CHAPTER 1

INTRODUCTION

In this chapter we introduce and contrast the matricial and geometric formulations of the so-called general linear model and introduce some notational conventions.

1. Orientation

Recall the classical framework of the *general linear model (GLM)*. One is given an n-dimensional random vector $\boldsymbol{Y}^{n\times 1} = (Y_1 \ldots, Y_n)^T$, perhaps multivariate normally distributed, with covariance matrix $(\mathrm{Cov}(Y_i, Y_j))^{n\times n} = \sigma^2 \boldsymbol{I}^{n\times n}$ and mean vector $\boldsymbol{\mu}^{n\times 1} = E(\boldsymbol{Y}) = (EY_1, \ldots, EY_n)^T$ of the form

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{X}^{n\times p}$ is known and σ^2 and $\boldsymbol{\beta}^{p\times 1} = (\beta_1, \ldots, \beta_p)^T$ are unknown; in addition, the β_i's may be subject to linear constraints $\boldsymbol{R}\boldsymbol{\beta} = 0$, where $\boldsymbol{R}^{c\times p}$ is known. $\boldsymbol{X}$ is called the *design,* or *regression, matrix,* and $\boldsymbol{\beta}$ is called the *parameter vector.*

1.1 Example. In the classical *two-sample problem,* one has

$$\boldsymbol{X}^T = \begin{pmatrix} \underbrace{1 \;\; 1 \;\; \ldots \;\; 1}_{n_1 \text{ times}} & \underbrace{0 \;\; 0 \;\; \ldots \;\; 0}_{n_2 \text{ times}} \\ \underbrace{0 \;\; 0 \;\; \ldots \;\; 0}_{n_1 \text{ times}} & \underbrace{1 \;\; 1 \;\; \ldots \;\; 1}_{n_2 \text{ times}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},$$

that is,

$$E(Y_i) = \begin{cases} \mu_1, & \text{if } 1 \le i \le n_1, \\ \mu_2, & \text{if } n_1 < i \le n_1 + n_2 = n. \end{cases}$$

•

1.2 Example. In *simple linear regression,* one has

$$\boldsymbol{X}^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_n \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} a \\ b \end{pmatrix},$$

that is,

$$E(Y_i) = a + bx_i \quad \text{for } i = 1, \ldots, n.$$

•

Typical problems are the estimation of linear combinations of the β_i's, testing that some such linear combinations are 0 (or some other prescribed value), and the estimation of σ^2.

1.3 Example. In the two-sample problem, one is often interested in estimating the difference $\mu_2 - \mu_1$ or in testing the null hypothesis that $\mu_1 = \mu_2$. •

1.4 Example. In simple linear regression, one seeks estimates of the intercept a and slope b and may want to test, for example, the hypothesis that $b = 0$ or the hypothesis that $b = 1$. •

If you have had some prior statistical training, you may well have already encountered the resolution of these problems. You may know, for example, that provided $\boldsymbol{X}$ is of full rank and no linear constraints are imposed on $\boldsymbol{\beta}$, the *best (minimum variance) linear unbiased estimator (BLUE)* of $\sum_{1\le i\le p} c_i\beta_i$ is $\sum_{1\le i\le p} c_i\hat\beta_i$, where

$$(\hat\beta_1, \dots, \hat\beta_p)^T = \boldsymbol{C}\boldsymbol{X}^T\boldsymbol{Y}, \quad \text{with} \quad \boldsymbol{C} = \boldsymbol{A}^{-1}, \quad \boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{X};$$

this is called the *Gauss-Markov theorem.*

In this book we will be studying the GLM from a geometric point of view, using linear algebra in place of matrix algebra. Although we will not reach any conclusions that could not be obtained solely by matrix techniques, the basic ideas will emerge more clearly. With the added intuitive feeling and mathematical insight this provides, one will be better able to understand old results and formulate and prove new ones.

