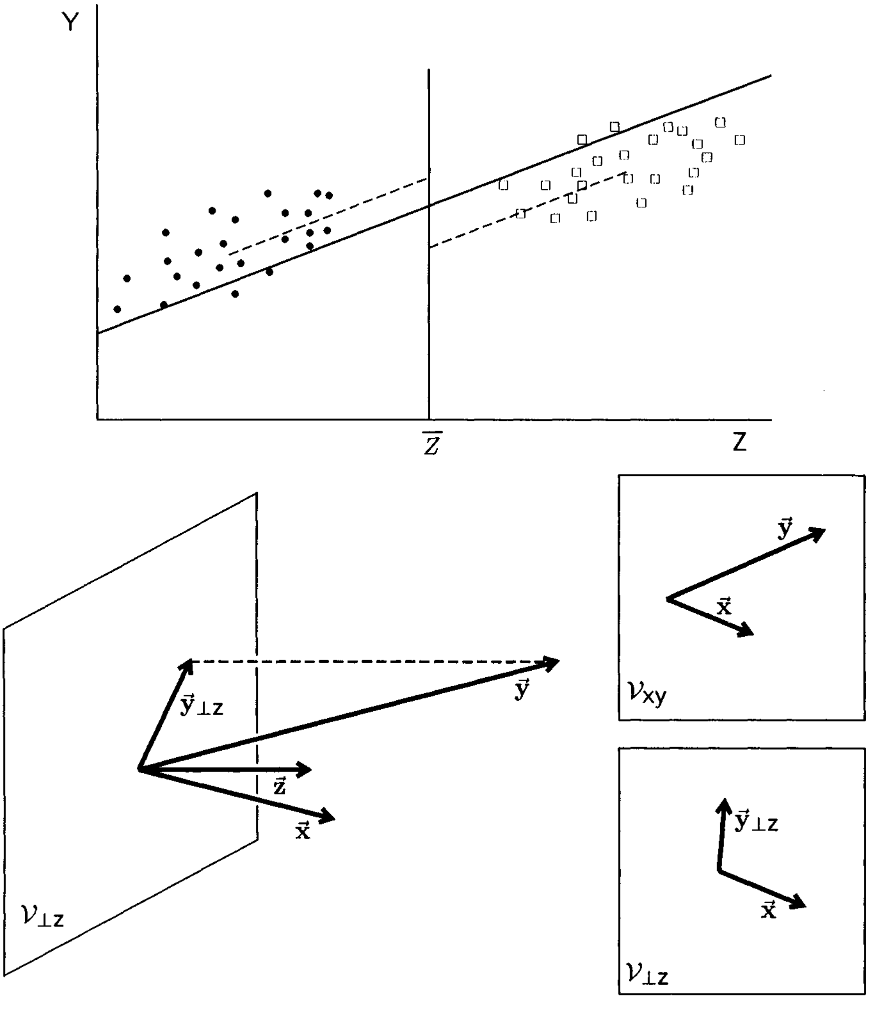


Figure 8.7: *Representations of an analysis of covariance when the covariate is strongly related to the grouping. The dashed lines in the top panel project the group means to $\overline{Z}$. The small panels at the lower right show the relationships between $\vec{x}$ and $\vec{y}$ and between $\vec{x}$ and $\vec{y}_{\perp z}$.*

are present, so that the space $\mathcal{V}_X$ is multidimensional. Even when the ordering of the groups is unchanged by the adjustment, the relationship of $\vec{y}$ to the contrast vectors spanning $\mathcal{V}_X$ may change in confusing ways. The analysis of covariance can best be used to eliminate accidental variability and to increase the statistical reliability of an effect, by reducing the size of $\vec{e} = \vec{y} - \vec{\hat{y}}$, not by changing $\vec{\hat{y}}$. The adjustment is unambiguous only when the covariate vector $\vec{z}$ is essentially orthogonal to the grouping vector $\vec{x}$.

For the analysis of covariance to apply accurately, it is essential that the relationship between Y and Z be the same in every group. The pooled variability is used to determine the adjustments made by the analysis, and these adjustments are inaccurate when the groups have different structures. As an extreme example, suppose that Y and Z are correlated with $r = -0.5$ in one group and $r = +0.8$ in another. When these groups are pooled, Y and Z will probably have a small positive correlation, but this correlation does not accurately represent the association in either group. Adjustments and projections made using the pooled association are not meaningful, and an analysis of covariance is inappropriate.

Differences in the covariance structure are most easily observed in a variable-space plot of the data. Scatterplots of homogeneous groups have the same size and orientation, except for sampling variation. In subject space, a homogeneous structure implies that the vectors $\vec{y}^{(k)}$ and $\vec{z}^{(k)}$ calculated from the kth group alone are similar for all k. They are also similar to the common vectors $\vec{y}_{\perp X}$ and $\vec{z}_{\perp X}$ that result from projecting $\vec{y}$ and $\vec{z}$ onto the orthogonal complement of the group structure defined by $\mathcal{V}_X$. This operation amounts to removing any differences among the group means and is similar to the illustration of partial correlation in Figure 7.1. Various procedures have been developed to investigate whether the covariance structure in a set of groups is homogeneous and are described in other multivariate texts.

Exercises

1. Verify the coefficients of $\vec{x}_{\text{quad}\perp\text{lin}}$ in Equation 8.4 by subtracting from $\vec{x}_{\text{quad}}$ its projection onto $\vec{x}_{\text{lin}}$.

2. Suppose that the three groups of an analysis of variance design are defined by a numerical factor that takes the values 1, 2, and 4. What are the contrasts that express linear and quadratic trend on these unequally spaced groups? Turn these contrasts into orthogonal vectors in distinct spaces as in Equations 8.4, and draw a picture of the effect space showing the vectors before and after orthogonalization.

3. Data are collected for Problem 2 as follows

Group 1	5	2	2							
Group 2	7	11	6	5	9	7	5			
Group 4	11	9	10	12	15	11	13	10	11	8

a. How do the differing sizes of the groups alter the placement of the vectors $\vec{\mathrm{x}}_{\mathrm{lin}}$ and $\vec{\mathrm{x}}_{\mathrm{quad}\perp\mathrm{lin}}$ from Problem 2?

b. Calculate the unweighted means vectors $\vec{\mathrm{w}}_{\mathrm{lin}}$ and $\vec{\mathrm{w}}_{\mathrm{quad}\perp\mathrm{lin}}$. What is their relationship to $\vec{\mathrm{x}}_{\mathrm{lin}}$ and $\vec{\mathrm{x}}_{\mathrm{quad}\perp\mathrm{lin}}$? Draw a sketch showing these vectors.

c. Run the unweighted means analysis for the data above. Give the regression equation.

4. List the effects and degrees of freedom for the following designs. Give equations comparable to Equation 8.12 that express the sums of squares for an unweighted means analysis in terms of correlations involving the unweighted subjects vectors.

a. A two-way design in which factor A has 3 levels and is crossed with factor B with 2 levels.

b. A three-way design in which factors A, B, and C have 3, 2, and 4 levels, respectively.

5. Suppose that the group sizes in a 2×2 design are

	b_1	b_2
a_1	20	3
a_2	8	12

a. Define dummy variables for the A, B, and AB effects in an unweighted individuals analysis and find the angles between $\vec{\mathrm{x}}_{\mathrm{a}}$, $\vec{\mathrm{x}}_{\mathrm{b}}$, and $\vec{\mathrm{x}}_{\mathrm{ab}}$.

b. Do the same for an unweighted groups analysis and the vectors $\vec{\mathrm{w}}_{\mathrm{a}}$, $\vec{\mathrm{w}}_{\mathrm{b}}$, and $\vec{\mathrm{w}}_{\mathrm{ab}}$.

c. How long are the vectors $\vec{\mathrm{U}}_{jk}$ and $\vec{\mathrm{U}}^{\mathrm{W}}_{jk}$?

d. Without worrying about vector length, locate the vectors $\vec{\mathrm{x}}_{\mathrm{a}}$, $\vec{\mathrm{x}}_{\mathrm{b}}$, and $\vec{\mathrm{x}}_{\mathrm{ab}}$ from the unweighted individuals analysis with pointers in three-dimensional space. Where is the vector $\vec{\mathrm{w}}_{\mathrm{a}}$? Give an example of a vector $\vec{\mathrm{y}}$ for which the unweighted subjects effect is larger than the unweighted groups effect and of a vector for which the reverse is true.

6. One reason to adopt the regression procedure for testing for differences among groups is that it allows one to use both categorical and continuous predictors simultaneously. Consider the following miniature data set consisting of two variables X and Y measured in two groups:

Group 1	(1, 3)	(4, 5)	(5, 6)	(3, 5)	(3, 5)	(2, 6)
Group 2	(5, 4)	(7, 4)	(4, 2)	(4, 3)	(7, 3)	(3, 2)

a. Define a group dummy vector Z and analyze these data using X and Z as predictors of Y. Draw a vector diagram illustrating the configuration of the variables.

b. How does this analysis differ from an analysis of covariance? In particular, how does the conditional test of Z given X differ from an analysis of covariance using X as a covariate? Illustrate the differences with diagrams.

7. Modify Figure 8.6 so that in subject space the group means are not identical and in variable space $\vec{x}$ and $\vec{z}$ are not exactly orthogonal. Make the modifications small, such as might arise from sampling accidents when there is no true group-covariate relationship. In variable space show the projections that create the adjusted means. In subject space show projections $\vec{\vec{y}}$ and $\vec{\vec{y}}_{\perp z}$. These vectors will be similar but not identical.

Chapter 9

Principal-component analysis and factor analysis

This chapter turns from the predictor-outcome procedures of multiple regression to the analysis of a single set of variables. When there are substantial correlations among the members of a set of variables, that set carries, in a sense, less information than it would appear from the number of variables it contains. Near multicollinearity is an example of this phenomenon. Although p vectors may be measured, they lie very close to a space of fewer than p dimensions. Both principal-component analysis and factor analysis take a collection of variables, examine its correlational structure, and extract its principal dimensions of variation. By reducing a large set of variables to a smaller one, they can locate patterns in the data and considerably simplify the subsequent analysis of the variables.

9.1 Principal-component vectors

A linearly independent set of p variables X_1, X_2, ..., X_p has p dimensions of variation, and the corresponding vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$ span a p-dimensional space $\mathcal{V}_{\mathsf{X}}$. Whenever the correlations among the variables are nonzero, these vectors do not point uniformly throughout the space $\mathcal{V}_{\mathsf{X}}$, but are concentrated more in some directions than in others. Thus, there is a larger horizontal component than a vertical component in the pair of correlated vectors

In many sets of real data, the actual variability is concentrated in only a few dimensions within the space $\mathcal{V}_{\mathsf{X}}$. For example, suppose that one has recorded the scores of a group of students on three measures of verbal ability, V_1, V_2, and V_3. These measures are correlated and have much in common. Because they vary together, the systematic component of their variability is enhanced by summing them to create a new variable

$$\mathsf{S} = \mathsf{V}_1 + \mathsf{V}_2 + \mathsf{V}_3.$$

In contrast, because individuals tend to have comparable scores on the three parts, there is much less systematic variability in the differences between pairs of tests:

$$\mathsf{D}_1 = \mathsf{V}_1 - \mathsf{V}_2 \qquad \text{and} \qquad \mathsf{D}_2 = \mathsf{V}_2 - \mathsf{V}_3.$$

If the sampling variability is large, these two variables will be very noisy and unreliable. The variables S, D_1, and D_2 span the same space as V_1, V_2, and V_3, but they do so in a different way. When working with such data, one may decide to combine the original variables into the reliable variable S and to ignore the unstable variable D_1 and D_2.

