The Geometry of the General Linear Model

by

Erik D. Jacobson

(Under the direction of Malcolm R. Adams, Edward Azoff, and Theodore Shifrin.)

Abstract

Two complementary geometric interpretations of data are used to discuss topics from elementary statistics including random variables, vectors of random variables, expectation, mean, variance, and the normal, F, and t probability distributions. The geometry of the general linear model and the associated hypothesis testing is developed, followed by a geometrically oriented discussion of the analysis of variance, simple regression, and multiple regression using examples. Geometry affords a rich discussion of orthogonality, multicollinearity, and suppressor variables, as well as multiple, partial, and semi-partial correlation. The last chapter describes the mathematical application of homogeneous coordinates and perspective projections in the computer program used to generate the representations of data vectors for several figures in this text.

Index words: Projections, geometry, homogeneous coordinates, statistics, linear models, regression

The Geometry of the General Linear Model

by

Erik D. Jacobson

B.A., Dartmouth College, 2004

A Thesis Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Master of Arts

Athens, Georgia

2011

The Geometry of the General Linear Model

by

Erik D. Jacobson

Approved:

Major Professor:	Theodore Shifrin
Committee:	Malcolm R. Adams Edward Azoff

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2011

Acknowledgments

I recognize the insight, ideas, and encouragement offered by my thesis committee: Theodore Shifrin, Malcolm R. Adams, and Edward Azoff and by Jonathan Templin. All four gamely dug into novel material, entertained ideas from other disciplines, and showed admirable forbearance as deadlines slipped past and the project expanded. I cannot thank them enough.

Contents

List of Figures

List of Tables

Introduction

Many of the most popular and useful techniques for developing statistical models are subsumed by the *general linear model*. This thesis presents the general linear model and the accompanying strategy of hypothesis testing from a primarily geometric point of view, detailing both the standard view of data as points in a space defined by the variables (data-as-points) and the less common perspective of data as vectors in a space defined by the individuals (data-as-vectors). I also develop the relevant statistical ideas geometrically and use geometric arguments to relate the general linear model to analysis of variance (ANOVA) models, and to correlation and regression models.

My approach to this material is original, although the central ideas are not new. The standard treatment of this material is predominantly algebraic with a minimal (in the case of regression) or nonexistent (in the case of ANOVA) discussion of the data-as-points geometry and no mention of the data-as-vectors approach. In addition, these models are often taught separately in different courses and never related. Only a very few texts present the data-as-vectors approach (although historically this was the geometry emphasized by early statisticians), and all most all of these texts are written for graduate students in statistics and presume sophisticated understandings of many statistical and mathematical concepts. Another major limitation of these texts is the quality of the drawings, which are schematic and not based on examples of real data. By contrast, I emphasize geometry, particularly the data-as-vectors perspective, in order to introduce common statistical models and to explain how they are in fact closely related to one another. I am able to use precise drawings of examples with real data because of the DataVectors computer program I developed to generate interactive representations of vector geometry (available upon request: ejacobsn@uga.edu). I am not aware of any other program that can generate these kinds of

representations. Although it is unlikely they would be useful representations for reports of original research, the drawings produced by the program (and the interactivity of the program itself) have potential as powerful pedagogical tools for those learning to think about linear models using the data-as-vectors geometric perspective.

I have drawn on many sources in preparing this manuscript. In particular, Thomas D. Wickens' (1995) text *The Geometry of Multivariate Statistics* and David J. Saville's and Graham R. Wood's (1991) *Statistical Methods: The Geometric Approach* introduced me to the data-as-vectors geometry. Other statistical texts that were particularly helpful include Ronald Christensen's (1996) *Plane Answers to Complex Questions*; S. R. Searle's (1971) *Linear Models*; Michael H. Kutner and colleagues' (2005) *Applied Linear Statistical Models*; and George Casella's and Roger L. Berger's (2002) text titled *Statistical Inference*. Using *R* to work out examples was made possible by Julian J. Faraway's (2004) *Linear Models with R*. It is worth mentioning that Elazar J. Pedhazur's (1997) *Multiple Regression in Behavioral Research: Explanation and Prediction* led me to initially ask the questions that this manuscript addresses and informed my understanding of applied multiple regression techniques. The discussion of projections and homogeneous coordinates in Theodore Shifrin and Malcolm R. Adams' (2002) *Linear Algebra: A Geometric Approach* and Ken Shoemake's landmark 1992 talk on the arcball control at the Graphics Interface conference in Vancouver, Canada were invaluable for writing the DataVectors program. In the work that follows, all of the discussion, examples, and figures are my own. Except where noted, the proofs are my own work or my adaptation of standard proofs that are used without citation in two or more standard references.

The statistical foundations of the subsequent chapters are developed in Chapter 1. The topics addressed include random variables, vectors of random variables, expectation, mean, variance, and the normal, F-, and t- probability distributions. Two complimentary geometric interpretations of data are presented and contrasted throughout. In Chapter 2, we develop the geometry of the general linear hypothesis testing by examining a simple experiment. In Chapter 3, we turn to several examples of analyses from the ANOVA framework and illustrate the geometric meaning of dummy variables and contrasts. The relationships between these contrasts, the F-ratio, and hypothesis testing are explored.

Chapter 4 describes simple regression and correlation analysis using two geometric interpretations of data. Variable transformations provide a way to make simple regression models more general and enable the development of models for non-linear data. In Chapter 5, we discuss multiple regression from a geometric point of view. The geometric perspectives developed in the previous chapters afford a rich discussion of orthogonality, multicollinearity, and suppressor variables, as well as multiple, partial, and semi-partial correlation. Chapter 6 takes us through a tour of four probability distributions that are referenced in the text and complements the statistical foundations presented in the first chapter.

The last chapter describes the mathematical basis of the DataVectors program used to generate some of the figures in this text. Points in $\mathbb{R}^3$ can be represented using homogeneous coordinates which facilitate affine translations of these points via matrix multiplication. Perspective projections of $\mathbb{R}^3$ to an arbitrary plane can also be realized via matrix multiplication directly from Euclidean or homogeneous coordinates. Computer systems producing perspective representations of 3-dimensional objects often perform an intermediate transform of the viewing frustum in $\mathbb{R}^3$ to an appropriately scaled parallelepiped in perspective space, retaining some information about relative depth.

Chapter 1

Statistical Foundations and Geometry

Applied statistics is fundamentally concerned with finding quantitative relationships among *variables*, the observed features of individuals in a population. Usually, data from an entire population are not available and instead these relationships must be inferred from a *sample*, a subset of the population. We denote the size of the sample by n. Variables are categorized as independent or dependent; the independent variables are used to predict or explain the dependent variables. Much of the work of applied statistics is finding appropriate models that relate independent and dependent variables and making disciplined decisions about the hypothesized parameters of these models. After discussing some foundational ideas of statistics from two different geometric perspectives (including random variables, expectation, mean, and variance), statistical models and hypothesis testing are introduced by means of an illustrative example.

	Variables				
	Dependent	Independent			
Individuals	Var_1	Var_2	Var_3	$\cdots$	Var_m
Individual_1	$\text{Obs}_{1,1}$	$\text{Obs}_{1,2}$	$\text{Obs}_{1,3}$	$\cdots$	$\text{Obs}_{1,m}$
Individual_2	$\text{Obs}_{2,1}$	$\text{Obs}_{2,2}$	$\text{Obs}_{2,3}$	$\cdots$	$\text{Obs}_{2,m}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Individual_n	$\text{Obs}_{n,1}$	$\text{Obs}_{n,2}$	$\text{Obs}_{n,3}$	$\cdots$	$\text{Obs}_{n,m}$

Table 1.1: A generic data set with one dependent variable, $m-1$ independent variables, and n observations for each variable.

1.1 Two geometric interpretations of data

One canonical representation of a data set is a table of observations (see Table 1.1). The rows of the table correspond to individuals and the columns of the table correspond to the variables of interest. Each entry in the table is a single observation (a real number) of a particular variable for a particular individual. This representation can be taken as an $n \times m$ matrix over $\mathbb{R}$ where n is the number of individuals and m is the number of variables comprising the data set. Then the columns and rows of the data matrix can be treated as vectors in $\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.

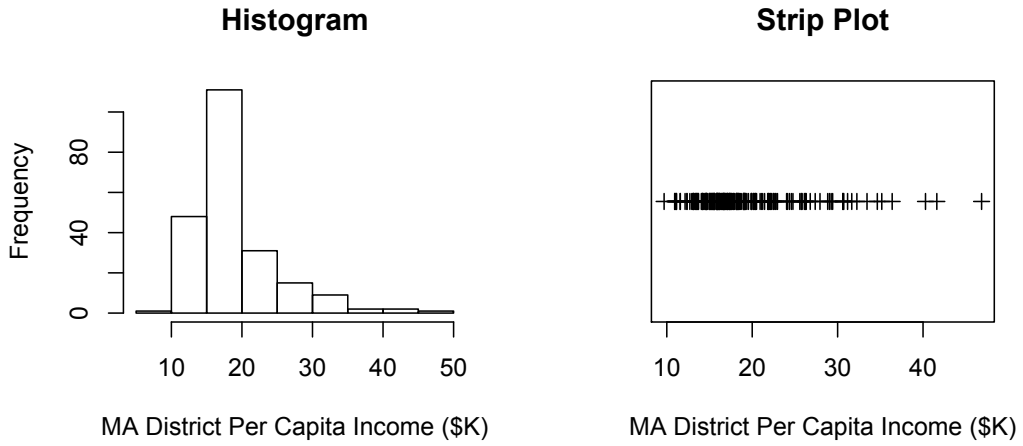


Figure 1.1: Strip plots (a one-dimensional scatterplot) can illustrate observations of a single variable, but histograms convey the variable's distribution more clearly.

Data are usually interpreted geometrically by considering each row vector as a point in Euclidean space with coordinate axes that correspond to the variables in the data set. Thus, each individual can be located in $\mathbb{R}^m$ with the i^{th} coordinate given by that individual's i^{th} variable observation. This space is called *variable space*. When only one variable is involved, the space is one dimensional and a strip chart (and more commonly the histogram which more clearly conveys the variable's distribution) illustrates this representation (see Figure 1.1). Scatterplots can illustrate data sets with up to three variables (see Figure 1.2).

There is a second geometric interpretation. The column vectors of the data matrix can be understood as vectors in $\mathbb{R}^n$ and used to generate useful *vector diagrams* (see Figure 1.3). Vector diagrams represent (subspaces) of *individual space*, and offer a complementary geometric

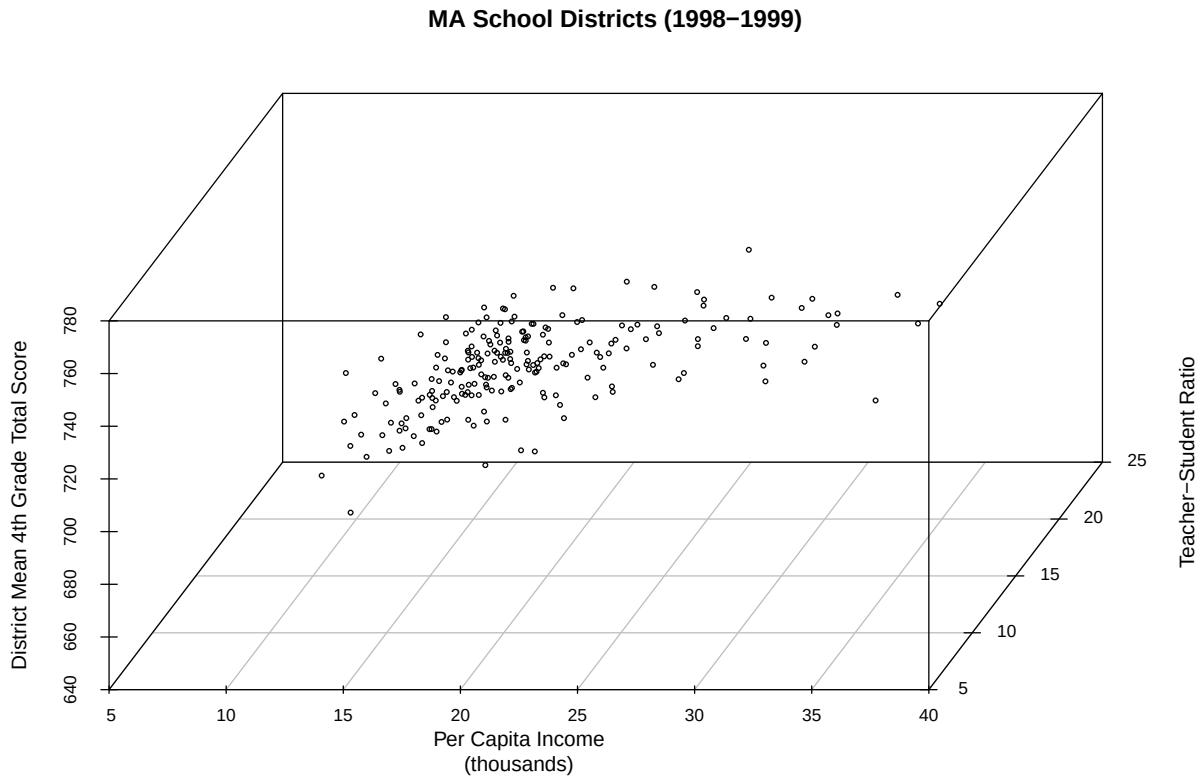


Figure 1.2: Scatterplots can illustrate multivariate data.

interpretation of the data. We use boldface letters to denote the vectors, and it is customary for $\mathbf{y}$ to denote the dependent variable(s) (called Var_1 in Table 1.1), and for $\mathbf{x}_i$ where i ranges between 1 and $m-1$, to denote the independent variables (called Var_2, Var_3, etc. in Table 1.1). The reader likely notices one immediate hurdle we face when interpreting the data matrix as vectors in $\mathbb{R}^n$—the dimension of the individual space is equal to the number of individuals, and, for almost every useful data set, this far exceeds that which we visualize, let alone reasonably illustrate on a two-dimensional page or computer screen. In practice, we are limited to illustrating up to 3-dimensional subspaces of individual space. In cases where the number of variables (and hence the dimension of smallest-dimensioned subspace of interest) is greater than three, planes and lines must represent higher-dimensional subspaces and vector diagrams become schematic.