From a geometric perspective, the GLM may be described as follows, using some terms that will be defined in subsequent chapters. One is given a *random vector* Y taking values in some given *inner product space* $(V, \langle\cdot,\cdot\rangle)$. It is assumed that Y has a *weakly spherical covariance operator* and the *mean* μ of Y lies in a given *manifold* M of V; for purposes of testing, it is further assumed that Y is *normally distributed.* One desires to estimate μ (or *linear functionals* of μ) and to test hypotheses such as $\mu \in M_0$, where M_0 is a given *submanifold* of M. The Gauss-Markov theorem says that the BLUE of the linear functional $\psi(\mu)$ is $\psi(\hat\mu)$, where $\hat\mu$ is the *orthogonal projection* of Y onto M. As we will see, this geometric description of the problem encompasses the matricial formulation of the GLM not only as it is set out above (take, for example, $V = \mathbb{R}^n$, $\langle\cdot,\cdot\rangle$ = dot-product, $Y = \boldsymbol{Y}$, $\mu = \boldsymbol{\mu}$, and M = the subspace of $\mathbb{R}^n$ spanned by the columns of the design matrix $\boldsymbol{X}$), but also in cases where $\boldsymbol{X}$ is of less than full rank and/or linear constraints are imposed on the β_i's.

2. An illustrative example

To illustrate the differences between the matricial and geometric approaches, we compare the ways in which one establishes the independence of

$$\bar Y = \hat\mu = \frac{\sum_{1\le i\le n} Y_i}{n} \quad \text{and} \quad s^2 = \hat\sigma^2 = \frac{1}{n-1}\sum_{1\le i\le n}(Y_i - \bar Y)^2$$

in the one-sample problem

$$\boldsymbol{Y}^{n\times 1} \sim N(\mu \boldsymbol{e}, \sigma^2 \boldsymbol{I}^{n\times n}) \quad \text{with} \quad \boldsymbol{e} = (1, 1, \ldots, 1)^T. \tag{2.1}$$

(The vector $\boldsymbol{e}$ is called the *equiangular vector*.)

The classical matrix proof, which uses some facts about multivariate normal distributions, runs like this. Let $\boldsymbol{B}^{n\times n} = (b_{ij})$ be the matrix

$$\begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} & 0 & \cdots & 0 & 0 \\ \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{1}{\sqrt{12}} & \frac{-3}{\sqrt{12}} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & \frac{1}{\sqrt{n(n-1)}} & \frac{-(n-1)}{\sqrt{n(n-1)}} \end{pmatrix}.$$

Note that the rows (and columns) of $\boldsymbol{B}$ are orthonormal $\left(\sum_{1\le k\le n} b_{ik}b_{jk} = \delta_{ij} \equiv \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \ne j \end{cases}\right)$ and that the first row is $\frac{1}{\sqrt{n}}\,\boldsymbol{e}^T$. Set

$$\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}.$$

Then

$$\boldsymbol{Z} \sim N(\boldsymbol{\nu}, \boldsymbol{\Sigma})$$

with

$$\boldsymbol{\nu} = \boldsymbol{B}(\mu\boldsymbol{e}) = \mu\boldsymbol{B}\boldsymbol{e} = (\sqrt{n}\,\mu, 0, \ldots, 0)^T$$

and

$$\boldsymbol{\Sigma} = \boldsymbol{B}(\sigma^2\boldsymbol{I})\boldsymbol{B}^T = \sigma^2\boldsymbol{B}\boldsymbol{B}^T = \sigma^2\boldsymbol{I};$$

that is, $Z_1, Z_2, \ldots, Z_n$ are independent normal random variables, each with variance σ^2, $E(Z_1) = \sqrt{n}\,\mu$, and $E(Z_j) = 0$ for $2 \le j \le n$. Moreover,

$$Z_1 = \frac{1}{\sqrt{n}} \sum\nolimits_{1\le i\le n} Y_i = \sqrt{n}\,\bar{Y}, \text{ or } \bar{Y} = \frac{Z_1}{\sqrt{n}},$$