Where the content of the variables dictates a way to combine them, as it does with the variables V_1, V_2, and V_3, forming the new variables is easy. One calculates the meaningful combinations and analyzes them. However, sometimes one lacks a theoretical basis for the combination and wishes to derive it from the observed structure of the variables. *Principal-component analysis* uses the correlations among a set of variables to reformulate them so as to concentrate their variability into as few dimensions as possible.

Figure 9.1 illustrates this transformation with a pair of variables in subject space. A two-dimensional space is spanned by the correlated vectors $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$. These vectors are replaced by a new pair of vectors $\vec{\mathbf{u}}_1$ and $\vec{\mathbf{u}}_2$. The new vectors are orthogonal, and they are chosen so that the longer vector $\vec{\mathbf{u}}_1$ points in the direction in which the original vectors vary the most and the shorter vector $\vec{\mathbf{u}}_2$ points in the direction in which they vary the least. The variables corresponding to these *principal-component vectors* are known as *principal components*. Because they span the same space as the original vectors, they describe the same effects. However, by capturing the greatest part of the variation in the first component, they represent the original variables in a more ordered and convenient way. Not least, the orthogonality of $\vec{\mathbf{u}}_1$ and $\vec{\mathbf{u}}_2$ makes them easier to use in a multiple regression than $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$.

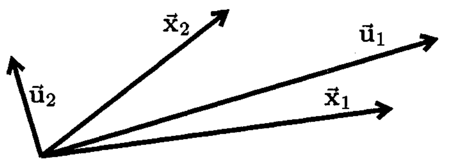


Figure 9.1: *A two-dimensional principal-component analysis. The vectors* $\vec{x}_1$ *and* $\vec{x}_2$ *are replaced by the vectors* $\vec{u}_1$ *and* $\vec{u}_2$ *spanning the same space.*

The idea illustrated in Figure 9.1 guides a principal-component analysis in any number of dimensions. The variables X_1, X_2, ..., X_p are replaced by a more convenient set U_1, U_2, ..., U_p. The new variables are linear combinations of the old; in vector form,

$$\vec{u}_k = a_{1k}\vec{x}_1 + a_{2k}\vec{x}_2 + \cdots + a_{pk}\vec{x}_p.$$

To hold the new vectors to the same scale as the original set, the sum of the squared coefficients is fixed to unity:

$$a_{1k}^2 + a_{2k}^2 + \cdots + a_{pk}^2 = 1. \tag{9.1}$$

The new variables are constructed so that they have four important properties.

1. No information is lost by the change of variables. Geometrically, the $\vec{u}_k$ span the same space as the $\vec{x}_j$.

2. The two sets of variables express the same total variability. In statistical terms, the sums of the two sets of variances are identical:

   $$\mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_p) = \mathrm{var}(U_1) + \mathrm{var}(U_2) + \cdots + \mathrm{var}(U_p).$$

   Geometrically, the sums of the squared lengths are the same:

   $$|\vec{x}_1|^2 + |\vec{x}_2|^2 + \cdots + |\vec{x}_p|^2 = |\vec{u}_1|^2 + |\vec{u}_2|^2 + \cdots + |\vec{u}_p|^2.$$

3. The derived variables are uncorrelated. Geometrically, their vectors are orthogonal. Thus, each vector represents a distinct aspect of the original space.

4. Each derived variable captures as much of the variability of the X_j as possible. Variable U_1 is chosen so that it has the largest possible variance and $\vec{u}_1$ is as long as possible (subject to Equation 9.1), variable U_2 has the next largest variance (subject both to Equation 9.1 and orthogonality to U_1), and so forth.

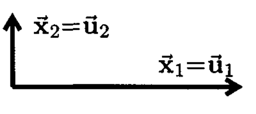
These four properties determine the length and position of the $\vec{u}_k$. The maximum-length criterion is central here, since it distinguishes the principal-component vectors from other orthogonal bases. The combination of conditions, particularly maximum length and orthogonality, puts another constraint on the coefficients; the coefficients of any two different principal components, treated as vectors, are orthogonal:

$$\vec{a}_j \cdot \vec{a}_k = a_{1j}a_{1k} + a_{2j}a_{2k} + \cdots + a_{pj}a_{pk} = 0. \tag{9.2}$$

Conceptually, it is helpful to think of finding the principal-component vectors one after the other. Start with the set $\mathcal{U}$ containing all vectors formed by linear combinations of the $\vec{x}_j$ whose coefficients satisfy Equation 9.1. To find the first principle component $\vec{u}_1$, one searches through $\mathcal{U}$ and picks the longest vector—if there are several candidates, any will do. Now one confines one's attention to the orthogonal complement of $\vec{u}_1$, that is, to the vectors in $\mathcal{U}$ that are orthogonal to $\vec{u}_1$. The longest of these is the second principal-component vector $\vec{u}_2$. For the third principal component, one further restricts the search to vectors that are orthogonal to both $\vec{u}_1$ and $\vec{u}_2$. The process of restriction and selection continues until $\mathcal{U}$ is exhausted. Because the space of candidate vectors loses one dimension at each step, the number of vectors that can be extracted equals the dimension of the space $\mathcal{V}_X$ spanned by the $\vec{x}_j$. Normally, a set of p vectors gives p principal-component vectors, but fewer are found when the $\vec{x}_j$ are linearly dependent and $\dim(\mathcal{V}_X) < p$.

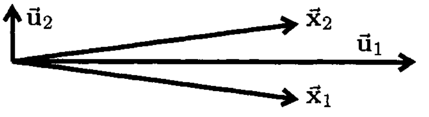
To understand how the structure of the original variables determines the principal-component vectors, it is helpful to look at several two-dimensional configurations of vectors to see how the position and length of the principal-component vectors depend on $\vec{x}_1$ and $\vec{x}_2$.

- When $\vec{x}_1 \perp \vec{x}_2$, a principal-component analysis is unnecessary. The variables already have the desired structure. If $|\vec{x}_1| \neq |\vec{x}_2|$, then $\vec{u}_1$ aligns with the larger of the two and $\vec{u}_2$ aligns with the shorter:



  When $|\vec{x}_1| = |\vec{x}_2|$, the configuration is circularly symmetrical, and the principal-component vectors have no preferred direction. Any pair of orthogonal vectors with lengths equal to that of the $\vec{x}_j$ satisfies the criteria.

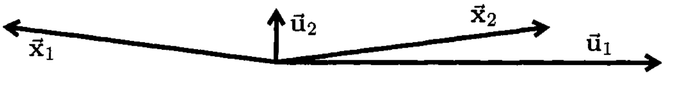
- When $\vec{x}_1$ and $\vec{x}_2$ are equally long and positively correlated, they are combined with equal weights. The first principal-component vector lies exactly between the original vectors, and the second is at right angles to it:



With the constraint on the sum of squares (Equation 9.1), the vectors are

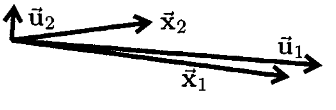
$$\vec{u}_1 = \frac{\vec{x}_1}{\sqrt{2}} + \frac{\vec{x}_2}{\sqrt{2}} \qquad \text{and} \qquad \vec{u}_2 = \frac{\vec{x}_1}{\sqrt{2}} - \frac{\vec{x}_2}{\sqrt{2}}. \tag{9.3}$$

The length of these vectors are $\sqrt{1 \pm \cos\angle(\vec{x}_1, \vec{x}_2)}$ times the common length of the original vectors. When the original vectors are negatively correlated, the sign of one coefficient in Equation 9.3 is reversed, giving the configuration



The principal-component vectors $\vec{u}_1$ and $\vec{u}_2$ are the same as in the positively correlated configuration.

- When one vector is longer than the other, it receives more weight in the combination. The first principal-component vector $\vec{u}_1$ is angularly closer to the longer of the original vectors than it is to the shorter:



In detail, the squared lengths of the principal-component vectors are

$$|\vec{u}_j|^2 = \left(\frac{|\vec{x}_1|^2 + |\vec{x}_2|^2}{2}\right)^2 \pm \sqrt{\left(\frac{|\vec{x}_1|^2 - |\vec{x}_2|^2}{2}\right)^2 + (\vec{x}_1 \cdot \vec{x}_2)^2}, \tag{9.4}$$

and the angles between the new and old vectors are

$$\cos\angle(\vec{x}_j, \vec{u}_k) = \sqrt{\frac{(\vec{x}_1 \cdot \vec{x}_2)^2 |\vec{u}_k|^2}{|\vec{x}_j|^2 \left[(\vec{x}_1 \cdot \vec{x}_2)^2 + (|\vec{u}_k|^2 - |\vec{x}_j|^2)^2\right]}}. \tag{9.5}$$

In most applications the squared vector lengths are variances and the dot product is a covariance.

The criteria that define the principal components determine their length and the line along which they lie, but not the sign of their direction. Either the vectors shown above or their 180° reversal are satisfactory principal components. An orientation that points in the same direction as one or more of the $\vec{\mathbf{x}}_k$ is often most natural, but either direction satisfies the mathematical criteria. Equation 9.5 gives the smaller angle (i.e., the positive cosine), and subtraction from 180° may be needed to fit with the particular orentation of $\vec{\mathbf{u}}_k$.

The particular configuration of principal-component vectors that one obtains is partly determined by the length of the original vectors, that is, by the variance of the original variables. In many circumstances, these lengths are determined by the units of measurement and are irrelevant to the variables' fundamental meaning. For example, a score recorded as the number of correct answers has both a larger mean and a larger variance than one measured as a proportion of correct responses, yet both express the same fundamental quantity. The larger variance gives the frequency measure a greater influence in a principal-component analysis than the proportion measure. To avoid biasing the outcome, one should be sure that all the variables are measured on comparable scales. When all variables are of the same type—for example, when all are proportions of correct responses—the scales may already be compatible and the variables can be analyzed as they stand. When the variables are measures of different types of quantities, the only way to put them on a similar footing may be to equate their variances by replacing the original observations with standard scores. All the vectors then have the same length. After standardization, the structure extracted by the principal-component analysis is determined entirely by the directions of the vectors, not by their lengths. Only the correlations are used in the analysis, not the covariances or standard deviations. This aspect of the variables—their angles or intercorrelations—is usually the most fundamental, and most principal-component analyses begin by converting the variables to standard form.

Calculating the coefficients of the linear combinations that create the principal components requires one to maximize $|\vec{\mathbf{u}}_j|$ subject to the constraints of Equations 9.1 and 9.2. Although the principles underlying this maximization are clear from the geometry, the numerical operations are less obvious. The conventional approach to the problem of maximizing the vector length is algebraic. There are standard methods in linear algebra for solving these problems. Numerically, the coefficients of the principal-component variables are known as the *eigenvectors* of the covariance matrix (also called characteristic vectors or latent vectors). The variances are known as *eigenvalues* (also called latent value or latent roots, or as characteristic values or roots). This terminology, which is not particularly

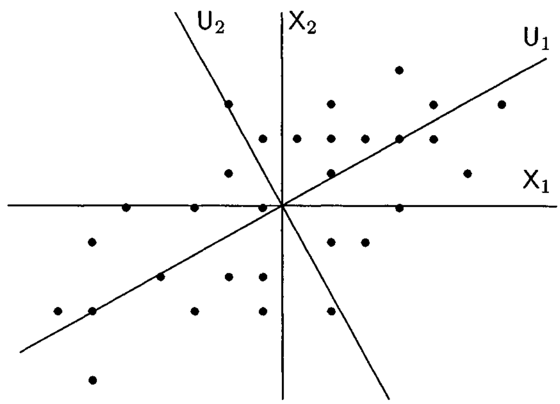


Figure 9.2: *A principal-component analysis in variable space. The rotated coordinates* U_1 *and* U_2 *align with the axis of the data distribution.*

relevant to this aspect of the geometry, often appears in discussions of principal-component analysis.