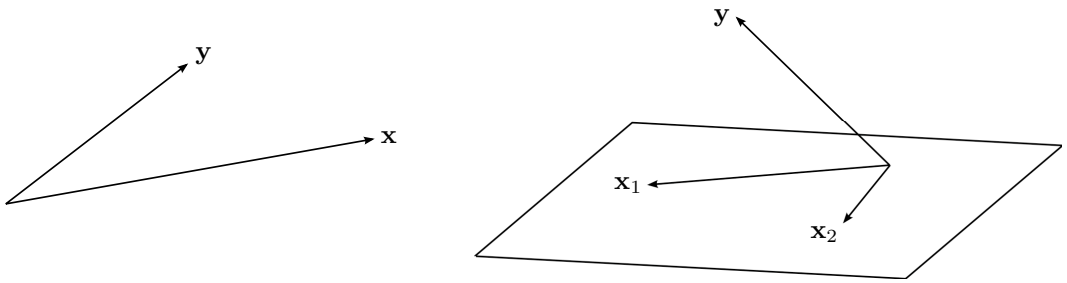


Figure 1.3: Vector diagram representations of variable vectors that span two and three dimensional subspaces of individual space ($\mathbb{R}^n$).

1.2 Random variables and expectation

In statistics, a random experiment is any process with an *outcome* that cannot be predicted before the experiment occurs. The set of outcomes for a random experiment is called the *sample space* and denoted Ω. The subsets of the sample space are called *events* and this collection is denoted $\mathcal{S}$. (Technical considerations may prohibit the inclusion of *all* subsets of Ω; the collection of events $\mathcal{S}$ must be a σ-algebra.) A *probability measure* P is a real-valued function defined on $\mathcal{S}$, such that $P(A) \geq 0$, for all $A \in \mathcal{S}$ and $P(\mathcal{S}) = 1$. In addition, P satisfies the *countable additivity axiom*: If $\{A_n : n \in \mathbb{N}\}$ is a countable, pairwise disjoint collection of events then

$$P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n).$$

In other words, $(\Omega, \mathcal{S}, P)$ is a measure space with probability measure P.

A *real random variable*, often denoted with capital Roman letter, is a real-valued function defined on Ω, e.g., $Y : \Omega \to \mathbb{R}$. Each random variable is associated with its *cumulative distribution function*, $F_Y : \mathbb{R} \to \mathbb{R}$ where $F_Y(y) = P(Y \leq y)$, where $Y \leq y$ indicates the set $\{\omega \in \Omega : Y(\omega) \leq y\}$. The cumulative distribution function allows the computation of the probability of any set in $\mathcal{S}$. The function Y is a *continuous* random variable if there is a function $f_Y : \mathbb{R} \to \mathbb{R}$ satisfying $F_Y(y) = \int_{-\infty}^{y} f_Y(x)dx$; f_Y is called the *probability density function.* In an analogous way, a *discrete* random variable has an associated density function $m_Y : \mathbb{R} \to \mathbb{R}$ that is non-zero only at countably many points and satisfies $F_Y(y) = \sum_{x\leq y} m_Y(x)$.

One of the most important concepts in statistics is that of the *expected value* $E(Y)$ of a random variable. In the case where Y is discrete, the expected value can be defined as a weighted average, $E(Y) = \sum_{y \in \mathbb{R}} y m_Y(y)$. For example, the expected value of the random variable that assigns each die roll event to its face-value is $E(Y) = \sum_{y \in \{1,2,\ldots,6\}} \frac{y}{6} = \frac{21}{6}$. If Y is continuous, then we define $E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$.

Expectation has the following three properties, stated without proof.

Theorem 1.2.1. *For any random variables X and Y, any a in $\mathbb{R}$, and any function $g : \mathbb{R} \to \mathbb{R}$,*

- $E(X + Y) = E(X) + E(Y)$
- $E(aY) = aE(Y)$
- $E(g \circ Y) = \int_{-\infty}^{\infty} g(y) f_Y(y) dy$

In general, the expectation of a linear combination of random variables is the corresponding linear combination of the respective expected values, but for our purposes we only need the weaker result involving a single random variable that is stated next.

Corollary 1.2.2. *For any random variable Y, any a in $\mathbb{R}$, and any functions g_1 and g_2*

- $E(g_1 \circ Y + g_2 \circ Y) = E(g_1 \circ Y) + E(g_2 \circ Y)$, *and*
- $E(ag_1 \circ Y) = aE(g_1 \circ Y)$.

Expectation is useful for understanding several other important concepts. Suppose $X : \Omega_X \to \mathbb{R}$ and $Y : \Omega_Y \to \mathbb{R}$ are two random variables with associated probability measures P_X and P_Y. The joint probability of the event (A and B), where $A \subset \Omega_X$ and $B \subset \Omega_Y$, is defined $P_X \times P_Y(A \text{ and } B) = P_X(A)P_Y(B)$, and the variables have a *joint probability distribution* defined $F_{XY}(x, y) = P_1 \times P_2(X \leq x \text{ and } Y \leq y)$. The expected value of a joint probability distribution is related to the expected value of the component variables. Two random variables are said to be *independent* if $F_{XY}(A \text{ and } B) = P_X(A)P_Y(B)$. Whenever two random variables, X and Y, are independent, the following statement holds:

$$E(XY) = E(X)E(Y)$$

Note that *independent* is used with several different meanings in this text. *Independent variables* in models are those used to predict or explain the dependent variable, but we will see that our methods require us to assume that the sample of n observations of the dependent variable are the realized values of the *independent random variables* $Y_1, \ldots, Y_n$. We also will discuss *independent hypothesis tests*, a usage which follows from the probabilistic definition just provided for random variables. If two hypothesis tests are independent, then the outcome of one test does not affect the likelihood of the possible outcomes of the second.

Expectation is also used to define the *covariance* of two random variables. The covariance of the random variables X and Y is

$$Cov(X, Y) = E\big((X - E(X))(Y - E(Y))\big).$$

It is a useful fact that the covariance of independent random variables is 0.

Probability distributions are families of functions indexed by *parameters*. Parameters are features of a distribution that can be adjusted to improve the correspondence of a mathematical probability model and the real-world situation it describes. They are often denoted by Greek letters, and many are defined using expectation. For example, the mean, μ_Y, of a random variable Y is a parameter for the normal distribution (see Section 6.1) and is defined as the expected value of Y:

$$\mu_Y = E(Y). \tag{1.1}$$

Later sections of this chapter provide more discussion of this definition and more examples of parameters.

Problems in inferential statistics often require one to estimate parameters using information from a *sample data vector*, a finite set of values that the random variable Y takes during an experiment with n trials. The sample is often denoted by a bolded, lowercase Roman letter. This notation is used because a sample can be understood as an n-tuple or a vector in $\prod_{i=1}^{n} \mathcal{V}_Y \subset \mathbb{R}^n$. For example, using the definition of Y in the previous paragraph (the face-value of a die roll), the experiment in which we roll a die 4 times might yield the sample $\mathbf{y} = (1, 3, 5, 2)^T \in \mathbb{R}^4$. Each value y_i is a realized value of the random variable Y. The sample data vector is geometrically

understood as a vector in individual space.

Equivalently, samples of n observations can be understood as the values of n independent, identically distributed random variables Y_i. The sample $\mathbf{y}$ is the realized value of a vector of random variables, called a *random variable vector* and denoted with boldface, capital Roman letter: $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)^T$. Because the roll of a die does not change the likelihood of later rolls, we say the 4 consecutive rolls are independent. In out example, we might perform the same experiment just as easily by rolling 4 identical dice at the same time and recording the 4 resulting face-values. The second view of samples, as the realization of random variables, is more common and will be used from now on in this document. Note that the same notation (a boldface, capital Roman letter) is used for matrices in standard presentations of linear algebra and in this text. In the following chapters, the context will indicate whether such a symbol represents a vector of random variables or a matrix of constants.

Functions of sample data are called *statistics*, and one important class of statistics are estimates of parameters called *estimators*. They are conceptually distinct from the (possibly unknown) parameters and are usually denoted by roman letters. For example, the sample mean, an estimator for the parameter μ_Y, is often symbolized $\overline{y}$. Another common notation for estimators, including vectors of estimators, is the hat notation. In the next chapter, for example, we will use the symbol $\hat{\mathbf{y}}$ to represent the least-squares estimator for $\mathbf{y}$.

1.3 Centered vectors, sample mean, and sample variance

In the previous section, we used expectation to define the mean of a random variable (see equation (**??**) and claimed that samples can be used to estimate these parameters. In this section we define the statistics $\overline{y}$ and s^2, provide geometric descriptions of both, and show they are estimators for the mean (μ) and the variance (σ^2), respectively. We begin with the useful concept of an *centered vector*.

1.3.1 Centered vectors

The centered (data) vector $\mathbf{v}_c$ of a vector $\mathbf{v} = (v_1, v_2, \cdots, v_n)^T$ is defined to be the vector

$$\mathbf{v}_c = (v_{c1}, v_{c2}, \cdots, v_{cn})^T = \mathbf{v} - \left(\frac{\mathbf{1} \cdot \mathbf{v}}{n}\right)\mathbf{1}, \tag{1.2}$$

where $\mathbf{1}$ denotes the vector $(1, 1, \cdots, 1)^T \in \mathbb{R}^n$.

1.3.2 The sample mean

The *sample mean*, $\overline{y}$, gives an indication of the central tendency of the sample $\mathbf{y}$, and is defined as the average value obtained by summing the observations and dividing by n, the number of observations (see equation 1.3). It is usually denoted by placing a bar over the lowercase letter denoting the vector, but it is not bolded because it is a scalar. One can gain intuition about the sample mean by imagining the observations as uniform weights glued to a weightless ruler according to their value. In this sense, the mean can be understood as the center of mass of the sample distribution. In variable space, the sample mean can be represented as a point on the axis of the variable.

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1.3}$$

Since the sample mean is a scalar, it has no direct vector representation in individual space. However, the *mean vector* is a useful individual-space object and is written $\overline{y}\mathbf{1}$. The definition for a centered vector (see equation 1.2) can be written more parsimoniously as the difference of two vectors, $\mathbf{y}_c = \mathbf{y} - \overline{y}\mathbf{1}$. This relationship can also be illustrated geometrically with vector diagrams (Figure 1.4).