while

$$\begin{aligned} (n-1)s^2 &= \sum\nolimits_{1\le i\le n} (Y_i - \bar{Y})^2 = \sum\nolimits_{1\le i\le n} Y_i^2 - n\bar{Y}^2 \\ &= \sum\nolimits_{1\le i\le n} Y_i^2 - Z_1^2 = \sum\nolimits_{1\le i\le n} Z_i^2 - Z_1^2 = \sum\nolimits_{2\le i\le n} Z_i^2 \end{aligned} \tag{2.2}$$

because

$$\sum\nolimits_{1\le i\le n} Z_i^2 = \boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{Y}^T\boldsymbol{Y} = \sum\nolimits_{1\le i\le n} Y_i^2.$$

This gives the independence of $\bar{Y}$ and s^2, and it is an easy step to get the marginal distributions: $\bar{Y} \sim N(\mu, \sigma^2/n)$ and $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

What is the nature of the transformation $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{Y}$? Let $\boldsymbol{b}_1 = \boldsymbol{e}/\sqrt{n}$, $\boldsymbol{b}_2, \ldots, \boldsymbol{b}_n$ denote the transposes of the rows of $\boldsymbol{B}$. The coordinates of $\boldsymbol{Y} = \sum_{1\le j\le n} C_j\boldsymbol{b}_j$ with respect to this new orthonormal basis for $\mathbb{R}^n$ are given by

$$C_i = \boldsymbol{b}_i^T\boldsymbol{Y} = Z_i, \quad i = 1, \ldots, n.$$

The effect of the change of coordinates $\boldsymbol{Y} \to \boldsymbol{Z}$ is to split $\boldsymbol{Y}$ into its components along, and orthogonal to, the equiangular vector $\boldsymbol{e}$.

Now I will show you the geometric proof, which uses some properties of (weakly) spherical normal random vectors taking values in an inner product space $(V, \langle\cdot,\cdot\rangle)$, here $(\mathbb{R}^n$, dot-product). The assumptions imply that $\boldsymbol{Y}$ is spherical normally distributed about its mean $E(\boldsymbol{Y})$ and $E(\boldsymbol{Y})$ lies in the manifold M spanned by $\boldsymbol{e}$. Let P_M denote orthogonal projection onto M and Q_M orthogonal projection onto the orthogonal complement $M^\perp$ of M. Basic distribution theory says that $P_M\boldsymbol{Y}$ and $Q_M\boldsymbol{Y}$ are independent. But

$$P_M\boldsymbol{Y} = \frac{\langle \boldsymbol{e}, \boldsymbol{Y}\rangle}{\langle \boldsymbol{e}, \boldsymbol{e}\rangle}\,\boldsymbol{e} = \bar{Y}\boldsymbol{e} \tag{2.3}$$

and

$$Q_M\boldsymbol{Y} = \boldsymbol{Y} - P_M\boldsymbol{Y} = \boldsymbol{Y} - \bar{Y}\boldsymbol{e} = (Y_1 - \bar{Y}, \ldots, Y_n - \bar{Y})^T;$$

it follows that $\bar{Y}$ and $(n-1)s^2 = \sum_{1\le i\le n}(Y_i - \bar{Y})^2 = \|Q_M\boldsymbol{Y}\|^2$ are independent. Again it is an easy matter to get the marginal distributions.

To my way of thinking, granted the technical apparatus, the second proof is clearer, being more to the point. The first proof does the same things, but (to the uninitiated) in an obscure manner.

3. Notational conventions

The chapters are organized into sections. Within each section of the current chapter, enumerated items are numbered consecutively in the form

$$(\textit{section\_number}.\textit{item\_number}).$$

References to items in a different chapter take the expanded form

$$(\textit{chapter\_number}.\textit{section\_number}.\textit{item\_number}).$$

For example, (2.4) refers to the 4^{th} numbered item (which may be an example, exercise, theorem, formula, or whatever) in the 2^{nd} section of the current chapter, while (6.1.3) refers to the 3^{rd} numbered item in the 1^{st} section of the 6^{th} chapter.