9.2 Variable-space representation

In some ways, the picture of principal component analysis in variable space is more obvious than it is subject space. In variable space, a pair of correlated variables is represented by a scatterplot with a cluster of points elongated in a diagonal direction, as in Figure 9.2. A principal-component analysis replaces the X_1 and X_2 axes by a rotated pair of axes U_1 and U_2 that are aligned with the elliptical cluster. Sometimes one or more axes are also reversed (i.e., reflected about zero). The rotation is made in such a way as to put U_1 along the longest axis of the ellipse and U_2 along the shorter axis. The cluster of points is unchanged; the only changes in the picture are in the axis on which they are represented. The rules for axis rotation given in Section 2.1 determine the coefficients. Equations 9.1 and 9.2 impose the same constraints as Equations 2.12, and the coefficients a_{ij} are those of the axis-defining vectors $\vec{a}_j$ for the variable-space rotation.

In more than two dimensions, normally distributed points in a variable-space scatterplot roughly form an ellipsoid or the higher-dimensional equivalent of an ellipsoid. A transformation to principal components is an axis rotation in this multidimensional space. The rotation lines the first new axis up with the longest dimension of the ellipsoid, lines the second new axis up with the next longest dimension, and so forth. As in the two-dimensional picture, the final axis is aligned with the shortest dimension of the configuration.

A comparison of principal-component extraction in subject space and variable space (Figures 9.1 and 9.2, respectively) points up the differences between the two representations. In variable space, the principal-component operation corresponds to a rotation of the axes to reexpress the points. In subject space, original vectors are not rotated, but are replaced by an equivalent set that spans the same space in a more convenient way. As usual, the variable-space picture emphasizes the pattern of the individual scores, and the subject-space picture emphasizes the relationships among the variables.

A picture such as Figure 9.2 tempts one to think of the line defined by the first principal axis as a regression line. This identification is wrong. The first principal axis is not the same as the regression line, either the one that predicts X_2 from X_1 or the one that predicts X_1 from X_2. The difference between regression and principal-component analysis is shown in Figure 9.3. The top portion of the figure shows both subject-space and variable-space representations of a one-predictor regression. In subject space the regression vector $\hat{\vec{\mathbf{y}}}$ lies in the space of the predictor variables. Although it is placed as close as possible to the outcome variable, it only expresses the variability of the predictors. In variable space, the regression predictor $\widehat{\mathsf{Y}}$ minimizes the sum of squared distances from the points to the regression line in a direction parallel to the Y axis, as shown by the dotted lines attached to a few of the points. The bottom portion of the figure shows the same pictures for a principal-component analysis. In subject space, the first principal-component vector combines both variables and lies between the two original vectors, outside the space spanned by either vector alone. In variable space the first principal axis is the major axis of an ellipse that roughly encloses the scattering of points. It is created by minimizing the squared deviations of the points in a direction perpendicular to the axis, again as shown by the dotted lines. As the subject-space picture shows, the first principal-component vector is always closer to $\vec{\mathbf{y}}$ than the regression predictor vector. In variable space, the regression line is always flatter than the principal axis, another manifestation of the phenomenon of *regression to the mean.*

9.3 Simplifying the variables

The full set of principal-component vectors $\vec{\mathbf{u}}_1, \vec{\mathbf{u}}_2, \ldots, \vec{\mathbf{u}}_p$ spans the same space $\mathcal{V}_{\mathsf{X}}$ as the original $\vec{\mathbf{x}}_j$. As a set, the new vectors may provide a

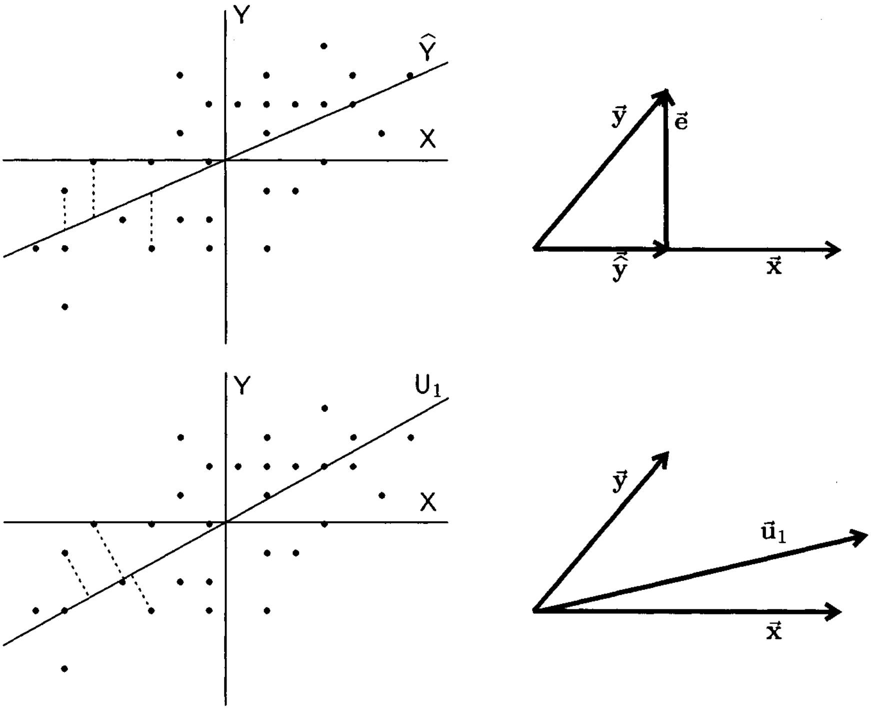


Figure 9.3: *A comparison of the first principal component and the regression line. The upper panels show regression; the lower, principal-component analysis. The left side illustrates subject space; the right, variable space. In variable space, the dotted lines show a few of the deviations whose sum of squares is minimized to fit the line.*

more convenient or rational basis for the space, but they do not change or restrict it in any way. However, the ordering of the principal-component vectors by length (i.e., by the variance of the principal-component variables) usually makes the first few principal components capture the bulk of the variability of the X_j. If this is so, then one can confine one's attention to the first few components without losing much of the original variability. In effect, one says that although the $\vec{x}_j$ span a space $\mathcal{V}_X$ of dimension p, all their important variation occurs in a subspace $\mathcal{V}_{pc} \subset \mathcal{V}_X$ of smaller dimensionality. For most applications, one can work almost as well in $\mathcal{V}_{pc}$ as in $\mathcal{V}_X$. Usually the smaller space is the more convenient. One can conduct any further multivariate analysis with a few orthogonal variables instead of many correlated variables.

Consider a regression problem with many correlated predictors. This analysis is subject to the problems of near multicollinearity described in Section 5.2, leading to unstable and inaccurate estimates of the coefficients. Now a principal-component analysis is run and the set of predictors is reduced to a few orthogonal variables that carry most of the original variability. A regression analysis that uses these variables has the simplicity and stability of the orthogonal analysis of Section 5.3.

Deciding how many variables to include in the reduced set is always a difficult matter. On the one hand, the point of the process is to simplify things by throwing out as many principal components as possible; on the other hand, each component that is dropped potentially eliminates some important information. Unfortunately, there are no surefire statistical tests to guide one here. Even when one makes the assumption of a normal sampling distribution, the distribution of any but the simplest (and least helpful) statistics is too complex to be useful. A helpful heuristic is a graph of the squared lengths of the principal-component vectors relative to the sum of squared lengths, $|\vec{\mathbf{u}}_j|^2/\sum|\vec{\mathbf{u}}_l|^2$. Using this graph, one tries to retain enough components to encompass the bulk of the total squared length, but without using components whose individual lengths are little different from those that follow it. The latter components are mainly determined by accidental variability and are unlikely to describe anything fundamental. Generally, the noisier the data, the fewer components are stable and substantial.

The simplification provided by selecting a few principal components for further analysis is no panacea, however. It is based on the supposition that sets of variables with concordant variability are more important than variables with isolated variability. This supposition may be wrong for a particular problem. A principal-component simplification risks passing over a unique variable that is nonetheless important in a later step of the analysis. The principal-component operation is most valuable when the variables have much in common and when these common elements dominate the effects of sampling noise, as with the three variables S, D_1, and D_2 mentioned at the start of this chapter.

To ease the interpretation of the reduced set of variables, one sometimes forms further combinations from them within the reduced space $\mathcal{V}_{\mathrm{pc}}$. The idea is to respan $\mathcal{V}_{\mathrm{pc}}$ with a new set of orthogonal vectors that are easier to interpret. These new variables are created by the same type of transformation that was used to create the principal components—an axis rotation in variable space and a respanning that preserves the total squared vector length in subject space. The new vectors are linear combination of the preserved $\vec{\mathbf{u}}_k$, subject to the restrictions of Equations 9.2 and 9.1. In this way, a set of ten original variables might be reduced to two by a principal-component analysis, then these two rewritten as two different variables that make greater intuitive sense. Although such a transformation loses the ordering of the variances that characterizes the original principal components, it can produce derived variables whose meaning is easier to understand. The major difficulty here is developing a set of principles that determine how to make the respannings or rotations. Several rotation algorithms of greater or lesser complexity and usefulness are available. One possibility is to attempt to orient the rotated vectors so that they have a "simple" relationship to the original vectors, for example, by giving the $\vec{x}_j$ only large or small projections onto the new $\vec{u}_k$. Another possibility is to adjust them to conform as nearly as possible to a different set of variables, perhaps deriving from previous research.

When using any of these reorientation procedures, including principal-component analysis itself, one should remember that the mathematical criteria that place a vector in $\mathcal{V}_{\mathrm{pc}}$ makes no use of the variables' meaning. Each transformation procedure optimizes some mathematical quantity, such a variance, but cannot assure that that criterion has any real application to the problem. As throughout statistical analysis, one should not accept any such mathematical result unquestioningly. Only a careful examination of the relationship of the derived variables to the original variables can determine whether a given configuration is an advance over the original set. One's caution should increase as one moves farther and farther from the original data.

9.4 Factor analysis

Like principal-component analysis, *factor analysis* is a procedure in which a structure of reduced dimensionality is extracted from a single set of variables. However, whereas principal-component analysis does this by respanning the space of the original variables with a simpler configuration, factor analysis formulates its space by fitting a model of common and unique parts to the observed pattern of vectors. There are many varieties of factor analysis. An important distinction is made between *exploratory factor analysis*, which takes a set of variables and extracts a structure from it, and *confirmatory factor analysis*, which examines a set of variables to see whether they are consistent with a proposed structure. This section looks briefly at the geometric representation of a simple form of exploratory factor analysis known as *common factor analysis*.