A few facts suggested by Figure 1.4 are worth demonstrating in generality. First, the mean vector $\overline{y}\mathbf{1}$ can be understood as (and obtained by) the orthogonal projection of $\mathbf{y}$ on the line in $\mathbb{R}^n$ spanned by the vector $\mathbf{1}$. We have

$$\text{Proj}_{\mathbf{1}}\mathbf{y} = \frac{\mathbf{y} \cdot \mathbf{1}}{\mathbf{1} \cdot \mathbf{1}}\mathbf{1} = \frac{\sum_{i=1}^{n} y_i}{n}\mathbf{1} = \overline{y}\mathbf{1}. \tag{1.4}$$

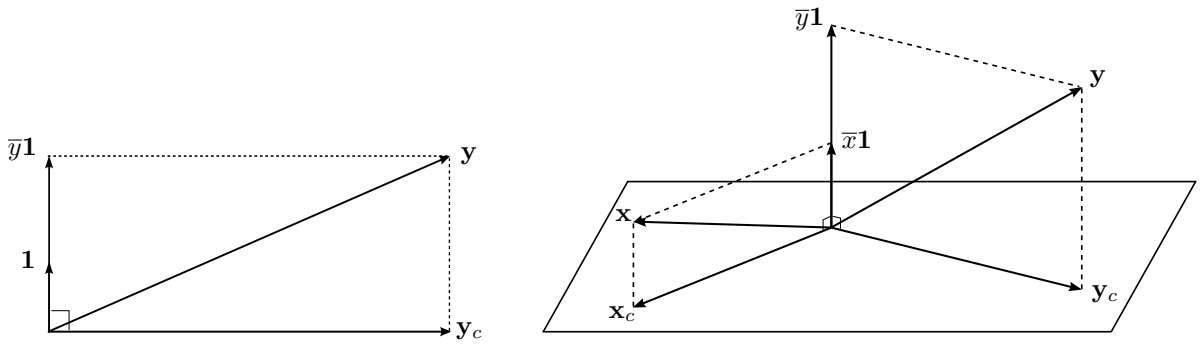


Figure 1.4: The centered vector $\mathbf{y}_c$ is the difference of the observation vector $\mathbf{y}$ and the mean vector $\overline{y}\mathbf{1}$, and the subspace spanned by centered vectors is orthogonal to the mean vectors.

It follows that a centered vector and the corresponding mean vector are orthogonal.

In the last section, we called estimators those statistics that can be used to estimate parameters. A special class of estimators are called *unbiased* because the expected value of the estimator is equal to the parameter. For example, we can write $E(\overline{Y}) = \mu_Y$ if $\overline{Y}$ is an unbiased estimator of μ_Y. To make sense of the statement $E(\overline{Y})$ we must think of $\overline{Y}$ as a linear combination of the n random variables associated with the sample $\mathbf{y}$, and consider $\overline{y}$ to be a realization of this random variable. We use the following definition of $\overline{Y}$ *as a random variable*:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \tag{1.5}$$

where Y_i is the random variable realized by the i^{th} observation in the sample $\mathbf{y}$. The definition of $\overline{Y}$ as a random variable demonstrates that random variables can be constructed as linear combinations of other random variables. We will see in later sections that, given distributional information about the component random variables Y_i, we can estimate the mean and variance of the composite random variables such as $\overline{Y}$.

It remains to show that $\overline{Y}$ is an unbiased estimator of μ, and the computation is straight-

forward:

$$\begin{aligned} E(\overline{Y}) &= E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(Y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \end{aligned}$$

Notice that this proof relies on the assumption that the random variables Y_i are identically distributed (in particular, they must all have the same mean, μ).

1.3.3 The variance and sample variance

The *variance* of a random variable Y, σ_Y^2, indicates its variability and helps answer the question: How much do the observed values of the variable vary from one to another? Returning to the physical model for the values in a distribution imagined as uniform weights glued to a massless ruler according to their value, the variance can be understood as the rotational inertia of the variable's distribution about the mean. The greater the variance, the greater the force would be required to change the rotation rate of the distribution by some fixed amount. A random variable with low variance has realized values that cluster tightly around the mean of the variable.

The variance is defined to be the expected value of the squared deviation of a variable from the mean:

$$\sigma_Y^2 = \mathrm{Var}(Y) = E\left((Y-\mu)^2\right). \tag{1.6}$$

If the variable's values are known for the entire population (of size n), then $\overline{y} = \mu$ and the variance can be computed as the mean squared deviation from the mean:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2. \tag{1.7}$$

In most analyses, however, only a sample of observations are available, and the formula (1.7) systematically underestimates the true population variance: $\frac{1}{n}\sum(y_i-\mu)^2 < \sigma^2$. This phenomenon is more noticeable with small samples.

An unbiased estimator, s^2, of the variance is obtained using $n-1$ instead of n in the denominator:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n} y_{ci}^2 = \frac{\|\mathbf{y}_c\|^2}{n-1}. \tag{1.8}$$

The proof that $E(s_n^2) \neq \sigma^2 = E(s^2)$ relies on the independence of the observations in the sample, one of the characteristics of a random sample and a common assumption made of empirical samples by researchers who perform statistical analyses. It is important to notice the similarity between equation 1.8 and the numerator and denominator of the F-ratio (see equation 2.15). In fact, we shall see that the estimate of the sample variance s^2 can be understood geometrically as the per-dimension length of the centered vector $\mathbf{y}_c$. This will be more fully explained in Section 2.1.3, but the key idea is that $\mathbf{y}_c$ lives in the orthogonal complement of $\mathbf{1}$, and this space is $(n-1)$-dimensional. The information in one dimension of the observation vector can be used to center the vector (and estimate the mean μ), and the information in each of the remaining $n-1$ dimensions provides an estimate of the variance σ^2. The best estimate is therefore the mean of these $n-1$ values which conveniently sum to $\|\mathbf{y}_c\|^2$.

In a manner analogous to $\overline{y}$, the sample variance s^2 can also be understood as the realization of the random variable S^2, which is defined as a linear combination of other random variables:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2. \tag{1.9}$$

In Chapter 2, we prove that S^2 is an unbiased estimator after a prerequisite result about the variance of random variables formed from linear combinations of random variables. Next, we briefly consider another statistic related to the sample variance that is frequently used in statistical analyses.

1.3.4 The standard deviation and Chebyshev's inequality

Sample variance is denoted s^2 because it is equal to the square of the *sample standard deviation*, s. Both the variance and the standard deviation address the variability of a random variable. However, the standard deviation is more easily interpreted than variance because the units for variance are the squared units of the variable. By taking the square root, the variance is

transformed to the metric of the variable and can be more easily compared with values in a data set. The standard deviation is also useful for estimating the probability that the variable will take a value in a particular interval of the domain.

The most general result of this kind is Chebyshev's inequality, which states that

$$P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}, \tag{1.10}$$

regardless of the distribution of Y. For sufficiently large samples, $\overline{Y}$ can be used to estimate μ and s can be used to estimate σ. For example, suppose that for some sample of 30 observations, $\overline{Y} = 1$ and $s = 2$. Then the probability that the next observation of Y deviates from the mean by more than $4 = 2s$ is at most $\frac{1}{2^2} = \frac{1}{4}$ or 25%.

The proof of this inequality follows easily from a more general statement called Markov's inequality: If $\tilde{Y} \geq 0$, then for any value of $\tilde{Y}$, denoted a, we have:

$$P(\tilde{Y} \geq a) \leq \frac{E(\tilde{Y})}{a} \tag{1.11}$$

The proof of Markov's inequality in the continuous case follows:

$$\begin{aligned} E(Y) &= \int_{\mathcal{V}_Y} y f_Y(y) dy \\ &= \int_{Y<a} y f_Y(y) dy + \int_{Y \geq a} y f_Y(y) dy \\ &\geq \int_{Y \geq a} y f_Y(y) dy \\ &\geq a \int_{Y \geq a} f_Y(y) dy \\ &= a\ P(Y \geq a) \end{aligned}$$

Chebyshev's inequality follows from Markov's inequality by replacing $\tilde{Y}$ with $(Y - \mu)^2 \geq 0$

and taking a to be $k^2\sigma^2$. Since $|Y - \mu| \geq k\sigma$ if and only if $\tilde{Y} \geq k^2\sigma^2$, we have

$$\begin{aligned} P(|Y - \mu| \geq k\sigma)) &= P(\tilde{Y} \geq a) \\ &\leq \frac{E(\tilde{Y})}{a} \\ &= \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2} \end{aligned}$$

Chebyshev's inequality illustrates the utility of standard deviations as a tool for describing the distribution of variables in variable space ($\mathbb{R}^m$). The standard deviation can be understood as the length of an interval on the axis of a variable that follows a normal distribution (see Section 6.1) and used to help represent the distribution (Figure 1.5). Nevertheless, one limitation of variable space representations of data is that standard deviations are hard to see in a scatterplot without calculating s and appropriately labeling the variable axes.

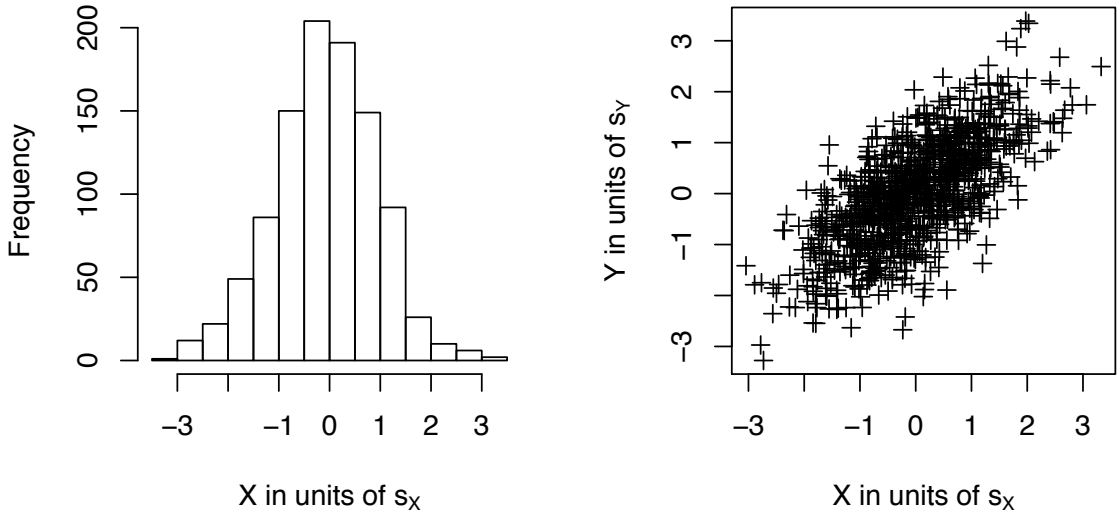


Figure 1.5: In variable space, the standard deviation s is a convenient, distribution-related unit for many variables; in this figure, the origin of each axis is shifted to the mean of the associated variable.

Standard deviations also have a useful interpretation in vector drawings of individual space ($\mathbb{R}^n$). Unlike scatterplots in variable space, vector drawings in individual space already represent the standard deviation by nature of their construction. From equation (1.8), it is clear that the variance of a variable can be expressed using the dot product of the centered variable with itself. It follows from the definition of the Euclidean norm ($\|v\| = \sqrt{v \cdot v}$) that the standard deviation

of a variable is proportional to the length of the centered vector.

$$s^2 = \frac{\mathbf{y}_c \cdot \mathbf{y}_c}{n-1} \iff s = \|\mathbf{y}_c\| \frac{1}{\sqrt{n-1}}$$

Since the constant of proportionality $\frac{1}{\sqrt{n-1}}$ depends only on the dimension of individual space, all centered variable vectors are scaled proportionally. Assuming comparable units, this means the ratio of the length of two centered vectors in individual space is equal to the ratio of the standard deviations of these variables. For example, in the left panel of Figure 1.3, the vector $\mathbf{y}$ has a smaller standard deviation than the vector $\mathbf{x}$ because it is shorter. Histograms of these two variables would show the y_i bunched more tightly around their mean.

1.4 An illustrative example

We conclude this chapter with an example to introduce statistical models and hypothesis testing.[1] The geometry of individual space is essential here; variable space does not afford a simple or compelling treatment of these ideas and their standard algebraic treatment at the elementary level masks rather than illuminates meaning. In fact, the presentation here is closer to the original formulation of the ideas by R.A. Fisher in the early twentieth century (Herr, 1980).

Suppose a tutor would like to discover if her tutoring improves her students' scores on a standardized test. She picks two of her students at random and for each calculates the difference in test score before and after a one month period of tutoring: $g_1 = 7$, $g_2 = 9$. Then she plots these gain-scores in individual space as the vector $\mathbf{y} = (g_1, g_2)$. (See Figure 1.6a)

To proceed, we must make a few reasonable assumptions about the situation. First, we assume that the students' scores are independent random variables. Among other things, this means that the gain-score of either student does not affect the gain-score of the other. Next, we assume that the gain-score for both students follows the same distribution. In particular, we want to assume that if we could somehow go back in time and repeat the month of tutoring, then the average gain-score over many repetitions would be the same for each student. Neither student is predisposed to a greater benefit from tutoring than the other. This assumption lets

[1] This example is inspired by something similar in Saville & Wood (1991). I have changed the context and values, used my own figures, and considerably developed the discussion.