Each exercise is assigned a difficulty level using the syntax

Exercise $[d]$,

where d is an integer in the range 1 to 5 — the larger is d, the harder the exercise. The value of d depends both on the intrinsic difficulty of the exercise and the length of time needed to write up the solution.

To help distinguish between the matricial and geometric points of view, matrices, including row and column vectors, are written in ***italic boldface*** type while linear transformations and elements of abstract vector spaces are written simply in *italic* type. We speak, for example, of the design matrix $\boldsymbol{X}$ but of vectors v and w in an inner product space V.

The end of a proof is marked by a ∎, of an example by a •, of an exercise by a ⋄, and of a part of problem set by a ∘.

CHAPTER 2

TOPICS IN LINEAR ALGEBRA

In this chapter we discuss some topics from linear algebra that play a central role in the geometrical analysis of the GLM. The notion of orthogonal projection in an inner product space is introduced in Section 2.1 and studied in Section 2.2. A class of orthogonal decompositions that are useful in the design of experiments is studied in Section 2.3. The spectral representation of self-adjoint transformations is developed in Section 2.4. Linear and bilinear functionals are discussed in Section 2.5. The chapter closes with a problem set in Section 2.6, followed by an appendix containing a brief review of the basic definitions and facts from linear algebra with which we presume the reader is already familiar.

1. Orthogonal projections

Throughout this book we operate in the context of a finite-dimensional *inner product space* $(V, \langle\cdot,\cdot\rangle)$ — V is a finite-dimensional vector space and $\langle\cdot,\cdot\rangle\colon V \times V \to \mathbb{R}$ is an *inner product:*

(i) *(positive-definiteness)* $\langle x, x\rangle \geq 0$ for all $x \in V$ and $\langle x, x\rangle = 0$ if and only if $x = 0$.

(ii) *(symmetry)* $\langle x, y\rangle = \langle y, x\rangle$ for all $x, y \in V$.

(iii) *(bilinearity)* For all $c_1, c_2 \in \mathbb{R}$ and $x_1, x_2, x, y_1, y_2, y \in V$, one has

$$\langle c_1 x_1 + c_2 x_2, y\rangle = c_1\langle x_1, y\rangle + c_2\langle x_2, y\rangle$$
$$\langle x, c_1 y_1 + c_2 y_2\rangle = c_1\langle x, y_1\rangle + c_2\langle x, y_2\rangle.$$

The canonical example is $V = \mathbb{R}^n$ endowed with the dot-product

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x} \cdot \boldsymbol{y} = \sum\nolimits_{1\leq i\leq n} x_i y_i = \boldsymbol{x}^T \boldsymbol{y}$$

for $\boldsymbol{x} = (x_1, \ldots, x_n)^T$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. Unless specifically stated to the contrary, we always view $\mathbb{R}^n$ as endowed with the dot-product.

Two vectors x and y in V are said to be *perpendicular*, or *orthogonal* (with respect to $\langle\cdot,\cdot\rangle$), if

$$\langle x, y\rangle = 0;$$

one writes

$$x \perp y.$$

The quantity

$$\|x\| = \sqrt{\langle x, x\rangle}$$

is called the *length*, or *norm*, of x. The squared length of the sum of two vectors is given by

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x, x+y\rangle + \langle y, x+y\rangle \\ &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2,\end{aligned}$$

which reduces to the *Pythagorean theorem*,

$$\|x+y\|^2 = \|x\|^2 + \|y\|^2, \tag{1.1}$$

when $x \perp y$.