To begin, consider a pair of standardized vectors $\vec{x}_1$ and $\vec{x}_2$ representing two correlated variables. The standardization gives these vectors the same

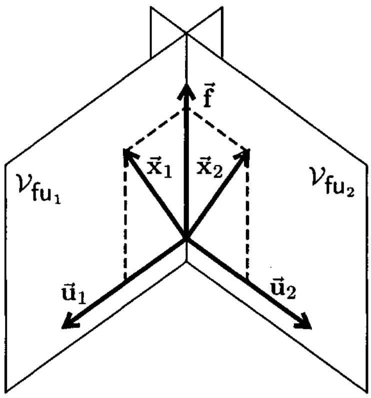


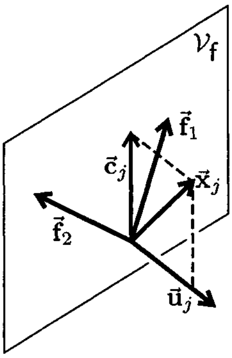
Figure 9.4: *Two observed variables decomposed into common and unique portions in a factor analysis.*

length, conveniently set to unity. Now ask how the correlation between $\vec{x}_1$ and $\vec{x}_2$ came about. One possibility is that it derives entirely from an unobserved latent variable that influences both their values. Each variable also has a unique component, so that the correlation between them is imperfect. Vector $\vec{x}_1$ is a linear combination of a vector $\vec{f}$ shared by both variables and a vector $\vec{u}_1$ unique to the first variable. Similarly, $\vec{x}_2$ combines a component along $\vec{f}$ with a different unique vector $\vec{u}_2$. The vector $\vec{f}$ is a *factor* derived from the two $\vec{x}_j$. Figure 9.4 illustrates the relationship of the variables. The two vectors $\vec{x}_1$ and $\vec{x}_2$ of equal length lie at an acute angle to each other. The vectors $\vec{f}$, $\vec{u}_1$, and $\vec{u}_2$ have the same unit length as $\vec{x}_1$ and $\vec{x}_2$ and are mutually orthogonal. Thus, each unique vector shares nothing with the common vector or with the unique vector of the other variable. The observed vectors lie in the subspaces generated by $\vec{f}$ and one of the unique vectors. Specifically, $\vec{x}_1$ lies in the space $\mathcal{V}_{\mathsf{fu}_1}$ generated by $\vec{f}$ and u_1, and $\vec{x}_2$ lies in the space $\mathcal{V}_{\mathsf{fu}_2}$ generated by $\vec{f}$ and $\vec{u}_2$:

$$\vec{x}_1 = a\vec{f} + \sqrt{1-a^2}\,\vec{u}_1 \qquad \text{and} \qquad \vec{x}_2 = a\vec{f} + \sqrt{1-a^2}\,\vec{u}_2. \tag{9.6}$$

The coefficient a is chosen to reproduced the observed correlation between X_1 and X_2. The three vectors of the model are mutually orthogonal, so $\vec{x}_1{\cdot}\vec{x}_2 = a^2\vec{f}{\cdot}\vec{f}$, and, as the vectors have unit length, a equals the correlation of X_1 and X_2, a result that answers Problem 2.4.

With two original vectors, one can reproduce the observed correlation with a single common factor and a unique component for each variable, as



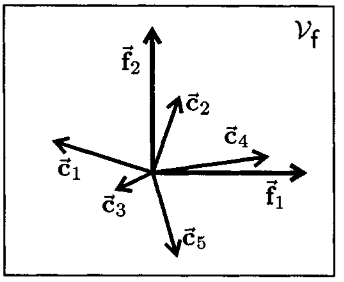


Figure 9.5: *A two-dimensional factor space. The left panel shows the space $\mathcal{V}_f$ and the vector $\vec{x}_j$ with its unique part $\vec{u}_j$. The right panel shows the projections of five variables onto $\mathcal{V}_f$.*

in Equations 9.6. With more variables, it may be necessary to postulate two or more factor dimensions to accommodate the observed correlations among the variables. One then can think of a *factor space* $\mathcal{V}_f$ spanned by m factor vectors $\vec{f}_1$, $\vec{f}_2$, ..., $\vec{f}_m$. At least in the initial portion of the analysis, these vectors are orthogonal. For each observed variable $\vec{x}_j$, there is a unique vector $\vec{u}_j$ that is orthogonal to all the $\vec{f}_k$ and to every other unique variable. Each observed variable is the sum of components in the factor space and a unique component:

$$\vec{x}_j = a_{j1}\vec{f}_1 + a_{j2}\vec{f}_2 + \cdots + a_{jm}\vec{f}_m + d_j\vec{u}_j. \tag{9.7}$$

Figure 9.5 illustrates the relationships when $m = 2$ factors are present. The orthogonal factor vectors $\vec{f}_1$ and $\vec{f}_2$ span a two-dimensional factor space $\mathcal{V}_f$, indicated by the plane in the left panel. This plane plays a role that is comparable to that of $\vec{f}$ in Figure 9.4. Orthogonal to this space and to each other are unique vectors $\vec{u}_1$, $\vec{u}_2$, ..., $\vec{u}_p$. Only one of these vectors is shown here, so that the entire diagram is equivalent to one of the planes in Figure 9.4. The vector $\vec{x}_j$ is decomposed into a component $\vec{c}_j$ in $\mathcal{V}_f$ and a component along $\vec{u}_j$, corresponding to its common and unique parts, respectively.

Each observed vector is represented by a picture such as the left panel of Figure 9.5 containing a vector $\vec{c}_j$ in the factor space $\mathcal{V}_f$ and a unique vector $\vec{u}_j$. A complete picture of an m-factor fit to p observed variables has $m+p$ dimensions. Visualization is not that difficult, however, since each unique vector $\vec{\mathbf{u}}_j$ is orthogonal to everything else except $\vec{\mathbf{x}}_j$ and can be considered separately. For interpretation it suffices to look at the factor space $\mathcal{V}_{\mathsf{f}}$ and the vectors $\vec{\mathbf{c}}_j$ and $\vec{\mathbf{f}}_k$ lying within it. With two extracted factors, the picture is planar, no matter how many original variables there are; for example, with five variables one might have the configuration in the right panel of Figure 9.5.

The length of the common component $\vec{\mathbf{c}}_j$ of $\vec{\mathbf{x}}_j$ indicates the extent to which the variability of X_j is shared with the other variables. This length is easily determined. Write the factor model of Equation 9.7 as a sum of common and unique parts,

$$\vec{\mathbf{x}}_j = \vec{\mathbf{c}}_j + d_j\vec{\mathbf{u}}_j,$$

with

$$\vec{\mathbf{c}}_j = a_{j1}\vec{\mathbf{f}}_1 + a_{j2}\vec{\mathbf{f}}_2 + \cdots + a_{jm}\vec{\mathbf{f}}_m.$$

The orthogonality of the factor structure and the unit length of the factor vectors means that

$$\begin{aligned} |\vec{\mathbf{c}}_j|^2 &= |a_{j1}\vec{\mathbf{f}}_1|^2 + |a_{j2}\vec{\mathbf{f}}_2|^2 + \cdots + |a_{jm}\vec{\mathbf{f}}_m|^2 \\ &= a_{j1}^2 + a_{j2}^2 + \cdots + a_{jm}^2. \end{aligned} \tag{9.8}$$

The squared length of $\vec{\mathbf{c}}_j$ is known as the *communality*. It is the systematic part of the representation and is the proportion of the variance of the $\vec{\mathbf{x}}_j$ that can be attributed to the factor structure. The length of the unique vector is

$$|d_j\vec{\mathbf{u}}_j|^2 = |\vec{\mathbf{x}}_j|^2 - |\vec{\mathbf{c}}_j|^2,$$

so that the proportion of the variability that cannot be explained by the factor structure is

$$d_j = \sqrt{1 - a_{j1}^2 - a_{j2}^2 - \cdots - a_{jm}^2}. \tag{9.9}$$

When trying to give meaning to the factors, it helps to look at the angles they make to the original vectors. On the right of Figure 9.5, the factor $\vec{\mathbf{f}}_1$ is much like the common portion of $\vec{\mathbf{x}}_4$, and $\vec{\mathbf{f}}_2$ is most closely related to $\vec{\mathbf{x}}_2$ and $\vec{\mathbf{x}}_4$, although the directions are reversed. The orthogonality of the factor structure means that the angular loadings are identical to the coefficients of the factors:

$$\begin{aligned} \vec{\mathbf{x}}_j \cdot \vec{\mathbf{f}}_k &= (a_{j1}\vec{\mathbf{f}}_1 + a_{j2}\vec{\mathbf{f}}_2 + \cdots + a_{jm}\vec{\mathbf{f}}_m + d_j\vec{\mathbf{u}}_j) \cdot \vec{\mathbf{f}}_k \\ &= a_{jk}\vec{\mathbf{f}}_k \cdot \vec{\mathbf{f}}_k = a_{jk} \end{aligned} \tag{9.10}$$

The equality of the coefficients and the angular loadings is a consequence of the orthogonality built into the factor model. It does not hold for multivariate techniques such as multiple regression where the vectors that make up the linear combinations are not necessarily orthonormal (for example, recall the discussion in Section 4.3).

The goal of a factor analysis is not simply to represent a set of variables by their common and unique parts. One wants to do so in a low-dimensional factor space. To investigate this possibility, one fits models with m-dimensional factor spaces for various values of m and chooses a representation that is a compromise between simplicity and a good fit to the observed correlations among the X_j. A low-dimensional solution, it is hoped, may point to the fundamental structure of the set of variables. As with the retention of principal components, the choice of m cannot be reduced to an algorithm and depends in part on the domain being studied and the investigator's intentions.

Once the factor space is obtained, the original set of factors that span it can be abandoned and new vectors constructed that span it in more convenient ways. The respanning of the space is an important stage of the typical factor analysis. Simple rotation, as mentioned in the previous section, is applicable here, as are a variety of other rotation procedures, both orthogonal and oblique. Typically, the transformations are designed to give the new factor variables a simple or otherwise comprehensible relationship to the original variables. The structure may also be turned to conform with some externally derived criteria, such as by matching the vectors of a factor solution obtained from other data. The choice of which type of rotation to use depends in part on the content of the analysis, and is as much an art as a science. The extensive lore on factor rotation should be consulted.

Computationally, factor analysis is complex. One must simultaneous estimate the number of factors to be used, the communalities of the observed variables, and the coefficients of the linear combinations that describe the systematic portion of each variable. These parameters are interrelated and cannot be estimated separately, as they are in a principal-component analysis. As might be expected when a $(m+p)$-dimensional model is fitted to a p-dimensional set of vectors, various ambiguities and indeterminancies can arise if the analysis is not carefully done.

A factor analysis typically has a different feel from a principal-component analysis. The latter procedure essentially derives the $\vec{\mathsf{u}}_k$ directly by respanning $\mathcal{V}_{\mathsf{X}}$, and any reduction in dimensionality is accomplished by throwing away the least substantial components. It can be seen as data simplification rather than as the fitting of a theoretical model. In contrast, a factor analysis involves fitting the theoretical structure implied by Equation 9.7 to the observed configuration of vectors. It has a more abstract quality, and its solutions are often interpreted as describing the underlying structure of the process that generated the variables. The difference between the two approaches is apparent in a comparison of two-dimensional representations in Figures 9.1 and 9.4. Without a strong theoretical model one could not transform two data vectors into a three-vector factor-analytic representation.

The constraints in confirmatory factor analysis are even more strongly expressed than they are in exploratory factor analysis. The confirmatory procedures fit data to a restricted factor structure, which is usually described either by a path diagram showing which variables are related to which others or by a set of linear combinations that the structure must obey. For example, one may suppose that certain of the observed variables depend only on a few of the total set of factors or that the factors derived from one set of variables are linear combinations of the factors derived from another set. Interpreted geometrically, these models constraint the placement of the vectors either by confining the projections of the original vectors $\vec{\mathbf{x}}_j$ to subspaces of the factor space or by restricting the placement of the location of factor vectors within $\mathcal{V}_{\mathfrak{f}}$. Often these models involve many variables and factors, making a complete geometric representation difficult or impossible to visualize. Nevertheless, the subject-space pictures are good ways to gain insight into the restrictions imposed and how they constrain the relationships among the variables. Frequently one can understand the effect of a constraint on a limited part of the model more clearly through the geometry than through a set of equations.