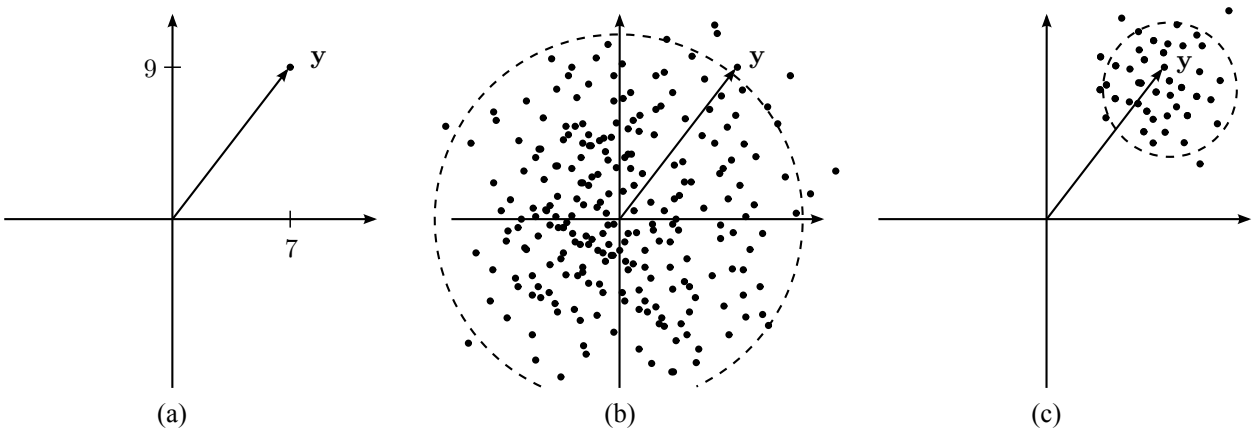


Figure 1.6: (a) The vector **y** is plotted in individual space; one must decide if (b) the vector **y** is more likely a sample from a distribution centered at the origin of individual space or instead (c) a sample from of a distribution centered away from the origin on the line spanned by (1,1).

us postulate a *true* mean gain-score μ for the population of students, and our goal is to use the data we have to estimate it. Finally, we make the assumption that the common distribution of gain-scores is normal with a mean of 0. This implies that the signed length of the vector follows a normal distribution with a mean of 0 (where the sign of the length is given by the sign of the mean gain-score) and standard deviation $\frac{s}{\sqrt{2-1}} = s$, where s is the standard deviation of gain-scores. Moreover, all directions are equally likely because of the assumed independence.

The tutor's question has two possible answers: The tutoring makes little difference in students' gain-scores or there is some effect. In the first case, we would expect many repetitions of her procedure to look like Figure 1.6b, and in the second case, many repetitions might look like Figure 1.6c, with the center of the distribution displaced from the origin. In both figures, the standard deviation of the length of the vector is indicated by a dashed circle.

1.4.1 The geometry of hypotheses

We call the first possibility the null hypothesis and write $H_0 : \mu = 0$, where μ is the mean gain-score resulting from tutoring. The second possibility is called the alternative hypothesis $H_1 : \mu \neq 0$. Certainly after plotting 100 repetitions of the tutor's procedure it would likely be easy to decide which hypothesis was most plausible; the challenge is to pick the most likely

hypotheses based only on a single trial.

The center of the distribution for the vector $\mathbf{y}$ must lie along the line spanned by the vector $\mathbf{1} = (1, 1)$. Geometrically, we can understand the vector $\mathbf{y}$ as the sum of the vector $\overline{y}\mathbf{1}$ and a second vector orthogonal to the first which can be written $\mathbf{e} = \mathbf{y} - \overline{y}\mathbf{1}$, where $\overline{y}$ is the length of the orthogonal projection of $\mathbf{y}$ on $\mathbf{1}$. This relationship is shown in Figure 1.7.

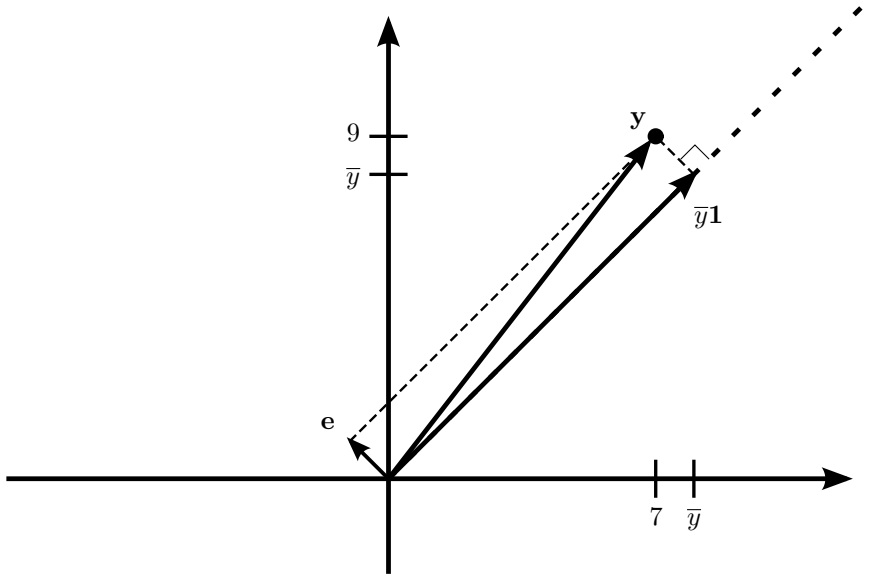


Figure 1.7: The vector $\mathbf{y}$ can be understood as the sum of $\overline{y}\mathbf{1}$ and a vector $\mathbf{e}$ that is orthogonal to $\mathbf{1}$.

The idea for testing the null hypothesis is to compare the lengths of $\overline{y}\mathbf{1}$ and $\mathbf{e}$ using the ratio

$$t = \frac{\text{sgn}\,(\overline{y})\|\overline{y}\mathbf{1}\|}{\|\mathbf{e}\|}.$$

To make sense of the hypothesis test geometrically, consider both parts of Figure 1.8. In both, the shaded region indicates the cone in individual space where the t-ratio is large. In Figure 1.8a, the vector $\mathbf{y}$ gives an estimate of the variance of the distribution of gain-scores under the null hypothesis, and the corresponding standard deviation is indicated by the dashed circle centered at the origin of individual space. In Figure 1.8b, it is instead the vector $\mathbf{e}$ that gives an estimate of the variance of the distribution of gain-scores relative to $\overline{y}\mathbf{1}$, and the corresponding standard deviation of this distribution is indicated by the radius of the dashed circle centered at $(\overline{y},\ \overline{y})$.

If the t-ratio is large, then the vector $\mathbf{y}$ is 'close' in some sense to the line spanned by $\mathbf{1}$.

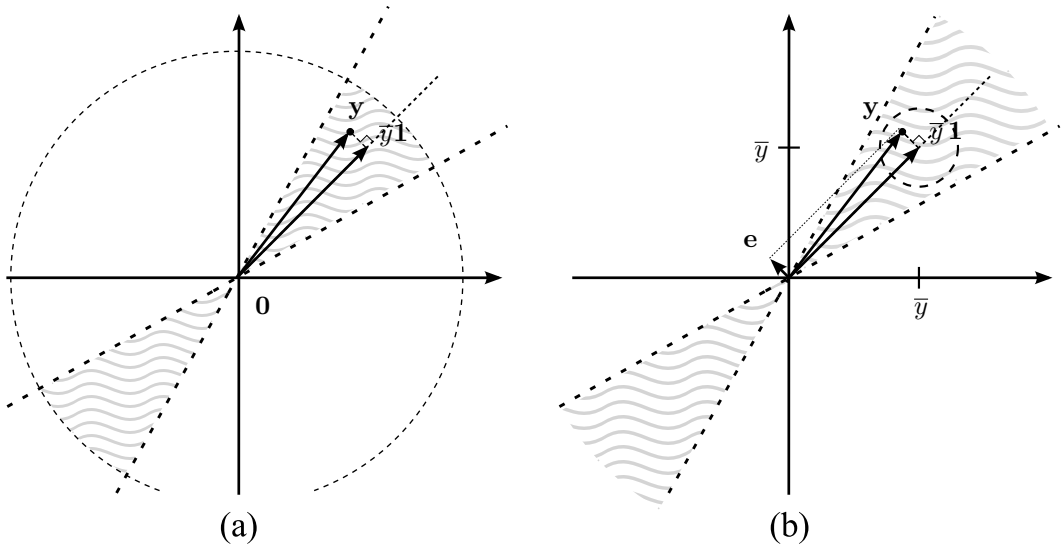


Figure 1.8: The distribution of the random variable $\mathbf{Y}$ has different centers and relies on a different estimates for the standard deviation under (a) the null hypothesis and (b) the alternative hypothesis.

In this case, we can see geometrically why the null hypothesis is unlikely. If $\mathbf{y}$ usually lands anywhere within the dashed circle in Figure 1.8a, then it is rare $\mathbf{y}$ will land in the shaded cone. The observation vector $\mathbf{y}$ is unusual under the null hypothesis and thus we can then reject the null hypothesis in favor of the more plausible alternative hypothesis. Notice that under the alternative hypothesis, the t-ratio will usually be large. Geometrically, we can see that the dashed circle in Figure 1.8b is entirely within the shaded cone. On the other hand, whenever the t-ratio is small we have no evidence with which we might reject the null hypothesis.

It can be shown that this ratio follows the t-distribution (see Section 6.4). Integrating the probability distribution function for the t-distribution between the values of -8 and 8 gives the quantity 0.921. This suggests that under the assumption of the null hypothesis, we can expect a t-ratio with an absolute value as high or higher than the one we observed only 8% of the time. Equivalently, we can expect a sample vector in individual space as close or closer to the line spanned by $\mathbf{1}$ only 8% of the time under the assumption of the null hypothesis.

1.4.2 The F-ratio and the t-test

Like the t-ratio, the F-ratio is used for testing hypotheses and follows the F-distribution (see Section 6.4). This example illustrates the relationship between the t-ratio and the F-ratio. The F-ratio is more versatile and will be used almost exclusively from here on for hypothesis testing. We begin by introducing the linear algebraic notation for describing the vector relationships depicted in Figure 1.7. This equation is called the *model* for the experiment. Models will be defined and elaborated in much more generality in the next chapter. Under this model, the observation vector $\mathbf{y}$ is assumed to be equal to the sum of the error vector $\mathbf{e}$ and the matrix product of mean vector $\overline{y}\mathbf{1}$ considered as a 2×1 matrix and the constant $\overline{y}$ considered as a 1×1 matrix.

$$\begin{bmatrix} 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 8 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

This model explains each sample as the sum of a constant mean gain-score (the product $\mathbf{Xb}$) plus a small amount of random error (the vector $\mathbf{e}$) that is different for each sample. Another way to understand the hypothesis test is to reframe our goal as the search for the estimate vector $\mathbf{b}$ of the true vector $\boldsymbol{\beta}$. The later describes the true relationship between $\mathbf{Y}$ and $\mathbf{X}$. The null hypothesis states that the vector we are trying to estimate $\boldsymbol{\beta}$ is zero, whereas the alternative hypothesis states it is non-zero. It is noteworthy that throughout this thesis, the alternative hypothesis is a negation of the null hypothesis instead of a hypothesis that specifies a particular value of $\boldsymbol{\beta}$.

The F-ratio is a comparison of the per-dimension squared lengths of (1) the projection $\overline{y}\mathbf{1}$ of the observation vector $\mathbf{y}$ onto the model vector $\mathbf{1}$ and (2) the error vector $\mathbf{e} = \mathbf{y} - \overline{y}\mathbf{1}$. In this case, the F-ratio is simply the square of the t-ratio (see Section 6.4) and it follows the

F-distribution (see Section 6.3). The F-ratio can be written

$$F = \frac{\|\overline{y}\mathbf{1}\|^2/1}{\|\mathbf{e}\|^2/1}.$$

(We mean by *per-dimension* a factor that is the inverse of the number of dimensions of the subspace containing the vector, after individual space has been partitioned into the model subspace and error subspace. More on this later; see Section 2.1.3.)