1.2 Exercise [1]. Let $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ be two nonzero vectors in $\mathbb{R}^2$ and let θ be the angle between them, measured counter-clockwise from v_1 to v_2. Show that

$$\cos(\theta) = \frac{\langle v_1, v_2\rangle}{\|v_1\|\,\|v_2\|}$$

and deduce that $\boldsymbol{v}_1 \perp \boldsymbol{v}_2$ if and only if $\theta = 90^\circ$ or 270°.
[Hint: Use the identity $\cos(\theta_2 - \theta_1) = \cos(\theta_1)\cos(\theta_2) + \sin(\theta_1)\sin(\theta_2)$.] ◇

1.3 Exercise [3]. Let $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ be the vectors in $\mathbb{R}^n$ given by

$$\boldsymbol{x}_i = 1,\ \boldsymbol{y}_i = i - \tfrac{n+1}{2},\ \text{and}\ \boldsymbol{z}_i = (i - \tfrac{n+1}{2})^2 - \tfrac{n^2-1}{12} \quad \text{for } 1 \le i \le n. \tag{1.4}$$

Show that $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are mutually orthogonal and span the same subspace as do $(1, 1, \ldots, 1)^T$, $(1, 2, \ldots, n)^T$, and $(1, 4, \ldots, n^2)^T$. Exhibit $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ explicitly for $n = 5$. ◇

1.5 Exercise [1]. Let $\mathcal{O}\colon V \to V$ be a linear transformation. Show that $\mathcal{O}$ preserves lengths, that is,

$$\|\mathcal{O}x\| = \|x\| \quad \text{for all } x \in V$$

if and only if it preserves inner products, that is,

$$\langle \mathcal{O}x, \mathcal{O}y\rangle = \langle x, y\rangle \quad \text{for all } x, y \in V.$$

Such a transformation is said to be *orthogonal*.
[Hint: Observe that

$$\langle x, y\rangle = \frac{\|x+y\|^2 - \|x\|^2 - \|y\|^2}{2} \tag{1.6}$$

for $x, y \in V$.] ◇

1.7 Exercise [2]. (1) Let v_1 and v_2 be elements of V. Show that $v_1 = v_2$ if and only if $\langle v_1, w\rangle = \langle v_2, w\rangle$ for each $w \in V$, or just for each w in some basis for V. (2) Let T_1 and T_2 be two linear transformations of V. Show that $T_1 = T_2$ if and only if $\langle v, T_1 w\rangle = \langle v, T_2 w\rangle$ for each v and w in V, or just for each v and w in some basis for V. ◇

As intimated in the Introduction, the notion of orthogonal projection onto subspaces of V plays a key role in the study of the GLM. We begin our study of projections with the following seminal result.

1.8 Theorem. *Suppose that M is a subspace of V and that $x \in V$. There is exactly one vector $m \in M$ such that the residual $x - m$ is orthogonal to M:*

$$(x - m) \perp y \quad \text{for all } y \in M, \tag{1.9}$$

or, equivalently, such that m is closest to x:

$$\|x - m\| = \inf\{\, \|x - y\| : y \in M \,\}. \tag{1.10}$$

The proof will be given shortly. The unique $m \in M$ such that (1.9) and (1.10) hold is called the *orthogonal projection of x onto M*, written $P_M x$, and the mapping P_M that sends $x \in V$ to $P_M x$ is called *orthogonal projection onto M*. In the context of $\mathbb{R}^n$ with $\boldsymbol{x} = (x_i)$ and $\boldsymbol{m} = (m_i)$, (1.9) reads

$$\sum\nolimits_{1\le i\le n}(x_i - m_i)y_i = 0 \quad \text{for all } \boldsymbol{y} = (y_i) \in M,$$

while (1.10) is the *least squares characterization* of $\boldsymbol{m}$:

$$\sum\nolimits_{1\le i\le n}(x_i - m_i)^2 = \inf\Big\{ \sum\nolimits_{1\le i\le n}(x_i - y_i)^2 : \boldsymbol{y} \in M \Big\}.$$