Exercises

1. Sketch a pair of vectors $\vec{\mathbf{x}}_1$ and $\vec{\mathbf{x}}_2$ with the following standard deviations and correlation. Add the principal-component vectors corresponding to variables

a. $s_1 = 1.0$, $s_2 = 1.0$, and $r = 0.25$.
b. $s_1 = 0.5$, $s_2 = 1.5$, and $r = 0.00$.
c. $s_1 = 1.0$, $s_2 = 1.0$, and $r = -0.65$.
d. $s_1 = 2.0$, $s_2 = 2.0$, and $r = -0.65$.
e. $s_1 = 1.0$, $s_2 = 1.0$, and $r = 0.85$.
f. $s_1 = 1.0$, $s_2 = 2.0$, and $r = 0.85$.
g. $s_1 = 1.0$, $s_2 = 4.0$, and $r = 0.85$.

2. Consider three standardized variables with correlations $r_{12} = r_{13} = 0.80$ and $r_{23} = 0.50$. Represent the three variables by pointers in three-dimensional space and identify the primary and secondary directions of variability. Indicate the placement of the principal-component vectors—their squared lengths are 2.41, 0.50, and 0.09. Discuss the dimensionality of this configuration.

3. Draw diagrams of regression and principal-component analysis for the special cases discussed at the end of Section 9.1.

4. Suppose that a researcher postulates that a set of nine variables gives a three-dimensional factor solution in which variables X_1 through X_4 are related to factors one and two and the remaining variables are related to factors two and three. Describe the pattern of the common components $\vec{c}_j$ implied by this confirmatory factor structure.

Chapter 10

Canonical correlation

In multiple regression, a linear combination of p predictors variables, X_1, X_2, ..., X_p is used to describe a single outcome variable Y. The procedure known as *canonical correlation* generalizes this idea to multivariate outcome measures. The p variables X_j are the same, but they are now matched to a set of q variables, $\mathsf{Y}_1, \mathsf{Y}_2, \ldots, \mathsf{Y}_q$. One wishes to discover both the magnitude and the nature of the relationship between the two sets of variables. There is a symmetry to the canonical-correlation problem that is absent in multiple regression. In regression, one moves from the X_j to Y; in canonical correlation, the two sets have an identical status.

10.1 Angular relationships between spaces

Geometrically, canonical correlation is a multidimensional extension of the measurement of the association between two vectors, either by the angle between them or by their correlation coefficient. In the canonical correlation problem the single vectors of simple correlation are replaced by the subspaces $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Y}$ generated by the sets of vectors $\vec{\mathbf{x}}_1$, $\vec{\mathbf{x}}_2$, ..., $\vec{\mathbf{x}}_p$ and $\vec{\mathbf{y}}_1$, $\vec{\mathbf{y}}_2$, ..., $\vec{\mathbf{y}}_q$, respectively. Because the dimensions of these spaces are typically greater than unity, the relationship between them is more complex than it is with single variables. Nevertheless, it is measured in essentially the same way, using the angles between vectors that are members of $\mathcal{V}_\mathsf{X}$ and vectors that are members of $\mathcal{V}_\mathsf{Y}$. The fundamental idea is to select pairs of *canonical vectors*, one from $\mathcal{V}_\mathsf{X}$ and one from $\mathcal{V}_\mathsf{Y}$, that have a particularly close association with each other and to characterize the relationship between the spaces by these vectors and the angles between them.

In one sense, canonical correlation is a generalization of multiple regression. In multiple regression, there are two spaces, the p-dimensional space

$\mathcal{V}_X$ generated by the predictors and unidimensional space $\mathcal{V}_Y$ generated by $\vec{\mathbf{y}}$. A linear combination of the predictors $\hat{\vec{\mathbf{y}}} \in \mathcal{V}_X$ is formed that approximates the outcome $\vec{\mathbf{y}} \in \mathcal{V}_Y$ as closely as possible. The discrepancy between prediction and outcome is minimized by selecting $\hat{\vec{\mathbf{y}}}$ so that its direction in the regression space $\mathcal{V}_X$ is angularly closest to $\mathcal{V}_Y$. This notation of angular closeness (or equivalently of maximal correlation) generalizes to the canonical-correlation problem in which the space $\mathcal{V}_Y$ is multidimensional.

It is hard to visualize angles between subspaces at first, since each subspace contains many vectors with many different orientations. However, the angle between any two vectors $\vec{\mathbf{x}} \in \mathcal{V}_X$ and $\vec{\mathbf{y}} \in \mathcal{V}_Y$ is easily imagined and can be calculated with the usual dot-product rule. Over the collection of vectors in the two spaces, some pairs are angularly more similar than others. The trick is to choose the two vectors $\vec{\mathbf{u}} \in \mathcal{V}_X$ and $\vec{\mathbf{v}} \in \mathcal{V}_Y$ that have the smallest angle between them. This angle measures the similarity of the spaces. The variables corresponding to these canonical vectors are known as *canonical variables*, and the correlation $R = \cos\angle(\vec{\mathbf{u}}, \vec{\mathbf{v}})$ is the *canonical correlation coefficient*. As members of the spaces $\mathcal{V}_X$ and $\mathcal{V}_Y$, the canonical vectors are linear combinations of the original vectors:

$$\begin{aligned} \vec{\mathbf{u}} &= a_1\vec{\mathbf{x}}_1 + a_2\vec{\mathbf{x}}_2 + \cdots + a_p\vec{\mathbf{x}}_p, \\ \vec{\mathbf{v}} &= b_1\vec{\mathbf{y}}_1 + b_2\vec{\mathbf{y}}_2 + \cdots + b_q\vec{\mathbf{y}}_q. \end{aligned} \tag{10.1}$$

The coefficients a_j and b_k that make up the combinations are known as *canonical coefficients*.

The minimum-angle criterion does not uniquely define the canonical vectors. It gives their direction but not their length. An additional constraint on the canonical coefficients is needed to make the vectors unique, at least up to reflection through the origin. Many versions of this constraint are possible, of which two have simple geometric interpretations. Viewed from subject space, the simplest approach is to fix the length of the vectors $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$, usually by giving the corresponding variables unit variance. The diagrams in this chapter use this equal-length convention. Viewed from variable space, the natural constraint is to fix the sum of squared coefficients to unity, as was done in principal-component analysis (Equation 9.1):

$$a_1^2 + a_2^2 + \cdots + a_p^2 = b_1^2 + b_2^2 + \cdots + b_q^2 = 1. \tag{10.2}$$

These coefficients create a unit-length vector in variable space and define a new axis in that space.

A picture that shows the relationship between the spaces in canonical correlation is hard to draw. For the problem not to reduce to multiple regression, both $\mathcal{V}_X$ and $\mathcal{V}_Y$ must be at least two dimensional. Their combination is a four-dimensional space, and many interesting examples involve

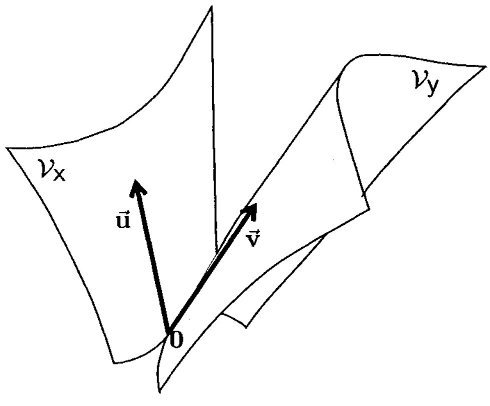


Figure 10.1: *The canonical correlation problem represented by two planes in space. It will probably be necessary to construct a three-dimensional representation of this four-dimensional configuration, using hands, sheets of paper, or the like.*

at least five dimensions. Still, these configurations are not impossible to imagine. Sketching them in two dimensions is more problematic; for what it may be worth, Figure 10.1 attempts an illustration. Without dwelling too much on three-dimensional reality, think of the two-dimensional spaces $\mathcal{V}_X$ and $\mathcal{V}_Y$ as planes that touch only at the origin. To allow the two spaces to touch at one point but nowhere else in three-dimensional space, they must be curved. This curvature is not really present, but substitutes for one's inability to see a four-dimensional space. The vectors canonical $\vec{u}$ and $\vec{v}$ are the two vectors in these planes that are angularly closest to each other. The canonical correlation coefficient measures the angle between them.

Picking the canonical vectors involves simultaneously selecting optimal directions in both $\mathcal{V}_X$ and $\mathcal{V}_Y$. To understand how this selection is made, forget for the moment about the simultaneous aspect and treat the problem of finding $\vec{u}$ and $\vec{v}$ separately, in the manner of multiple regression. First think of finding $\vec{u}$ in the space $\mathcal{V}_X$. Suppose that somehow one has already identified a vector $\vec{v} \in \mathcal{V}_Y$ that is angularly nearest to $\mathcal{V}_X$. Among all unit-length vectors in $\mathcal{V}_X$, some are closer to $\vec{v}$ than others. From these, select the vector $\vec{u}$ that makes the smallest angle with $\vec{v}$. An illustration of this choice (Figure 10.2) looks much like the picture of multiple regression in Figure 4.2, but with a few important differences. The vector $\vec{v}$ has the same role as $\vec{y}$ in regression, and $\vec{u} \in \mathcal{V}_X$ has a role somewhat like $\hat{\vec{y}}$. However,

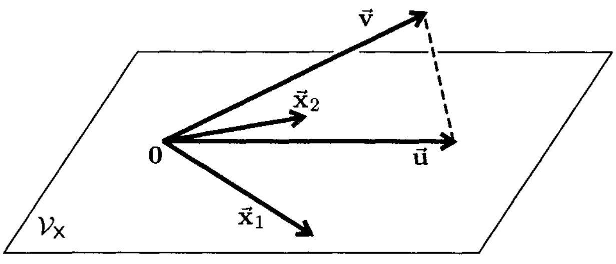


Figure 10.2: *The placement of the canonical vector in the space $\mathcal{V}_X$. A comparable picture expresses the selection of $\vec{v}$ from $\mathcal{V}_Y$. Compare this diagram to Figure 4.2.*

while in regression the length of $\hat{\vec{y}}$ is chosen to be the projection of $\vec{y}$ onto $\mathcal{V}_X$, in canonical correlation only the direction of $\vec{u}$ is determined from $\vec{v}$. Its length is fixed by one of the constraints mentioned above. The direction of $\vec{u}$ is the same as that of a regression predictor, but its tip is not directly under the tip of $\vec{v}$, which is why there is no error vector $\vec{e} \perp \mathcal{V}_X$ in the regression sense. The dashed line connecting $\vec{u}$ and $\vec{v}$ in Figure 10.2 indicates the link between them, but is not a projection and is not orthogonal to either vector.

The description in the last paragraph took the target vector $\vec{v} \in \mathcal{V}_Y$ as given. In fact, it must also be found. The selection of $\vec{v}$ is like the selection of $\vec{u}$ and is illustrated by a picture almost identical to Figure 10.2. The roles of $\vec{u}$ and $\vec{v}$ are exchanged, and the search takes place in $\mathcal{V}_Y$ instead of $\mathcal{V}_X$. Each of these two optimizations is half the problem. Although it is helpful to think of the two searches separately, they are interdependent and must be done simultaneously. The mathematical algorithm that is used for canonical correlation works this way, finding both the vector $\vec{u} \in \mathcal{V}_X$ closest to $\mathcal{V}_Y$ and the vector $\vec{v} \in \mathcal{V}_Y$ closest to $\mathcal{V}_X$ simultaneously.