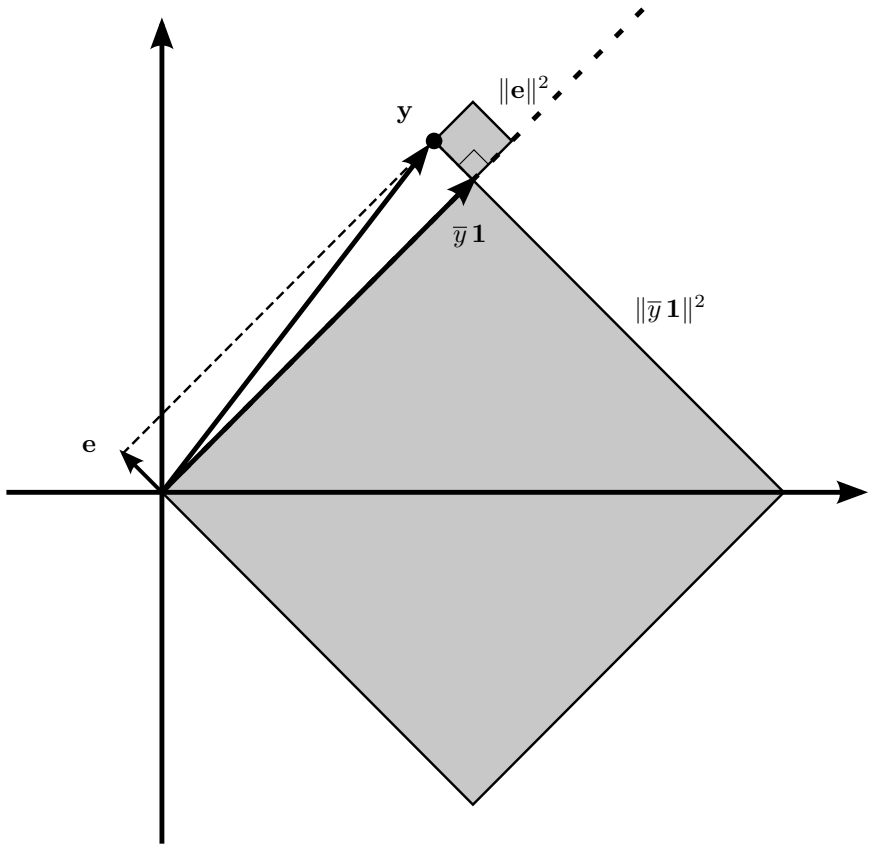


Figure 1.9: The t-ratio of $\|\overline{y}\,\mathbf{1}\|$ to $\|\mathbf{e}\|$ under the t-distribution provides the same probability information about the likelihood of the observation under the null hypothesis as the F-ratio of $\|\overline{y}\,\mathbf{1}\|^2$ to $\|\mathbf{e}\|^2$ under the F-distribution.

The values $\|\overline{y}\mathbf{1}\|^2$ and $\|\mathbf{e}\|^2$ are illustrated in Figure 1.9. Again it is quite evident from the geometry that this ratio is much greater than one. The F-distribution tells us how unusual such an observation would be, agreeing with our prior result. Under the null hypothesis, we would expect a sample with an F statistic this large or larger only 8% of the time.

Chapter 2

The General Linear Model

2.1 Statistical models and the general linear model

A statistical model is analogous to a mathematical function. It expresses the dependent variable $\mathbf{Y}$ as a function of the independent variables $\mathbf{x}_i$. One difference between functions and models is that models incorporate random error. For a fixed set of input values, the model does not always return the same value because each output contains a random component.

A model is called the *true model* when it is assumed to express the real relationship between the variables. However, if the data collected is restricted from the population to a sample, the true model can never be discovered with certainty. Only approximations of the true model are possible. To make the problem of finding an approximation tractable, a family of models is defined using a small number of parameters, θ_1, θ_2, $\cdots$, θ_k (see equation 2.1). The true model can then be expressed with fixed (but unknown) parameter values, and the sample data can be used to estimate them.

$$\mathbf{Y} = f_{\theta_1,\theta_2,\cdots,\theta_k}(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) \tag{2.1}$$

We define the *design matrix* $\mathbf{X}$ of a linear model to be the matrix $[\mathbf{1}\ \mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_p]$. With this notation and a general framework for statistical models in mind, we are prepared to state the general linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \tag{2.2}$$

where $\mathbf{Y},\ E \in \mathbb{R}^n$ and $\mathbf{X}$ is an $n \times (p+1)$ matrix over $\mathbb{R}$. In the general linear model, the vector of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ is analogous to the parameters $\theta_1,\ \theta_2,\ \cdots,\ \theta_k$ used in equation (2.1). The vector $\mathbf{E} = (E_1, E_2, \cdots, E_n)^T \in \mathbb{R}^n$ is the random error component of the model and is the difference between the observed values and those predicted by the product $\mathbf{XB}$.

The example model from the first chapter (see equation 1.12) is a simple case of the general linear model. The design matrix for this model is simply the vector $\mathbf{X} = \mathbf{1}$, $\mathbf{Xb}$ is the vector $\overline{y}\mathbf{1}$, and the vector $\mathbf{e}$ is the centered observation vector $\mathbf{y}_c$. In general, the matrix product $\mathbf{Xb}$ is the projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ (which in this case is the line spanned by $\mathbf{1}$). Recall that the use of a lower-case, bold $\mathbf{y}$ and $\mathbf{e}$ indicates vectors of observed values from a particular sample. The capital, bold $\mathbf{Y}$ and $\mathbf{E}$ in equation (2.2) indicate the corresponding random variable. Just as Greek characters refer to individual population parameters and Roman characters refer to corresponding sample statistics (e.g., μ and $\overline{x}$), we use $\boldsymbol{\beta}$ for the vector of model parameters but $\mathbf{b}$ for the vector of corresponding sample statistics.

An equivalent statement of the general linear model is a system of equations for each dependent random variable Y_i. Thus, for all i, we have:

$$Y_i = \mathbf{X}_i\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + E_i, \tag{2.3}$$

where $\mathbf{X}_i$ is the i^{th} row of the design matrix $\mathbf{X}$. As demonstrated in later chapters, the general linear model subsumes many of the most popular statistical techniques including analysis of variance (ANOVA) models in which the design matrix columns $\mathbf{x}_i$ are categorical variables (for example, gender or teacher certification status), simple regression and correlation models comparing two continuous variables, and multiple regression models in which there are two or more continuous, independent variables (e.g., teacher salary and years of experience).

The formulation of the general linear model presented here treats the observed values of the independent variables as fixed constants instead of as realized random variables. In experimental data sets, this is often appropriate because researchers control the values of the independent variables (e.g., dosage in drug trials). In observational data sets, it often makes more sense

to treat the observations of independent variables as realized random variables, because these observations are determined by choice of the sample rather than experimental design. Under certain conditions, the general linear model can be used when both independent and dependent variables are treated as random variables. For example, the requirement that all random variables have a multivariate normal distribution is sufficient but not strictly necessary. Even though the results hold in greater generality, all independent variables are treated as fixed in order to simplify the presentation.

2.1.1 Fitting the general linear model

Given a data set and a model, the next step of statistical analysis is to use the sample data to estimate the parameters of the model. This process is called fitting the model. Our goal is that the vector of parameter estimates, $\mathbf{b} = (b_0, b_1, \cdots, b_p)^T$, be as close as possible to $\boldsymbol{\beta}$, the vector of putative true parameter values. Such a comparison is unfortunately impossible because $\boldsymbol{\beta}$ is unknown. However, since the model separates realized values of each random variable Y_i into a systematic component $\mathbf{X}_i\boldsymbol{\beta}$ (where $\mathbf{X}_i$ is the i^{th} row of the matrix $\mathbf{X}$) and a random component E_i, a feasible goal is to find a vector $\mathbf{b}$ so that $\mathbf{X}_i\mathbf{b}$ is as close as possible to Y_i, for each i, $1 \leq i \leq n$. To accomplish this, we need a number that can summarize the model deviation over all the observed values in the sample. A natural choice for this number is the length of the error vector $\mathbf{E}$, because the Euclidean norm depends on the value of each coordinate. Therefore, it suffices to find a vector $\mathbf{b}$ which minimizes the expected length of $\mathbf{E}$.

With a sample of observations $\mathbf{y}$ in hand, the best estimate available for $\mathbf{Y}$ is simply the sample $\mathbf{y}$. Restating the goal identified above in terms of the sample, we say we are looking for a vector $\mathbf{b}$ that will minimize the difference between $\mathbf{Xb}$ and $\mathbf{y}$. This difference is the best available estimate for $\mathbf{E}$ and is denoted $\mathbf{e}$. It follows that the fitted model can be written

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}, \tag{2.4}$$

and similarly for each observed y_i we can write

$$y_i = \mathbf{X}_i\mathbf{b} + e_i.$$

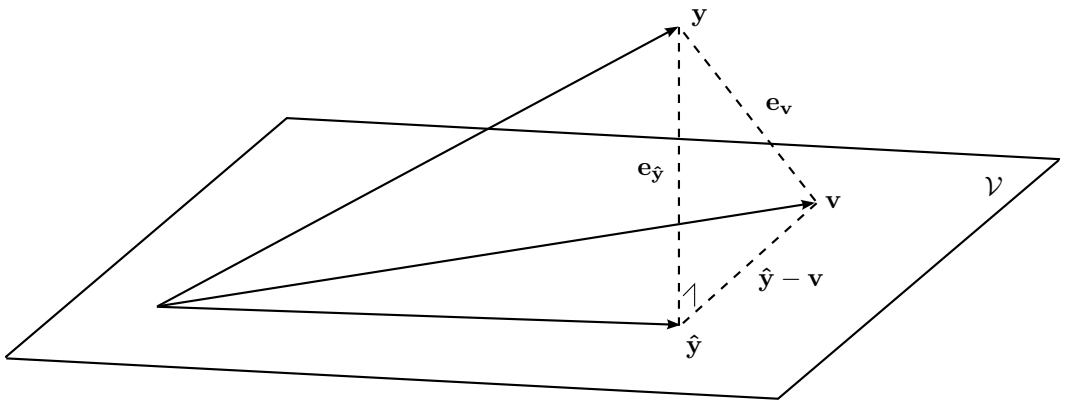


Figure 2.1: The vector $\hat{\mathbf{y}}$, the projection of $\mathbf{y}$ onto $\mathcal{V} = C(\mathbf{X})$, is seen to be the unique vector in $\mathcal{V}$ that is closest to $\mathbf{y}$.

The strategy for finding $\mathbf{b}$ can be understood geometrically by imagining $\mathbf{y}$ in n-dimensional Euclidean space. The vector $\boldsymbol{\beta}$ is assumed to lie in the $(p+1)$-dimensional subspace spanned by $\mathbf{1}$ and the vectors $\mathbf{x}_i$. We are looking for a vector $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ in the column space of $\mathbf{X}$, $C(\mathbf{X})$, such that the length of $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is minimized. (The hat notation indicates an estimate; $\hat{\mathbf{y}}$ is an estimate for the observed vector of y-values $\mathbf{y}$.) The situation is represented in Figure 2.1 and the proof of Lemma 2.1.1 follows easily.

Lemma 2.1.1. *Let $\mathcal{V} \subset \mathbb{R}^n$ be a subspace and suppose $\mathbf{y} \in \mathbb{R}^n$. For each vector $\mathbf{v} \in \mathcal{V}$, let $\mathbf{e}_\mathbf{v} = \mathbf{y} - \mathbf{v}$. Then there is a unique vector $\hat{\mathbf{y}} \in \mathcal{V}$ such that $0 \leq \|\mathbf{e}_{\hat{\mathbf{y}}}\| < \|\mathbf{e}_\mathbf{v}\|$ for all $\mathbf{v} \neq \hat{\mathbf{y}}$.*

Proof. We can write $\mathbb{R}^n = \mathcal{V} \oplus \mathcal{V}^\perp$, and claim that $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto $\mathcal{V}$, the unique vector in $\mathcal{V}$ such that $\mathbf{e}_{\hat{\mathbf{y}}} \in \mathcal{V}^\perp$. Let $\mathbf{v} \neq \hat{\mathbf{y}} \in \mathcal{V}$. By the Pythagorean Theorem, $\|\mathbf{e}_\mathbf{v}\|^2 = \|\mathbf{e}_{\hat{\mathbf{y}}}\|^2 + \|\hat{\mathbf{y}} - \mathbf{v}\|^2$, which implies $0 \leq \|\mathbf{e}_{\hat{\mathbf{y}}}\| < \|\mathbf{e}_\mathbf{v}\|$. □

The projection of $\mathbf{y}$ onto $C(\mathbf{X})$, the vector $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$, gives the desired estimate for $\mathbf{y}$. It remains to state and prove the general method for obtaining $\mathbf{b}$ using the sample data vector $\mathbf{y}$ and the matrix $\mathbf{X}$. The trivial case in which $\mathbf{y} \in C(\mathbf{X})$ does not require estimation because the model (equation 2.4) can then be solved directly for $\mathbf{b}$—it is not addressed here.