1.11 Exercise [1]. Let M be a subspace of V. (a) Show that

$$\|x - P_M x\|^2 = \|x\|^2 - \|P_M x\|^2 \tag{1.12}$$

for all $x \in V$. (b) Deduce that

$$\|P_M x\| \le \|x\| \tag{1.13}$$

for all $x \in V$, with equality holding if and only if $x \in M$. ◇

Proof of Theorem 1.8. (1.9) *implies* (1.10): Suppose $m \in M$ satisfies (1.9). Then for all $y \in M$ the Pythagorean theorem gives

$$\|x - y\|^2 = \|(x - m) + (m - y)\|^2 = \|x - m\|^2 + \|m - y\|^2,$$

so m satisfies (1.10).

(1.10) *implies* (1.9): Suppose $m \in M$ satisfies (1.10). Then, for any $0 \ne y \in M$ and any $\delta \in \mathbb{R}$,

$$\|x - m\|^2 \le \|x - m + \delta y\|^2 = \|x - m\|^2 + 2\delta\langle x - m, y\rangle + \delta^2\|y\|^2;$$

this relation forces $\langle x - m, y\rangle = 0$.

Uniqueness: If

$$m_1 + y_1 = x = m_2 + y_2$$

with $m_i \in M$ and $y_i \perp M$ for $i = 1$ and 2, then the vector

$$m_1 - m_2 = y_2 - y_1$$

lies in M and is perpendicular to M, and so it is perpendicular to itself:

$$0 = \langle m_1 - m_2, m_1 - m_2 \rangle = \|m_1 - m_2\|^2,$$

whence $m_1 = m_2$ by the positive-definiteness of $\langle \cdot, \cdot \rangle$.

Existence: We will show in a moment that M has a basis $m_1, \ldots, m_k$ consisting of mutually orthogonal vectors. For any such basis the generic $m = \sum_{1 \le j \le k} c_j m_j$ in M satisfies

$$(x - m) \perp M$$

if and only if $x - m$ is orthogonal to each m_i, that is, if and only if

$$\langle m_i, x \rangle = \langle m_i, \sum\nolimits_{1 \le j \le k} c_j m_j \rangle = \sum\nolimits_{1 \le j \le k} \langle m_i, m_j \rangle c_j = \langle m_i, m_i \rangle c_i$$

for $1 \le i \le k$. It follows that we can take

$$P_M x = \sum\nolimits_{1 \le i \le k} \frac{\langle m_i, x \rangle}{\langle m_i, m_i \rangle}\, m_i. \tag{1.14}$$

To produce an orthogonal basis for M, let $m_1^*, \ldots, m_k^*$ be any basis for M and inductively define new basis vectors $m_1, \ldots, m_k$ by the recipe $m_1 = m_1^*$ and

$$\begin{aligned} m_j &= m_j^* - P_{[m_1^*, \ldots, m_{j-1}^*]} m_j^* = m_j^* - P_{[m_1, \ldots, m_{j-1}]} m_j^* \\ &= m_j^* - \sum\nolimits_{1 \le i \le j-1} \frac{\langle m_j^*, m_i \rangle}{\langle m_i, m_i \rangle}\, m_i \end{aligned} \tag{1.15}$$

for $j = 2, \ldots, k$; here $[m_1^*, \ldots, m_{j-1}^*]$ denotes the span of $m_1^*, \ldots, m_{j-1}^*$ and $[m_1, \ldots, m_{j-1}]$ denotes the (identical) span of $m_1, \ldots, m_{j-1}$ (see Subsection 2.6.1). ∎

The recursive scheme for cranking out the m_j's above is called *Gram-Schmidt orthogonalization.* As a special case of (1.14) we have the following simple, yet key, formula for projecting onto a one-dimensional space:

$$P_{[m]} x = \frac{\langle x, m \rangle}{\langle m, m \rangle}\, m \qquad \text{for } m \ne 0. \tag{1.16}$$