There are two special cases where the relationship between the subspaces $\mathcal{V}_X$ and $\mathcal{V}_Y$ makes the canonical-correlation problem trivial. The first case occurs when the two spaces are mutually orthogonal, and the correlation of every $\vec{x} \in \mathcal{V}_X$ with every $\vec{y} \in \mathcal{V}_Y$ is zero. This configuration is not particularly interesting: some relationship between the spaces is necessary for canonical correlation to provide useful information. The second special case occurs when the spaces overlap and have vectors in common (other than the null vector $\vec{0}$). When this happens a vector from the common subspace serves for both $\vec{u}$ and $\vec{v}$, the minimum angle between members of the spaces is zero, and $R = 0$. The situation is comparable to a regression problem in which $\vec{y}$ falls in the space $\mathcal{V}_X$ and can be predicted perfectly. Overlapping spaces of this type are most likely to be caused by mistakes made when selecting variables for the analysis. Possibly the same variable has been inadvertently included in both sets or some unheeded linear dependency is present. For example, placing several scores in one set and their sum in the other creates such an overlap. The canonical correlation analysis here is less trivial than in its regression counterpart, however. Although $\mathcal{V}_X$ and $\mathcal{V}_Y$ share some vectors and the angle between the canonical vectors is zero, the spaces need not be identical in other respects. As will be described in the next section, higher-order canonical components may be present that describe the parts of the spaces that do not intersect. Nevertheless, canonical correlation is most useful and interesting when it connects two sets of variables that derive from distinctly different measurements.

10.2 The sequence of canonical triplets

The process just described produces a canonical triplet consisting of the vector $\vec{u}$, the vector $\vec{v}$, and the maximum correlation R or minimum angle between them. This single triplet does not capture the full multivariate character of the relationship between the spaces $\mathcal{V}_X$ and $\mathcal{V}_Y$. When both spaces are multidimensional, their relationship cannot be expressed by a single linear combination. Except in the simplest problems, other canonical triplets can be found.

Conceptually, these additional canonical variables are found by applying the canonical-correlation procedure recursively, as was done in principal-component analysis. At each step, a new search is made in the portions of the spaces that are orthogonal to the vectors already found. Suppose that a canonical correlation analysis has been performed on two spaces with dimensions p and q, both greater than unity, and that the triplet $(\vec{u}_1, \vec{v}_1, R_1)$ has been obtained. The vector $\vec{u}_1$ specifies one direction in the space $\mathcal{V}_X$, but it does not span the whole space. The orthogonal complement $\mathcal{V}_{X\perp u_1}$ of $\vec{u}_1$ in $\mathcal{V}_X$ has dimension $p-1$. It describes the variation of the $\vec{x}_j$ that is unrelated to $\vec{u}_1$. By the same logic, there is a subspace $\mathcal{V}_{Y\perp v_1}$ of dimension $q-1$ in $\mathcal{V}_Y$ that is orthogonal to the canonical vector $\vec{v}_1$. One now measures the relationship between $\mathcal{V}_{X\perp u_1}$ and $\mathcal{V}_{Y\perp v_1}$ by repeating the canonical-correlation analysis. Among all vectors in $\mathcal{V}_{X\perp u_1}$ and $\mathcal{V}_{Y\perp v_1}$, the two vectors $\vec{u}_2 \in \mathcal{V}_{X\perp u_1}$ and $\vec{v}_2 \in \mathcal{V}_{Y\perp v_1}$ are chosen that have the smallest angle between them. The angle between $\vec{u}_2$ and $\vec{v}_2$ is no less than that between the original canonical variates $\vec{u}_1$ and $\vec{v}_1$ (otherwise they would have been extracted as the first pair), but unless the two residual subspaces are orthogonal this angle is less than 90°. The variables $\vec{u}_2$ and $\vec{v}_2$ capture a portion of the relationship between $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$ that is not expressed by $\vec{\mathbf{u}}_1$ and $\vec{\mathbf{v}}_1$.

The process of reducing the spaces to orthogonal complements and reapplying the canonical-correlation selection procedure can be repeated until at least one of the spaces is exhausted. From these steps, one gets a series of triplets, each composed of a canonical variate from each space and the canonical angle (or correlation) between them:

$$(\vec{\mathbf{u}}_1, \vec{\mathbf{v}}_1, R_1), \quad (\vec{\mathbf{u}}_2, \vec{\mathbf{v}}_2, R_2), \quad \ldots, \quad (\vec{\mathbf{u}}_s, \vec{\mathbf{v}}_s, R_s).$$

Because each step removes a dimension from the spaces, there are as many canonical triplets as the dimension of the smaller space, and

$$s = \min[\dim(\mathcal{V}_{\mathsf{X}}),\ \dim(\mathcal{V}_{\mathsf{Y}})].$$

If there are no within-set linear dependencies, then s equals the number of variables in the smaller set.

The way that the triplets are determined gives them two important properties. First, the angles between the vectors are ordered, with the angle between each pair of components no smaller than the one that precedes it. The sequence of canonical correlations never increases and, except in a few special cases, declines at each step from R_1 to R_s. Second, the vectors in each set are mutually orthogonal. For any pair of different indices j and k, $\vec{\mathbf{u}}_j \perp \vec{\mathbf{u}}_k$ and $\vec{\mathbf{v}}_j \perp \vec{\mathbf{v}}_k$. The canonical variables within a set are uncorrelated. The vectors from a complete set of canonical triplets span the smaller of the original spaces and create an orthogonal basis for that space. They create a matching subspace in the larger space and give it an orthogonal basis.

Both canonical correlation and principal-component analysis respan a vector space with an ordered orthogonal basis. They differ in the criteria by which these vectors are chosen. The structure extracted by a principal-component analysis derives from the variability within the space, while that extracted by canonical correlation derives from the relationship to another space. Different structures can be implied by the different criteria. The vector that captures the greatest part of the relationship between two spaces can be an unimportant part of the within-space variability, and the largest principal component in one space can have little relationship to the variability in another space. Nevertheless, if the data arise from a low-dimensional structure combined with random noise, the dominant within-space and between-space structures are likely to be similar.

Interpreted in variable space, one way to view a canonical-correlation analysis is as an axis rotation. If the coefficients are standardized to give a unit sum of squares (as in Equation 10.2), then each set corresponds to a unit vector $\vec{\mathbf{a}}_j$, and the transformation shifts the space to one measured with these axes (as described in connection with Equation 2.14). The transformation differs from the axis rotation that appeared in principal-component analysis, however. In canonical correlation, the coefficient vectors are not in general orthogonal. The new variable-space axis corresponding to $\vec{\mathbf{u}}_1$ is not necessarily orthogonal to the axis corresponding to $\vec{\mathbf{u}}_2$, and so forth. Thus, the axis rotation here is oblique, not orthogonal.

Although it is helpful to think of a canonical-correlation analysis as the sequential matching of one dimension at a time, one does not actually calculate it that way. Like principal-component analysis, mathematically canonical correlation is a constrained maximization problem. The canonical vectors and their correlations are found algebraically as the solution to an eigenvalue-eigenvector problem. The coefficients a_{jk} and b_{jk} of the linear combinations that determine $\vec{\mathbf{u}}_k$ and $\vec{\mathbf{v}}_k$ are eigenvectors, and the squared correlations R_k^2 are the eigenvalues. As in principal-component analysis, this terminology pervades many treatments of canonical correlation.

Frequently one does not need the full set of $s = \min[\dim(\mathcal{V}_\mathsf{X}), \dim(\mathcal{V}_\mathsf{Y})]$ canonical components to express the meaningful part of the relation between $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Y}$. The logic here is similar to that used with principal-component analysis or factor analysis. The size of the canonical correlation falls off with each successive triplet. Because each sequence of vectors is orthogonal, the squares of these coefficients can be interpreted as additive sums of squares or variances. Typically, the bulk of the variance relating the spaces resides in the subspace spanned by the first few canonical variables. In such cases, one is usually justified in considering only the important canonical variables and ignoring the rest. In effect, one concludes that the connection between $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Y}$ resides in a pair of subspaces that are only small parts of the original spaces.

Once one has selected a set of canonical variables to express the association between spaces, one usually wants to give them meaning. The discussion of the interpretation of regression coefficients in Section 4.3 applies here. Useful information comes from the *canonical loadings*, which are the angles between the original vectors and the derived canonical vectors, $\angle(\vec{\mathbf{x}}_j, \vec{\mathbf{u}}_k)$ in $\mathcal{V}_\mathsf{X}$ and $\angle(\vec{\mathbf{y}}_j, \vec{\mathbf{v}}_k)$ in $\mathcal{V}_\mathsf{Y}$. Small angles between vectors indicate that the variables measure much the same thing. Alternatively, the complete makeup of a linear combination may suggest how that component is formed. As with any other linear combination of nonorthogonal variables, the individual coefficients should be interpreted only with great caution.

The cautions that apply to the interpretation of multiple regression should also be observed with canonical correlation. The analysis selects vectors from one subspace based on their relationship to structures outside the space. Even when the vectors themselves are fairly well determined, multicollinearity among the original vectors may make the canonical vectors hard to interpret. With two spaces in which selection is taking place, the opportunity to capitalize on accidental relationships among the variables is much greater in canonical correlation than it is in the other multivariate procedures. The technique should not be applied unless the original variables are very well determined. Large sample sizes are essential to obtain the necessary stability.

10.3 Test statistics

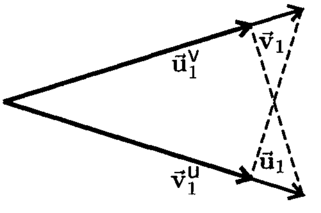
A full canonical-correlation analysis generates a sequence of s canonical triplets $(\vec{\mathbf{u}}_k, \vec{\mathbf{v}}_k, R_k)$ that characterize the relationship between the two sets of variables. One can treat these components purely descriptively, as a way to measure and represent the association. However, one often wants to determine whether the association that they describe is due to more than chance. Suppose that in an underlying population, the variables measured by the X_j and the Y_j are truly unrelated—the population vectors are orthogonal. Accidents of sampling can only decrease these right angles, so that some evidence of angular agreement appears in the sample. By maximizing the association between $\vec{\mathbf{u}}_k$ and $\vec{\mathbf{v}}_k$, canonical correlation captures the accidental variation and the observed R_k is positive. One wants a test to determine whether the magnitudes of the observed canonical correlations are larger than could be attributed to these accidental effects.

Several test statistics are available to probe for systematic effects in a canonical-correlation analysis. Since the relationship between the spaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$ is summarized by the angles between their canonical vectors, these statistics are based either on these angles directly or on functions that combine them. There are two basic test strategies: one can either look at the first canonical correlation alone or can combine all the correlations into an omnibus test statistic.

The simpler approach is to look only at the first canonical pair. Its vectors have the minimum possible angular separation and their correlation is maximal. If the relationship between the spaces is purely random, this angle is large, while if there is a systematic relationship, the angle is small. A threshold can be constructed and the hypothesis of random relationship rejected when R_1 exceeds this threshold. The conventional test here uses the *greatest characteristic root statistic*, which equals the square of the canonical correlation coefficient. The particular critical value of this statistic depends on the dimensions of the spaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$, the number of observations, and the desired significance level. Tables of these values appear in many multivariate texts. Taking advantage of the symmetries of the situation, they are usually organized by the number of canonical components, $s = \min(p, q)$, the differences in the dimensions of the two effect spaces, $|p - q|$, and the dimension of the error space, $N - p - q$.