Theorem 2.1.2. *Let* $\mathbf{X}$ *be an* $n \times k$ *matrix with linearly independent columns, let* $\mathbf{y} \in \mathbb{R}^n$ *be in the complement of* $C(\mathbf{X})$*, and let* $\mathbf{b}$ *be a vector in* $\mathbb{R}^k$*. Suppose that* $\mathbf{Xb} = \hat{\mathbf{y}}$ *and that* $\mathbf{y} - \hat{\mathbf{y}} \in C(\mathbf{X})^{\perp}$*. We claim that* $\mathbf{b} = (\mathbf{X}^T X)^{-1}\mathbf{X}^T\mathbf{y}$.

Proof. $C(\mathbf{X})^{\perp}$ is the null-space of the matrix $\mathbf{X}^T$, and so

$$\mathbf{X}^T(\mathbf{y} - \mathbf{Xb}) = \mathbf{0}.$$

Consequently,

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{y}.$$

The result follows as long as $\mathbf{X}^T\mathbf{X}$ is nonsingular, for in this case

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

To see that $\mathbf{X}^T\mathbf{X}$ is nonsingular, let $\mathbf{v} \in \mathbb{R}^k$ and suppose $\mathbf{X}^T\mathbf{X}(\mathbf{v}) = \mathbf{0}$. It follows that $\mathbf{X}^T(\mathbf{Xv}) \cdot \mathbf{v} = \mathbf{0}$, and therefore that $\|\mathbf{Xv}\|^2 = \mathbf{0}$. We conclude that $\mathbf{v} = \mathbf{0}$ since $\mathbf{Xv} = 0$ and $\mathbf{X}$ has rank k. □

This method produces a vector $\mathbf{b}$ of parameter estimates, called the *least squares estimate* because it minimizes the sum of the squared components of $\mathbf{e}$. Another method of finding a formula for $\mathbf{b}$ is to use calculus to find critical points of the function $S(\mathbf{b}) = \|\mathbf{e}\|^2$. The resulting formula can be used obtain the *normal equations*

$$(\mathbf{X}^T\mathbf{X})\mathbf{b} = \mathbf{X}^T\mathbf{y},$$

which is also an intermediate step in the proof of Theorem 2.1.2.

2.1.2 Centering independent variable vectors

The least squares solution $\mathbf{b}$ can be obtained from the orthogonal projection of the observed dependent variable vector $\mathbf{y}$ onto the column space of the design matrix $\mathbf{X} = [\mathbf{1}\ \mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_p]$. In fact, the same solution can be obtained from the projection of $\mathbf{y}$ onto the subspace of $\mathbb{R}^n$

spanned by the vector $\mathbf{1}$ and the centered vectors $\mathbf{x}_{ci} = \mathbf{x}_i - \overline{x}_i \cdot \mathbf{1}$. Let $\mathbf{X}'$ denote the block matrix $[\mathbf{1}\ \mathbf{X}_c]$ with $\mathbf{X}_c = [\mathbf{x}_{c1}\ \mathbf{x}_{c2}\ \cdots\ \mathbf{x}_{cp}]$. Centering a vector entails subtracting the projection of that vector onto the vector $\mathbf{1}$ and because $\mathbf{1}$ is in both $C(\mathbf{X})$ and $C(\mathbf{X}')$, these subspaces are equal.

From Theorem 2.1.2, we know that the least squares solution for $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$ is

$$\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T y},$$

and that the least squares solution for $\mathbf{y} = \mathbf{X'b} + \mathbf{e}$ is

$$\mathbf{b'} = ((\mathbf{X'})^{\mathbf{T}}\mathbf{X'})^{-1}(\mathbf{X'})^{\mathbf{T}}\mathbf{y}.$$

The following theorem relates these solutions.

Theorem 2.1.3. *Whenever* $\mathbf{b} = (b_0, \ldots, b_p)^T$ *is the least squares solution for the linear model of* $\mathbf{y}$ *with the design matrix* $\mathbf{X}$ *and* $\mathbf{b'} = (b'_0, \ldots, b'_p)^T$ *is the least squares solution for the linear model of* $\mathbf{y}$ *with the corresponding centered design matrix* $\mathbf{X}_c$*, then* $b_i = b'_i$ *for all* $1 \leq i \leq p$*. Moreover, the parameter estimate* b_0 *can also be obtained from* $\mathbf{b'}$*; in particular,* $b_0 = b'_0 - \sum_{i=1}^{p} b'_i \overline{x}_i$*.*

Proof. Since $C(\mathbf{X}) = C(\mathbf{X}')$ we know that $\mathbf{Xb} = \mathbf{X'b'}$. The result follows immediately from the equation

$$\mathbf{X}\begin{bmatrix} 1 & -\overline{x}_1 & \cdots & -\overline{x}_p \\ & & & \\ \mathbf{0} & & I_p & \\ & & & \end{bmatrix} = \mathbf{X}',$$

where I_p is the $p \times p$ identity matrix. □

The geometry of the result is instructive and readily apparent in the case where there is only one independent variable, that is, when $\mathbf{X} = [\mathbf{1}\ \mathbf{x}]$. It follows from Theorem 2.1.2 that $\hat{\mathbf{y}} = b_0\mathbf{1} + b_1\mathbf{x}$. Let $\mathbf{X}' = [\mathbf{1}\ \mathbf{x}_c]$. Now the vectors $\mathbf{x}$, $\mathbf{x}_c$, and $\overline{x}\mathbf{1}$ form a right triangle in $C(\mathbf{X}) = C(\mathbf{X}_c)$ that is similar to the triangle formed by $b_1\mathbf{x}$, $b'_1\mathbf{x}_c$, and $\mathbf{z} = b_1\mathbf{x} - b'_1\mathbf{x}_c = z\mathbf{1}$ (see Figure 2.2). It follows that $b_1 = b'_1$ and that $b_0 = \overline{y} - z$. Certainly $b'_0 = \overline{y}$, and by similarity we

have that $z = b'_1\overline{x}$. Thus, $b_0 = b'_0 - b'_1\overline{x}$ and we can therefore write $\hat{\mathbf{y}} = (b'_0 - b'_1\overline{x})\mathbf{1} + b'_1\mathbf{x}_c$, which is in terms of $\mathbf{b}'$ as desired.

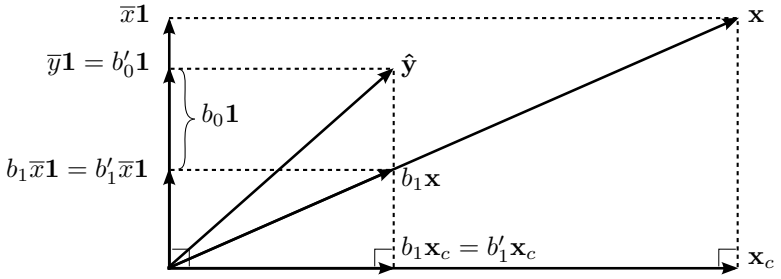


Figure 2.2: The vector $\hat{\mathbf{y}}$, the projection of the vector $\mathbf{y}$ into $C([\mathbf{1}\ \mathbf{x}])$, is equal to $b_0\mathbf{1} + b_1\mathbf{x}$ and also is equal to $b'_0\mathbf{1} + b'_1\mathbf{x}_c$.

2.1.3 Subspaces of individual space

It is often convenient to think of individual space ($\mathbb{R}^n$) partitioned into several subspaces that correspond to different parts of the assumed linear model. One important subspace in individual space is spanned by the vector $\mathbf{1}$. We denote this line $\mathcal{V}_{\mathbf{1}}$.

As we have already seen, the least squares solution $\mathbf{b}$ is derived by projecting the observation vector $\mathbf{y}$ into the column space of $\mathbf{X}$, $C(\mathbf{X})$. For analogous notation, we also use $\mathcal{V}_{\mathbf{X}}$ to denote this subspace of $\mathbb{R}^n$. Certainly $\mathcal{V}_{\mathbf{1}} \subset \mathcal{V}_{\mathbf{X}}$ and moreover, we can write $\mathcal{V}_{\mathbf{X}} = \mathcal{V}_{\mathbf{1}} \oplus \mathcal{V}_{\mathbf{X}_c}$, where $\mathcal{V}_{\mathbf{X}_c}$ denotes the orthogonal complement of $\mathcal{V}_{\mathbf{1}}$ in $\mathcal{V}_{\mathbf{X}}$. Finally, we let $\mathcal{V}_{\mathbf{e}}$ denote the orthogonal complement of $\mathcal{V}_{\mathbf{X}}$ in $\mathbb{R}^n$. Putting these statements together gives the equation

$$\mathbb{R}^n = \mathcal{V}_{\mathbf{1}} \oplus \mathcal{V}_{\mathbf{X}_c} \oplus \mathcal{V}_{\mathbf{e}}, \text{ (all orthogonal).} \tag{2.5}$$

From linear algebra, we know there is corresponding equation relating the dimensions of these subspaces to the dimension of individual space:

$$n = 1 + p + (n - p - 1). \tag{2.6}$$

The vectors that make up linear models (e.g., $\mathbf{y}$, $\hat{\mathbf{y}}$, $\mathbf{e}$, etc.) are each contained in precisely one of these subspaces. The *(ambient) dimension* of each vector is defined to be the dimension of its associated (smallest) ambient space. Of course, each vector is one-dimensional in the

traditional sense. A basic technique for estimating models and testing hypotheses is finding a convenient basis for individual space based on the vectors that make up the model. This new definition for *dimension* is directly related to this implicit basis imposed on $\mathbb{R}^n$ by the linear model. The vector $\mathbf{1}$ is 1-dimensional and is also (almost always) taken to be the first of the implicit basis vectors. Next we choose a set of p orthogonal basis vectors for $\mathcal{V}_{\mathbf{X}_c}$. If the vectors of centered predictors $\{\mathbf{x}_{ci} : 1 \leq i \leq p\}$ are all orthogonal, so much the better. The vector $\hat{\mathbf{y}}_c$ is contained in $\mathcal{V}_{\mathbf{X}_c}$ and therefore has (ambient) dimension p. Finally, we can pick any set of $n - p - 1$ vectors that span $\mathcal{V}_{\mathbf{e}}$ to complete the basis for $\mathbb{R}^n$. The vector $\mathbf{e}$, naturally, has the ambient space $\mathcal{V}_{\mathbf{e}}$ and therefore has dimension $n - p - 1$. The observation vector $\mathbf{y}$ is an n-dimensional vector in individual space. It is rarely necessary to specify these vectors explicitly but several arguments depend on their existence and orthogonality.

We already have used the name *individual space* for $\mathbb{R}^n$. The subspaces of individual space imposed by a linear model also have convenient names. The space $\mathcal{V}_{\mathbf{1}}$ is called *mean space*, the space $\mathcal{V}_{\mathbf{X}_c}$ is called the *effect space* or the *model space*, and the space $\mathcal{V}_{\mathbf{e}}$ is called the *error space*.

2.1.4 The assumptions of the general linear model

If we adopt the general linear model for a particular data set, then there are three assumptions we are required to make. The logic of fitting the model and testing hypotheses about estimated parameters depends on these assumptions. First, we assume that the sample $\mathbf{y} \in \mathbb{R}^n$ is a set of n observations of n independent random variables Y_i (see Section 1.2), and that each Y_i follows the normal distribution (see Section 6.1). Further we assume that the expected value of each Y_i is a linear combination of the variables x_i specified by the parameters in the vector $\boldsymbol{\beta}$, that is $\mu_{Y_i} = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$. Finally, we assume that the variables Y_i have a common variance σ^2. In summary, for all i, we assume $E\left((Y_i - \mu_{Y_i})(Y_j - \mu_{Y_j})\right) = 0$ for all $j \neq i$ (a consequence of independence), and that the random variable Y_i follows the normal distribution with a mean of $\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$ and a variance of σ^2.

The three assumptions about the random variables Y_i can be reframed as assumptions about

the error component of the general linear model. From equation (2.3), we can write

$$E_i = Y_i - \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}, \tag{2.7}$$

which illustrates the dependence between the random variable E_i and the random variable Y_i, for each $1 \leq i \leq n$. If the variance of Y_i is known, it is clear that the variance of E_i must be the same. Because the true parameter vector $\boldsymbol{\beta}$ minimizes $E(\|\mathbf{E}\|)$, we know that the expected value of each E_i must be zero. It follows that the following three assumptions about the random variables E_i are equivalent to the first set assumptions concerning the random variables Y_i. Therefore, by adopting a linear model we have assumed:

1. All E_i are independent random variables with normal distributions.
2. $E(E_i) = 0$ for all i.
3. $Var(E_i) = \sigma^2$ for all i.

The assumptions about the random error variables play a central role in justifying hypothesis tests of the parameter estimates, $\mathbf{b}$.