1.17 Example. In the context of $\mathbb{R}^n$ take $\boldsymbol{m} = \boldsymbol{e} \equiv (1, \ldots, 1)^T$. Formula (1.16) then reads

$$P_{[\boldsymbol{e}]} \boldsymbol{x} = \frac{\langle \boldsymbol{x}, \boldsymbol{e} \rangle}{\langle \boldsymbol{e}, \boldsymbol{e} \rangle}\, \boldsymbol{e} = \frac{\sum_i x_i}{\sum_i 1}\, \boldsymbol{e} = \bar{x}\boldsymbol{e} = (\bar{x}, \ldots, \bar{x})^T;$$

we used this result in the Introduction (see (1.2.3)). Formula (1.12) reads

$$\sum\nolimits_i (x_i - \bar{x})^2 = \|\boldsymbol{x} - \bar{x}\boldsymbol{e}\|^2 = \|\boldsymbol{x}\|^2 - \|\bar{x}\boldsymbol{e}\|^2 = \sum\nolimits_i x_i^2 - n\bar{x}^2;$$

this is just the computing formula for $(n-1)s^2$ used in (1.2.2). According to (1.10), $c = \bar{x}$ minimizes the sum of squares $\sum_i (x_i - c)^2$. •

1.18 Exercise [2]. Use the preceding techniques to compute $P_M \boldsymbol{y}$ for $\boldsymbol{y} \in \mathbb{R}^n$ and $M = [\boldsymbol{e}, (x_1, \ldots, x_n)^T]$, the manifold spanned by the columns of the design matrix for simple linear regression. ◇

1.19 Exercise [3]. Show that the transposes of the rows of the matrix $\boldsymbol{B}$ of Section 1.2 result from first applying the Gram-Schmidt orthogonalization scheme to the vectors $\boldsymbol{e}$, $(1, -1, 0, \ldots, 0)^T$, $(0, 1, -1, 0, \ldots, 0)^T$, $\ldots$, $(0, \ldots, 0, 1, -1)^T$ in $\mathbb{R}^n$ and then normalizing to unit length. ◇

1.20 Exercise [2]. Suppose $x, y \in V$. Prove the *Cauchy-Schwarz inequality*:

$$|\langle x, y \rangle| \le \|x\| \, \|y\|, \tag{1.21}$$

with equality if and only if x and y are linearly dependent. Deduce *Minkowski's inequality*:

$$\|x + y\| \le \|x\| + \|y\|. \tag{1.22}$$

[Hint: For (1.21), take $m = y$ in (1.16) and use part (b) of Exercise 1.11.] ◇

1.23 Exercise [4]. Define $d: V \times V \to \mathbb{R}$ by $d(x, y) = \|y - x\|$. Show that d is a metric on V such that the set $\{\, x \in V : \|x\| = 1 \,\}$ is compact. ◇

1.24 Exercise [3]. Show that for any two subspaces M and N of V,

$$\sup\{\, \|P_M x\| : x \in N \text{ and } \|x\| = 1 \,\} \le 1, \tag{1.25}$$

with equality holding if and only if M and N have a nonzero vector in common.
[Hint: A continuous real-valued function on a compact set attains its maximum.] ◇

To close this section we generalize (1.14) to cover the case of an arbitrary basis for M. Suppose then that the basis vectors $m_1, \ldots, m_k$ are not necessarily orthogonal and let $x \in V$. As in the derivation of (1.14),

$$P_M x = \sum\nolimits_j c_j m_j$$

is determined by the condition

$$m_i \perp (x - P_M x) \quad \text{for } i = 1, \ldots, k,$$

that is, by the so-called *normal equations*

$$\sum\nolimits_{1 \le j \le k} \langle m_i, m_j \rangle c_j = \langle m_i, x \rangle, \quad i = 1, \ldots, k. \tag{1.26}$$