A test of canonical correlation does not have the same geometry under the null hypothesis as the test of multiple regression that was illustrated in Figure 6.3. The difference lies in the random components. Instead of a fixed space $\mathcal{V}_{\mathsf{X}}$ and a random vector $\vec{\mathbf{y}}$ under the null hypothesis, both spaces are random. When there is no systematic relationship between the two sets of variables, both $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$ are randomly oriented subspaces in the $(N-1)$-dimensional space of deviations about the grand mean. To help visualize the null-hypothesis configuration, think of a line for $\mathcal{V}_{\mathsf{X}}$ and a line for $\mathcal{V}_{\mathsf{Y}}$ struck at random angles through the center of a sphere—this picture is rough and loses the multivariate nature of the spaces. Although the random axes are unlikely to be orthogonal, they are still less likely to be nearly parallel. The distribution of the statistic under the null hypothesis derives from this representation, although the multidimensionality of $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$ makes its form less obvious than it is in regression. One aspect of the regression picture carries over clearly. When the null hypothesis is false and a systematic component is present, the centers of variability about the population centers are displaced away from the origin, somewhat as shown in the right-hand panel of Figure 6.3. The displacement reduces the angle between the observed vectors $\vec{\mathbf{x}}_j$ and $\vec{\mathbf{y}}_k$ and increases their correlation.

The greatest root statistic can be interpreted as a vector length. When $R > 0$, the two canonical vectors $\vec{\mathbf{u}}_1$ and $\vec{\mathbf{v}}_1$ are not orthogonal and have projections $\vec{\mathbf{u}}_1^{\mathsf{V}}$ and $\vec{\mathbf{v}}_1^{\mathsf{U}}$ onto each other:



The two original vectors have the same length, so the two projections also have identical lengths. Where the spaces are closely related, these projections are long, and where the spaces are unrelated, they are short. Using the geometry of right triangles,

$$|\vec{\mathbf{u}}_1^{\mathsf{V}}| = |\vec{\mathbf{v}}_1^{\mathsf{U}}| = |\vec{\mathbf{u}}_1||\vec{\mathbf{v}}_1| \cos \angle(\vec{\mathbf{u}}_1, \vec{\mathbf{v}}_1).$$

If the canonical vectors are given unit length, then the correlation equals the vector length, and

$$|\vec{\mathbf{u}}_1^{\mathsf{V}}|^2 = |\vec{\mathbf{v}}_1^{\mathsf{U}}|^2 = R_1^2.$$

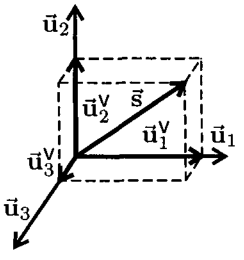
The greatest root statistic equals the length of these projections.

For many purposes, examining the first root is a completely satisfactory way to test for association. However, the fully general multivariate relationship between the spaces $\mathcal{V}_{\mathsf{X}}$ and $\mathcal{V}_{\mathsf{Y}}$ cannot be captured in any single dimension. Canonical triplets beyond the first carry components of the relationship that are not measured by R_1. To examine the relationships in these directions, one wants a way to test the correlations R_2, R_3, and so forth for deviation from zero. Unfortunately, exact test statistics for these components are not available. The problem arises from a fundamental limitation on one's knowledge of the situation. The sampling distribution of the kth canonical correlation depends on the true (population) values of canonical triplets 1 through $k-1$, both their correlation and their canonical vectors. For example, to test the null hypothesis that the second canonical correlation is zero, one needs to know—not just estimate from the data—the first canonical triplet. Because this population information is never available in any realistic circumstance, the sampling distributions needed for the tests cannot be constructed. Only the first canonical correlation R_1 does not have a predecessor, and its distribution does not depend on assumptions about ancillary population effects.

Instead of testing the higher-order terms individually, one turns to tests that combine information from the entire set of s components in a single test statistic. There are several such statistics. An easy way to create a composite statistic is to sum the unidimensional relationships as vectors. The projection vectors $\vec{\mathbf{u}}_1^{\mathsf{V}}$ and $\vec{\mathbf{v}}_1^{\mathsf{U}}$ in the first dimension have their counterparts $\vec{\mathbf{u}}_k^{\mathsf{V}}$ and $\vec{\mathbf{v}}_k^{\mathsf{U}}$ in the other dimensions. Each of these projections measures the association in one canonical direction, and they are mutually orthogonal. Together they form the vector

$$\begin{aligned} \vec{\mathbf{s}} &= \vec{\mathbf{u}}_1^{\mathsf{V}} + \vec{\mathbf{u}}_2^{\mathsf{V}} + \cdots + \vec{\mathbf{u}}_s^{\mathsf{V}} \\ &= \vec{\mathbf{v}}_1^{\mathsf{U}} + \vec{\mathbf{v}}_2^{\mathsf{U}} + \cdots + \vec{\mathbf{v}}_s^{\mathsf{U}}. \end{aligned}$$

The vector $\vec{\mathbf{s}}$ is the diagonal of the rectangular solid formed by the projection vectors and is, in this sense, a measure of the amount of overlap between the spaces:

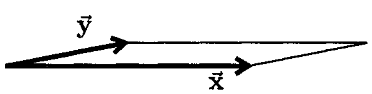


Since $\vec{\mathrm{s}}$ is the sum of orthogonal components, its squared length is

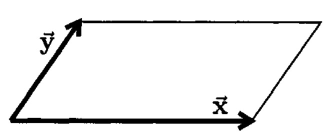
$$\begin{aligned} |\vec{\mathrm{s}}|^2 &= |\vec{\mathrm{u}}_1^{\mathrm{V}}|^2 + |\vec{\mathrm{u}}_2^{\mathrm{V}}|^2 + \cdots + |\vec{\mathrm{u}}_s^{\mathrm{V}}|^2 \\ &= R_1^2 + R_2^2 + \cdots + R_s^2. \end{aligned} \tag{10.3}$$

This vector is shorter when the spaces are unrelated than when they have much in common. In the multivariate literature this quantity is known as *Pillai's trace statistic.* As with the first root, the distribution of this statistic under the null hypothesis of unrelated spaces depends on the sizes of both effect spaces and of the error space. Its critical values are specially tabled or approximated by a function of an F statistic.

A second omnibus test statistic measures the common portion of the spaces in a different way. The basis for this statistic is the multivariate volume measure described in Section 2.1 and used to measure near multicollinearity in Section 5.2. Its use as a multivariate test statistic is easiest to see in a two-dimensional analogy. The similarity of two vectors $\vec{\mathrm{x}}$ and $\vec{\mathrm{y}}$ is inversely measured by the area of the parallelogram that they determine. Closely associated vectors make a small area and different vectors make a large area:



and



The area depends in part on the length of the vectors, and to standardize it between 0 and 1 (collinearity and orthogonality, respectively), it is divided by the product of the lengths of the two vectors:

$$\text{measure} = \frac{\text{area of parallelogram}}{|\vec{\mathrm{x}}||\vec{\mathrm{y}}|}.$$

This measure equals $\sqrt{1 - r_{\mathrm{XY}}^2}$, which is small when the vectors have much in common and near unity when they are unrelated. To turn this measure into a multivariate statistic, one replaces the areas and lengths by the generalized volume measure, which gives the multivariate volume of the parallelepiped formed by the set of vectors. The volume defined by the $p+q$ vectors in the entire problems is divided by the volumes measured separately within the spaces $\mathcal{V}_{\mathrm{X}}$ and $\mathcal{V}_{\mathrm{Y}}$. The conventional version of this statistic, known as *Wilks' lambda statistic*, is the square of this ratio:

$$\Lambda = \frac{[\mathrm{vol}(\vec{\mathrm{x}}_1, \ldots, \vec{\mathrm{x}}_p, \vec{\mathrm{y}}_1, \ldots, \vec{\mathrm{y}}_q)]^2}{[\mathrm{vol}(\vec{\mathrm{x}}_1, \ldots, \vec{\mathrm{x}}_p)]^2 [\mathrm{vol}(\vec{\mathrm{y}}_1, \ldots, \vec{\mathrm{y}}_q)]^2}. \tag{10.4}$$

Numerically, the squared volumes equal the determinants of the covariance matrices of the appropriate sets of variables. The Λ statistic is small when there is an association between the spaces, and it is nearly unity when they are unrelated. For completely orthogonal spaces, $\Lambda = 1$. As with the other multivariate statistics, its sampling distribution is specially tabled or approximated by an F ratio. A version of this statistic, applied to multiple regression, was mentioned in Problem 4.5.

The three statistics described in this section measure violations of the unrelatedness hypothesis in different ways. Each statistic is most responsive to a different alternative hypothesis. The statistic that depends on the first component is most clearly different from the other two. It is largest when the primary component of the relationship between the spaces is substantial. When a single common dimension dominates the relationship, it is the most powerful of the three statistics. The other two statistics combine information over all components and are less sensitive to unidimensional association. In compensation, they are more likely to detect patterns in which each of the first few population canonical components carries a moderate amount of the association but none is greatly more important that the others. They guard against the possibility that $\mathcal{V}_\mathsf{X}$ and $\mathcal{V}_\mathsf{Y}$ have a diffuse association spread over several dimensions.

10.4 The multivariate analysis of variance

The analysis of variance described in Chapter 8 investigates how the mean of a single variable Y differs among a set of groups. However, one often obtains data in which a set of potentially nonorthogonal outcome variables Y_1, Y_2, ..., Y_q is recorded from each subject. The multivariate analysis of variance generalizes the univariate procedure to investigate how the multivariate means of these variables differ among the groups. Just as multiple regression can be used to analyze univariate differences in means, canonical correlation can be used to examine the multivariate pattern of group means. The multivariate analysis of variance combines the dummy coding of groups from the analysis of variance with the measures of multivariate relationship from canonical correlation.

Broadly speaking, the univariate and multivariate analyses differ in two respects. From a descriptive viewpoint, the analysis takes the multivariate nature of the outcome variables into account. For example, a collection of highly correlated variables probably carries redundant information about the group differences. Although each variable does not provide separate information about the groups, a combination of the variables may give a clearer picture of how the groups differ than does any variable alone. From an inferential viewpoint, the sources of random variation are more complicated than in the ordinary analysis of variance. The sampling variability affects all the interconnected outcome variables and the correlations among these variables must be accommodated in the analysis.

The multivariate analysis of variance extends the univariate analysis of variance in the same way that canonical correlation extends multiple regression. The groups are represented by dummy variables, as they are in the analysis of variance. The structure of a g-group design is expressed by a set of $g-1$ dummy vectors $\vec{x}_1$, $\vec{x}_2$, ..., $\vec{x}_{g-1}$ to create the space $\mathcal{V}_X$. These vectors are selected in whatever way is convenient and interpretable. The outcome vectors $\vec{y}_1$, $\vec{y}_2$, ..., $\vec{y}_q$ create the space $\mathcal{V}_Y$. Because both of these spaces are potentially multidimensional, canonical correlation is used to assess their relationship.

A multivariate analysis of variance provides two distinct types of information about the group differences. First, it measures the magnitude of the relationship of the groups to the variables. This information is embodied in the size of the canonical correlations R_k or in test statistics such as $|\vec{s}|^2$ or Λ (Equations 10.3 and 10.4). Second, it gives information about the particular way that the groups differ. The canonical vectors $\vec{u}_k$ and $\vec{v}_k$ and their coefficients a_{jk} and b_{jk} identify patterns that discriminate among the groups.

The simplest form of multivariate group difference occurs with two groups. Here $\mathcal{V}_X$ is a one-dimensional space spanned by a single dummy vector, as in the two-group tests of Section 3.4. The space $\mathcal{V}_Y$ is q-dimensional. Figure 10.3 illustrates the relationship when $q = 2$. The space $\mathcal{V}_X$ is a line, and the space $\mathcal{V}_Y$ is a plane that crosses it. The picture is similar to that of regression, but with the dimensions of $\mathcal{V}_X$ and $\mathcal{V}_Y$ reversed. Because the effect space $\mathcal{V}_X$ is unidimensional, there is a single canonical component, and the vector $\vec{u} \in \mathcal{V}_X$ is determined up to a reflection—it could point either left or right. The vector $\vec{v}$ lies in $\mathcal{V}_Y$ in the position closest to $\vec{u}$. It defines the combination of variables that maximally discriminates between the groups, and so is known as a *discriminant function.* The relationship of $\vec{v}$ to the original vectors helps one interpret how the groups differ. For example, when $\vec{v}$ makes a positive angle to all the $\vec{y}_j$, it suggests that all the variables contribute jointly to the discrimination. The remarks in Section 4.3 on the interpretation of multiple regression vectors apply here.