2.2 Linear combinations of random variables

To test hypotheses about the parameters of the general linear model requires that we understand the distributions of linear combinations of random variables such as $\overline{Y}$ and S^2, which are linear combinations of the random variables $Y_1, \cdots, Y_n$. Assuming that the Y_is are all random variables with the same normal distribution, it is reasonable to ask for the distribution of linear combinations of these variables such as $\overline{Y}$ and S^2.

We saw in Section 1.3.2 that, given a sample $\mathbf{y}$, the statistic $\overline{y}$ is an unbiased estimator for μ. In this section, we prove that S^2 is an unbiased estimator of σ^2 (see equation 1.9) and find the distributions of $\overline{Y}$ and S^2. These results (and the methods developed to obtain them) afford a rigorous geometric foundation for the F-ratio developed in the next section.

We begin with the following lemma:

Lemma 2.2.1. *Let Y be a random variable with $|Var(Y)| < \infty$. Then for any a and b in $\mathbb{R}$,*

$$Var(aY + b) = a^2\, Var(Y).$$

Proof. Using the definition of variance (see equation 1.9) and the linearity of expectation (see Corollary 1.2.2), we have

$$\begin{aligned}
\mathrm{Var}(aY+b) &= \big(E\big((aY+b) - E(aY+b)\big)\big)^2 \\
&= \big(E\big(aY + b - aE(Y) - b\big)\big)^2 \\
&= \big(E\big(aY - aE(Y)\big)\big)^2 \\
&= a^2\left(E\big(Y - E(Y)\big)\right)^2 \\
&= a^2\mathrm{Var}(Y)
\end{aligned}$$

□

Next, we consider the variance of sums of random variables.

Lemma 2.2.2. *Let $Y_1, \cdots, Y_n$ denote n random variables and suppose the random variables are independent (i.e. $E\left((Y_i - \mu_{Y_i})(Y_j - \mu_{Y_j})\right) = 0$, for all $i \neq j$). Then*

$$Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i).$$

Proof. We apply the definition of variance, the linearity of expectation, and the hypothesis of

independence:

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right) &= E\left(\left(\sum_{i=1}^{n} Y_i - n\mu_i\right)^2\right) = E\left(\left(\sum_{i=1}^{n}(Y_i - \mu_i)\right)^2\right) \\
&= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(Y_i - \mu_i)(Y_j - \mu_j)\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} E\big((Y_i - \mu_i)(Y_j - \mu_j)\big) \\
&= \sum_{i=1}^{n} E(Y_i - \mu_i)^2 + \sum_{i=1}^{n}\sum_{j\neq i} E\big((Y_i - \mu_i)(Y_j - \mu_j)\big) \\
&= \sum_{i=1}^{n} \mathrm{Var}(Y_i)
\end{aligned}$$

□

The main result follows.

Theorem 2.2.3. *Let W be a random variable and suppose $W = \sum_{i=1}^{n} a_i Y_i$, where $a_i \in \mathbb{R}$, for all i, and $Y_1, \cdots, Y_n$ are independent random variables with finite variance. Then*

$$Var(W) = \sum_{i=1}^{n} a_i^2 \, Var(Y_i).$$

Proof. The result is a straightforward consequence of Lemma 2.2.1 and Lemma 2.2.2. □

2.2.1 The variance of the sample mean

The fact that the variance of a sum of independent random variables is the sum of their variances (Lemma 2.2.2) yields another useful fact. We claim that the variance of the sample mean, $\mathrm{Var}(\overline{Y})$, is $\dfrac{\sigma^2}{n}$. This follows easily by applying the definition of $\overline{Y}$ and recalling the assumption that the observations of a sample are assumed to be independent. Since $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, we

have:

$$\begin{aligned}
\mathrm{Var}(\overline{Y}) &= E\big((\overline{Y} - E(\overline{Y}))^2\big) && (2.8)\\
&= \frac{1}{n^2} E\big((\sum Y_i - n\mu)^2\big) && (2.9)\\
&= \frac{1}{n^2} \mathrm{Var}\big(\sum Y_i\big) && (2.10)\\
&= \frac{1}{n^2} \sum \mathrm{Var}(Y_i) = \frac{\sigma^2}{n} && (2.11)
\end{aligned}$$

□

Recall that the variance of a variable is the square of the standard deviation. Considering the variance of $\overline{Y}$, we conclude that the standard deviation of $\overline{Y}$ is

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}, \tag{2.12}$$

where σ is the common standard deviation of the random variables Y_i (see Section 2.1.4).

2.2.2 Sample variance is an unbiased estimator

Whenever the variance of a random variable is calculated from a random sample $\mathbf{y} = (y_1, \cdots, y_n)$, the unbiased estimator stated in equation (1.8) is used. Recall that this differed from the naive estimate s_n^2 (see equation 1.7) by a correction factor of $\frac{n}{n-1}$. It remains to show that the random variable S^2 corresponding to the estimate s^2 is in fact unbiased. We wish to show that $E(S^2) = \sigma^2$ and in the process demonstrate that s_n^2 calculated from sample data is a biased estimate of variance.

We start with a standard algebraic argument that S^2 is unbiased. As with the proof of

Lemma 2.2.2, this argument relies on partitioning a sum of squares.

$$\begin{aligned} E(S^2) &= E\left(\frac{1}{n-1}\sum (Y_i - \overline{Y})^2\right) \\ &= \frac{1}{n-1}E\big(\sum (Y_i - \mu + \mu - \overline{Y})^2\big) \\ &= \frac{1}{n-1}E\big(\sum (Y_i - \mu)^2 - 2n(\overline{Y} - \mu)^2 + n(\overline{Y} - \mu)^2\big) \\ &= \frac{1}{n-1}\Big(\sum E\big((Y_i - \mu)^2\big) - nE\big((\overline{Y} - \mu)^2\big)\Big) \\ &= \frac{1}{n-1}\Big(\sum \mathrm{Var}(Y_i) - n\mathrm{Var}(\overline{Y})\Big) \\ &= \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2 \end{aligned}$$

Although this algebraic argument that S^2 is an unbiased estimator does not support a geometric interpretation, a geometric argument is possible. Given a sample $\mathbf{y}$, to find s^2 we would like to calculate the per-dimension squared length of $\mathbf{y} - \mu\mathbf{1}$, a vector with ambient space $\mathbb{R}^n$. Thus, when $\mu = \mu_{Y_i}$ for all $1 \leq i \leq n$, is known, the statistic $\frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2$ gives an unbiased estimate of σ^2. In the present case, however, the common mean, μ, is unknown. Instead of using the true parameter μ we use the sample mean $\overline{y}$. The key idea here is that one dimension of individual space $\mathbf{y} \in \mathbb{R}^n$ is used to estimate the mean by calculating $\overline{y}$. The true population parameter μ is independent of the sample $\mathbf{y}$, so $\mathbf{y} - \mu\mathbf{1}$ has ambient dimension of n. The situation is different for $\overline{y}$ which instead depends directly on the sample $\mathbf{y}$. Once $\mathbf{y}$ has been decomposed as the sum of the mean vector $\overline{y}\mathbf{1}$ and the centered vector $\mathbf{y}_c = \mathbf{y} - \overline{y}\mathbf{1}$, the centered vector no longer has ambient dimension of n. Instead, the vector $\mathbf{y}_c$ has ambient dimension $n-1$. The desired result follows from precisely the same concept—the statistic s^2 is the average *per-dimension* squared length of the centered data vector $\mathbf{y}_c$. In order to demonstrate this rigorously, we use an approach that will be useful for geometrically justifying hypothesis testing using the F-ratio: we find a convenient orthonormal basis for $\mathbb{R}^n$.

We have seen that $\overline{y}$ can be obtained by projecting $\mathbf{y}$ on the line spanned by $\mathbf{1}$ (see equation 1.4). Let $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ be an orthonormal basis for $\mathbb{R}^n$ chosen so that $\mathbf{u}_1 = \frac{1}{\sqrt{n}}\mathbf{1}$ and with the other basis vectors fixed but unspecified. We can write each basis vector in terms of the original coordinates as $\mathbf{u}_i = (a_{i1}, \cdots, a_{in})^T$. Since $\{\mathbf{u}_i\}$ is a basis, $\mathbf{y} = \sum c_i\mathbf{u}_i$ for some $c_1, \ldots, c_n \in \mathbb{R}$. We

can find each c_i by taking the (signed) length of the projection of $\mathbf{y}$ on $\mathbf{u}_i$, where c_i is negative if the projection of $\mathbf{y}$ on $\mathbf{u}_i$ has the opposite direction as $\mathbf{u}_i$. We know that the coefficient $c_i = \sum_{j=1}^{n} a_{ij} y_j$ for each i, and in particular that $c_1 = \sqrt{n}(\overline{y})$. Since each $\mathbf{u}_i$ is a unit vector, we also have $\sum_{j=1}^{n} a_{ij}^2 = 1$, for all i.

Next we consider c_i for all $1 \leq i \leq n$, as the realized value of the corresponding linear combination of random variables $C_i = \sum_{j=1}^{n} a_{ij} Y_j$. Now if $E(Y_i) = 0$, for all $1 \leq i \leq n$, we have that $E(C_i) = 0$. (This assumption is analogous to the assumption that each coordinate of the error vector $\mathbf{E}$ in the general linear model has an expected value of 0; that is, $E(E_i) = 0$.) By Theorem 2.2.3, we have that

$$\begin{aligned} \mathrm{Var}(C_i) &= \sum a_i \mathrm{Var}(Y_i) \\ &= \sum a_i^2 \sigma^2 \\ &= \sigma^2 \sum a_i^2 \\ &= \sigma^2 \end{aligned}$$

We have proved

Lemma 2.2.4. *Let Y_i, $1 \leq i \leq n$, be independent random variables such that Y_i follows a normal distribution with a mean of 0 and a variation of σ^2 for all i, and suppose that $\mathbf{u} \in \mathbb{R}^n$ is a unit vector. If $C\mathbf{u}$ is the projection of the random variable vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ onto $\mathbf{u}$, then Var$(C) = \sigma^2$.*

By the definition of variance, we see that

$$\begin{aligned} \mathrm{Var}(C_i) &= E\big(\big(C_i - E(C_i)\big)^2\big) \\ &= E\big(C_i^2\big). \end{aligned}$$

It is clear now that

$$E(C_i^2) = \sigma^2, \tag{2.13}$$

and it follows that, with a particular vector of observations $\mathbf{y}$ (the realized values of the random

variable vector $\mathbf{Y}$), the mean of the squared lengths of the vectors $c_i\mathbf{u}_i$ for $1 < i \leq n$ will give the best available estimate of σ^2. It almost goes without saying that the sum $\sum_{i\neq 1} c_i\mathbf{u}_i = \mathbf{y}_c$ because $c_1\mathbf{u}_1 = \overline{y}\mathbf{1}$ (see equation 1.2). We now see the utility of the per-dimension squared length of $\mathbf{y}_c$ for estimating σ^2. To summarize, we have

$$s^2 = \frac{\sum_{i=2}^{n} c_i^2}{n-1} = \frac{\sum_{i=2}^{n} \|c_i\mathbf{u}_i\|^2}{n-1} = \frac{\|\mathbf{y}_c\|^2}{n-1}$$

We are happy to observe that this method of estimating variance agrees with the definition of sample variance (see equation 1.8).

2.3 Testing the estimated model

Statistical analyses, in addition to making estimates of the parameters defining putative true models for variables in a data set, often provide statistical tests for these estimates. *Hypothesis testing* entails a comparison of the estimated model with a simpler model that is described by the *null hypothesis*. Such a test, for example, can provide evidence that the true model is more similar to the estimated model than to a model, say, in which the dependent variable has no systematic relationship with the independent variables.

The general strategy for hypothesis testing with the general linear model is to compare a *restricted model* obtained from a hypothesis that constrains one or more parameters with the *unrestricted model*, which corresponds to an alternative hypothesis without these constraints. The null hypothesis is often the most restricted model (the parameter vector $\boldsymbol{\beta}$ is set to the zero vector so there is no systematic portion in the model). On the other hand, if parameter estimates can vary, the estimated model is the best choice because it has the smallest squared error vector. If the observed sample under the distribution implied by the restricted model is so unusual that this null hypothesis is untenable, then the null hypothesis is rejected in favor of the alternative hypothesis and the estimated model.