Geometrically, the two-group configuration in Figure 10.3 differs from that in multiple regression only in the reversed roles of $\mathcal{V}_X$ and $\mathcal{V}_Y$. Formally, the regression projection equations discussed in Section 4.1 can be used to find $\vec{v}$. A set of normal equations analogous to Equations 4.1 and 4.4 is written, with $\vec{x}$ and $\vec{y}$ exchanged. Solving this system gives the coefficients of $\vec{v}$. Under the assumption of normally distributed variability, the test-

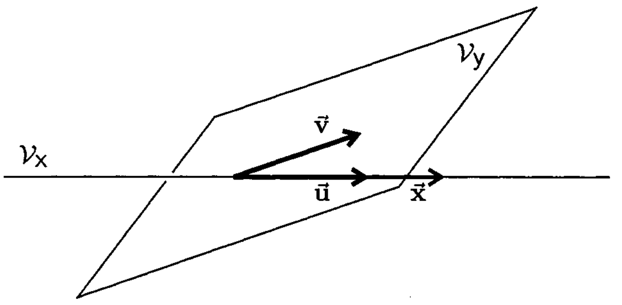


Figure 10.3: *The relationship between the dummy variable $\vec{x} \in \mathcal{V}_X$ and the outcome space $\mathcal{V}_Y$ in a two-group multivariate analysis of variance. The two outcome vectors that define $\mathcal{V}_Y$ are not shown.*

ing arguments of Chapter 6 also apply. Although the random orientation applies to the multidimensional space $\mathcal{V}_Y$ instead of the univariate space $\mathcal{V}_X$, the angular relationship of the spaces is the same—it does not matter whether the line or the space wobbles when one is looking only at the angle between them. The symmetry of normal sampling variability makes the two situations equivalent.

An analysis with more than two groups needs the full machinery of canonical correlation. Both $\mathcal{V}_X$ and $\mathcal{V}_Y$ are multidimensional spaces, and a series of canonical triplets is produced, each triplet composed of a vector that contrasts the groups, a discriminant function on the outcome variables, and a canonical correlation. Mechanically, the analysis is like other canonical-correlation problems. Programs specifically designed for the multivariate analysis of variance work in the same way as programs for canonical correlation, although they usually include special features to treat the group structure, such as automatic coding of the groups and interactions or ways to plot the discriminant functions.

Often the hardest part of a multivariate analysis of variance is figuring out what it means. The multivariate outcome makes the relationship among the groups more complicated than it is in the univariate analysis, and there is no uniformly best way to proceed. A plausible strategy is to do the interpretation in three steps. The first step is to decide how many canonical triplets to use. If the sequence of canonical correlations is dominated by one or two terms, then one can limit one's attention to these—certainly there is no obligation to look at every canonical triplet that could be extracted. The second step is to assign meaning to the canonical variables.

Because the group vectors $\vec{\mathbf{u}}_k$ are linear combinations of dummy vectors, they are themselves dummy vectors. After computing the coefficients of each group in each derived vector, one may recognize these as similar to those of a known contrast variable, such as a group comparison, an average, or polynomial trend. If so, then the differences can be described in those terms; even if not, a single dummy vector is more comprehensible than is a combination of variables. The discriminant function vectors in $\mathcal{V}_{\mathsf{Y}}$ express linear combinations of the outcome variables. The process of interpretation here is like that for any other derived variable. Examining both the coefficients of the combinations and the angles they make to the original vectors is important. The final step is to describe the relationship between the two sets of variables. This task is made easier by having already interpreted the canonical variables separately. Each retained canonical triplet says that some pattern of the groups (the vector $\vec{\mathbf{u}}_k$) is most closely tied to some composite measured variable (the vector $\vec{\mathbf{v}}_k$). This relationship can be described and interpreted.

Where several canonical triplets have been retained, it is useful to examine the results in variable space. Each canonical variable $\vec{\mathbf{v}}_k$ defines a new axis in this space, either by a simple axis rotation if the coefficients are given a unit sum of squares or by a rotation coupled with stretching or shrinking if the coefficients are chosen to fix the new variables' variance. Together, the discriminant functions give a new set of axes, although not usually a rigid rotation of the space, and create a new multidimensional space in which the canonical variables can be represented. For example, with two triplets, one plots values of V_1 and V_2. Plotting the scores themselves is usually less valuable than plotting the centers of the groups, each represented by the transformation of its mean. This plot, particularly if only two or three dimensions are retained, is a useful complement to the vector analysis.

Like the analysis of covariance, the appropriateness of the multivariate analysis of variance depends on the homogeneity of the variability—both variances and covariances—across the groups. Outcome vectors $\vec{\mathbf{y}}_j^{(k)}$ calculated from the kth group alone must be comparable in length and orientation from group to group and they must be similar to the vectors $\vec{\mathbf{y}}_{j\perp\mathsf{X}}$ obtained by projecting the $\vec{\mathbf{y}}_j$ onto the orthogonal complement of $\mathcal{V}_{\mathsf{X}}$. In variable space, the scatterplots from the separate groups must be similar and must be similar to the scatterplot that results when all the groups are shifted to center them at a common mean. Without this consistency, the discriminant-function vectors, although solving the formal canonical-correlation problem, do not represent a construct that has the same meaning in the different groups. For example, if the $\vec{\mathbf{y}}_j^{(k)}$ differ in orientation among the groups, then their angular loadings with the $\vec{\mathbf{v}}_k$ are inconsistent. It is important to examine this homogeneity assumption before finally describing the group differences that are extracted by a multivariate analysis.

In a multifactor design, the dummy vectors defining the various effects create a set of subspaces. In a two-way design, three subspaces $\mathcal{V}_\mathsf{a}$, $\mathcal{V}_\mathsf{b}$, and $\mathcal{V}_\mathsf{ab}$ are produced, corresponding to the A and B main effects and the AB interaction. As discussed in Chapter 8, these spaces may be defined to give either the groups or the individuals equal importance. These subspaces give structure to the subsequent analysis. The most direct approach is to treat each effect subspace separately by examining its relationship to the dependent-variable space $\mathcal{V}_\mathsf{y}$. Each effect in the analysis gives a set of triplets, each composed of a combination of the groups in that effect space, a discriminant function in $\mathcal{V}_\mathsf{y}$, and the canonical correlation between them. A two-factor design is treated as three separate analyses, one relating $\mathcal{V}_\mathsf{a}$ to $\mathcal{V}_\mathsf{y}$, one relating $\mathcal{V}_\mathsf{b}$ to $\mathcal{V}_\mathsf{y}$, and one relating $\mathcal{V}_\mathsf{ab}$ to $\mathcal{V}_\mathsf{y}$. Test statistics such as $|\vec{\mathsf{s}}|^2$ or Λ give composite tests of the different effects. The resulting vectors and effects are interpreted like those of a single-factor analysis. Many computer packages use this approach.

This approach sometimes runs into difficulty. When the various effects are analyzed independently, one is likely to get different sets of discriminant vectors in $\mathcal{V}_\mathsf{y}$ from the different analyses. The combinations that best discriminate the A factor may be different from those that best discriminate factor B or the AB interaction. Sometimes the potential to express the different effects with different combinations of the outcome variables is helpful, but more often it is a curse. On the positive side, with different discriminant vectors, each factor can have a different relationship to the outcome, as may be appropriate for treatments with different actions. On the negative side, using different discriminant vectors for the different effects gives a less unified picture. Interpreting the interaction is particularly confusing. The natural way to think of an interaction is as the modulation of one main effect over the levels of the other, for example, as a simple A effect that changes with the level of B. However, when the main effects and the interaction are defined on different combinations of the outcome variables, as may happen when separate canonical correlations are run, this interpretation is impossible. The discriminant-function vector in $\mathcal{V}_\mathsf{y}$ that is modulated by factor B as as part of the interaction analysis may not be the same vector that is influenced by the main effect of A. This disparity makes the interaction hard to interpret and the interpretation hard to explain. The situation is particularly confusing when the data are noisy, for then it is difficult to tell which differences are real and which are accidental.

One way to escape from the inconsistencies caused by multiple canonical vectors in the outcome space is to use the same discriminant vectors $\vec{\mathsf{v}}_k$ to analyze all effects. This analysis must be done in several steps. First, a one-way multivariate analysis of variance is run to determine the directions in $\mathcal{V}_Y$ that are related to the grouping. In this analysis the structure of the groups is ignored; for example, a two-by-three design is treated as a one-way six-group analysis. This analysis gives a set of discriminant functions that apply to the full group structure. Next, the canonical triplets from this analysis are examined and the important ones retained. If possible, the vectors $\vec{v}_k \in \mathcal{V}_Y$ in the outcome space are given reasonable interpretations. In the final step, the extracted dimensions are related to the factorial structure. A univariate factorial analysis of variance is run for each derived variable. Since these analyses use the same outcome vector for every main effect and interaction, the results are more readily understandable. Because the different $\vec{v}_k$ are orthogonal, each of these analyses describes a distinct aspect of the group differences.

Of course, this approach, like any other, is only one way to understand one's data. Multivariate data are intrinsically rich, and, as with all multivariate problems, there is no sure road to a comprehensible interpretation. Particularly with a design as complicated as a multifactor multivariate analysis of variance, one usually needs to examine several versions of the analysis before settling on successful description of the data.

Exercises

1. Consider a canonical-correlation analysis relating a set of $p = 4$ variables to a set of $q = 5$ variables. Express the correlations R_j given below as vector lengths, and comment on how the spaces $\mathcal{V}_X$ and $\mathcal{V}_Y$ are related.

a. 0.82, 0.34, 0.08, and 0.05.

b. 0.65, 0.53, 0.48, and 0.40.

c. 1.00, 1.00, 0.38, and 0.05.

2. Suppose that $\dim(\mathcal{V}_X) = 4$ and $\dim(\mathcal{V}_Y) = 5$ and that they have two dimensions in common. What patterns of canonical correlations are possible?

3. Consider two miniature groups of bivariate scores:

Group 1	Y_1	1	2	2	2	3	3	3	4	4	5
	Y_2	3	3	4	5	4	6	7	5	7	7
Group 2	Y_1	2	3	3	4	4	4	4	5	5	6
	Y_2	1	2	3	2	3	5	6	3	6	5

A multivariate analysis of variance on these data gives $R = 0.78$ and the canonical vector $\vec{\mathbf{v}} = -0.82\vec{\mathbf{y}}_1 + 0.57\vec{\mathbf{y}}_2$.

a. Plot the scores as a scatterplot in variable space.

b. Draw the vector $\vec{\mathbf{a}}$ in variable space defined by the coefficients of the discriminant function, and draw a new axis along this vector. Projections onto this axis define the variable V.

c. The correlation of Y_1 and Y_2 is 0.38, and their correlations with the dummy vector $\vec{\mathbf{x}} = \vec{\mathbf{U}}_1 - \vec{\mathbf{U}}_2$ are -0.44 and 0.42, respectively. Use these values to describe the subject-space representation of the analysis. Where are $\vec{\mathbf{u}}$ and $\vec{\mathbf{v}}$?

4. Consider two groups of bivariate scores, Y_1 and Y_2. These data can be used either in a multivariate analysis of variance or in an analysis of covariance with Y_1 as the outcome variable and Y_2 as the covariate. Explain, using subject-space diagrams, how these two analyses differ.

Index