The foremost concern, given the sample $\mathbf{y}$ and a vector of parameter estimates $\mathbf{b}$, is to determine how close the estimated model is to the true model. Recall that the general linear

model decomposes each observed y_i as the sum of (1) a linear combination of the x_is (this is the systematic portion of the model) and (2) the term e_i, which is the random portion of the model. Furthermore, by adopting the linear model, we assume that, for all $1 \leq i \leq n$, the expected value of E_i is zero and the variance of E_i is $\sigma^2 > 0$ (see Section 2.1.4). We cannot compare $\mathbf{b}$ with $\boldsymbol{\beta}$ without knowing the true model. Instead, we will compare the estimated model with the model that has no systematic portion.

The model with no systematic portion is precisely the model in which the parameters are each constrained to zero: $\boldsymbol{\beta} = \mathbf{0}$. This is the same as saying that we expect all of the random variables Y_i to behave like the random variables E_i, for if $\boldsymbol{\beta} = \mathbf{0}$ then $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$ certainly implies that $\mathbf{Y} = \mathbf{E}$. It follows that for all $1 \leq i \leq n$, the expected value of Y_i is zero and the variance of Y_i is $\sigma^2 > 0$.

Under any basis for $\mathbb{R}^n$, the coordinates of $\mathbf{Y}$ are each estimates of σ (see the discussion in the previous section, for example), and so the expected value of the squared length of the random vector $\mathbf{Y}$ is $n\sigma^2$. By contrapositive, we observe that if we can show that the expected value of the squared length of $\mathbf{Y}$ is unlikely to be $n\sigma^2$, then we are able to conclude that $\boldsymbol{\beta}$ is unlikely to be the zero vector. In this case, then $\mathbf{b}$ is the best estimate for $\boldsymbol{\beta}$ in the column space of $\mathbf{X}$ (i.e. when the parameters are allowed to vary), and so it is reasonable to conclude that $\mathbf{b}$ is close to $\boldsymbol{\beta}$. This logic is used in every hypothesis test of the estimated model. The crux of the argument is using the observed data to show that we can (or cannot) reasonably expect the expected value of the squared length of $\mathbf{Y}$ to be $n\sigma^2$ given the evidence from the sample $\mathbf{y}$.

Let us call the model with no systematic portion the *null model*. We want to know how likely the sample $\mathbf{y}$ is if reality actually corresponds to the null hypothesis that the linear model is no better than chance at predicting or explaining the observed data $\mathbf{y}$. In other words, the observed data are due merely to chance and have no systematic relationship with the independent variables.

For the sake of argument, we first assume the null model holds and therefore hypothesize that expected squared length of $\mathbf{Y}$ is $n\sigma^2$. Next we suppose that $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n\}$ is an orthonormal basis for $\mathbb{R}^n$ chosen so that $\{\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_{p+1}\}$ span $C(\mathbf{X})$. If $\mathbf{Y}$ is written as $\sum C_i\mathbf{u}_i$, where C_i is a random variable constructed by the appropriate linear combination of the random variables

Y_i, $1 \leq i \leq n$, then the expected value of C_i^2 is σ^2 for all i because we have assumed that $E(Y_i^2) = \sigma^2$ for all i. Thus, we expect the per-dimension squared length of $\hat{\mathbf{Y}}$ to be σ^2:

$$\frac{1}{n}\|\mathbf{Y}\|^2 = \frac{1}{n}\|\sum_{i=1}^{n} C_i\mathbf{u}_i\|^2 = \sigma^2$$

Furthermore, by the least-squares estimate for $\mathbf{b}$, we can express $\mathbf{Y}$ as the sum of the vector $\hat{\mathbf{Y}} \in C(\mathbf{X})$ and the random variable vector $\mathbf{E} \in C(\mathbf{X})^{\perp}$. It follows that the expected per-dimension squared length of each term of $\mathbf{Y}$ will be σ^2:

$$\frac{1}{p+1}\|\hat{\mathbf{Y}}\|^2 = \frac{1}{p+1}\|\sum_{i=1}^{p+1} C_i\mathbf{u}_i\|^2 = \sigma^2$$

and that

$$\frac{1}{n-p-1}\|\mathbf{E}\|^2 = \frac{1}{n-p-1}\Big|\sum_{i=p+2}^{n} C_i\mathbf{u}_i\Big|^2 = \sigma^2.$$

In fact, we expect that the per dimension squared length of $\hat{Y}$ and $\mathbf{E}$ to be the same:

$$\frac{\|\hat{\mathbf{Y}}\|^2}{p+1} = \sigma^2 = \frac{\|\mathbf{E}\|^2}{n-p-1}. \tag{2.14}$$

This justifies why the F-ratio,

$$F = \frac{\|\hat{\mathbf{y}}\|^2/(p+1)}{\|\mathbf{e}\|^2/(n-p-1)}, \tag{2.15}$$

is an appropriate statistic for evaluating the likelihood of the observed data under the assumption of the null model.

If null hypothesis is correct, then we would expect the per-dimension squared length of the sample least squares estimate $\hat{\mathbf{y}}$ to be similar to the per-dimension squared length of the sample error vector $\mathbf{e}$ because of equation (2.14) In this case, we would expect the value of F to be close to 1. On the other hand, if the F-ratio is large then the average squared lengths of the projections of $\hat{\mathbf{y}}$ onto the arbitrary basis vectors $\{\mathbf{u}_i\}$ would be greater than σ^2, implying that the random variables Y_i do not have a mean of zero. Moreover, if F is sufficiently large, then the sample is unlikely under the assumption of the null hypothesis. For this reason, when F is sufficiently large, we have some confidence in rejecting the null hypothesis and accepting the

alternative hypothesis that $\boldsymbol{\beta}$ is not the zero vector. It is important to stress that this procedure does not 'prove' that $\boldsymbol{\beta}$ is close to $\mathbf{b}$. Having decided that in all likelihood $\boldsymbol{\beta} \neq \mathbf{0}$, the estimate $\mathbf{b}$ provides the best guess we can make of the unknown parameter $\boldsymbol{\beta}$ given the sample $\mathbf{y}$.

For example, another common model (also called the *null model* in some texts) is the model where β_1 is left free to be estimated but the rest of the parameters are constrained to 0. The null hypothesis for a test of the estimated model against this model is written

$$H_0 : \boldsymbol{\beta} = (1\ \mathbf{0})^T, \tag{2.16}$$

where $\mathbf{0}$ is a p-dimensional row vector. Under this hypothesis, the restricted model can be written

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1} \end{bmatrix} \begin{bmatrix} b_0 \end{bmatrix} + \mathbf{E}.$$

The alternative hypothesis can be expressed $H_a : \boldsymbol{\beta} \neq 0$ and the corresponding unrestricted model is the full general linear model expressed in equation (2.2). Once the null and alternative hypotheses have been articulated, the corresponding models are fit using the least-squares method and compared using the F-ratio:

$$F = \frac{\|\hat{\mathbf{y}}_c\|^2/(p)}{\|\mathbf{e}_c\|^2/(n-p-1)}.$$

Because one dimension of individual space is spanned by the vector $\mathbf{1}$, the vector $\hat{\mathbf{y}}_c$ has an ambient dimension of p, and $\mathbf{e}$ has an ambient dimension of $(n-p-1)$.

Chapter 3

Geometry and Analysis of Variance

Analysis of variance (ANOVA) is a term that describes a wide range of statistical models appropriate for answering many different kinds of questions about a wide range of experimental and observational data sets. What unites these techniques is that the independent variable(s) in ANOVA are always categorical variables (for example, gender or teacher certification status) and the dependent variable(s) are continuous. Since ANOVA techniques were developed separately from the regression techniques that will be discussed in the next chapter, different vocabulary is used for characteristics of both kinds of models that are actually very similar or identical. For example, the independent variables in ANOVA models are usually called *factors*, whereas the independent variables in regression models are called *predictors*. In a similar way, the coefficients for the independent variables in an ANOVA model are called *effects* but are often called *parameters* in regression analyses. The hypotheses one can test using ANOVA generally concern the differences of means in the dependent variable at different *levels* of a factor, the finite set of discrete values attained by the factor. When a single independent variable or factor is used in the model, then the analysis is called *one-way ANOVA*, and if two factors are used, then the analysis is *two-way ANOVA*.

3.1 One-way ANOVA

In the example from section 1.4, we were interested in whether or not the mean gain-score on a standardized test after one month of tutoring was significantly different from 0. We were only interested in a single population, namely those students who had been tutored for one month. In many kinds of research comparisons between two or more treatment groups are required. For example, it is quite plausible that this particular tutor has no effect over and above the effect of studying alone for an extra month. One-way ANOVA models allow us to compare the means of different groups. When individuals are assigned to treatment groups randomly, it is defensible to conclude that membership in a particular treatment group is responsible for differences in outcomes.

Suppose a tutoring company wanted to research the efficacy of private tutoring to generate data for an advertising campaign. Because individuals who have already elected private tutoring may differ systematically from individuals who have not, they focus on the population of 42 tutees who participate in group tutoring sessions. Of these, three are randomly selected for 1 hour of private tutoring, three are randomly selected for 2 hours of private tutoring, and three are randomly selected as controls (they continue to participate in the group tutoring session). In this design, the treatment factor has three levels:

1. Weekly group tutoring session
2. Weekly 1 hour private tutoring session
3. Weekly 2 hour private tutoring session

The scores before and after one month of tutoring are used to calculate the gain-scores as before. Simulated data for this study are presented in Table 3.1.

3.1.1 Dummy variables

The factors (i.e., independent variables) in ANOVA models specify the factor level for each observation (instead of measurement data) and are called *dummy variables*. These vectors are

Table 3.1: Data for a 3-level factor recording tutoring treatment.

Levels	Observed gain-scores
Group tutoring	6.93
	6.13
	4.25
1 hour private tutoring	11.94
	7.43
	9.43
2 hours private tutoring	12.44
	14.64
	9.17

more easily interpreted if they are orthogonal and when they can be chosen to encode particular hypotheses of interest.

The simplest method of creating dummy variables is to first sort $\mathbf{y}$ by factor level so that all observations of the same level are consecutive. For example, the three gain-scores of students who attended group tutoring might be in the first three slots of $\mathbf{y}$, the three gains-scores of students who received 1 hour of private tutoring might be in the fourth, fifth, and sixth slots, and the scores of the final level might be in the seventh, eight, and ninth slots. We then create a dummy variable $\mathbf{X}_i$ to represent the i^{th} factor level under the convention that $\mathbf{X}_{ij} = 1$ if y_j is an observation of the i^{th} factor level and 0 elsewhere. The dummy variables constructed in this

manner for the tutoring example are presented in following fitted model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

$$\begin{bmatrix} 6.93 \\ 6.13 \\ 4.25 \\ 11.94 \\ 7.43 \\ 9.43 \\ 12.44 \\ 14.64 \\ 9.17 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5.77 \\ 9.60 \\ 12.08 \end{bmatrix} + \begin{bmatrix} 1.16 \\ 0.36 \\ -1.52 \\ 2.34 \\ -2.17 \\ -0.17 \\ 0.36 \\ 2.56 \\ -2.91 \end{bmatrix}.$$

Geometrically, this means that we are considering the orthogonal basis of individual space comprised of the columns of $\mathbf{X}$ and any 6 other arbitrary vectors that span the error space. Notice that with this simple form of dummy coding, the design matrix $\mathbf{X}$ does not include the mean vector $\mathbf{1}$. Moreover, if the mean vector were added as another column in the design matrix $\mathbf{X}$ then this matrix would be singular, since $\mathbf{1}$ is $\mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3$.

3.1.2 Hypothesis tests with dummy variables

Estimating this model using least squares gives the vector $\mathbf{b} = (5.77, 9.60, 12.08)^T$, which can be interpreted as the vector of factor level means. An overall test of the hypothesis that these means are significantly different from 0 can be accomplished by calculating the F-ratio with 3 and 6 degrees of freedom:

$$F = \frac{\|\mathbf{Xb}\|^2/3}{\|\mathbf{e}\|^2/6} \approx \frac{271.45}{4.8584} \approx 55.875$$

This value is so large that, under the assumption that all of the factor level means are 0 ($H_0 : \mu_1 = \mu_2 = \mu_3 = 0$), we would expect an F value this large or larger only 0.009 % of the time.

We can conclude that at least one factor mean is significantly different than zero.

In addition, since the dummy variables are orthogonal, each can be used to test a hypothesis that is independent of the rest of the model. In particular, we can test the hypothesis that each factor level mean is different than zero ($H_0 : \mu_1 = 0$; $H_0 : \mu_2 = 0$; $H_0 : \mu_3 = 0$). The F-ratio for each test is presented below, along with the corresponding *p-value*. The p-value of a hypothesis test is the probability of obtaining an F-ratio as large or larger under the assumption of the corresponding null hypothesis. Notice that the numerator degrees of freedom for each of these tests is 1 because each $\mathbf{X}_i$ is a vector in the chosen orthogonal basis for individual space.

$$\begin{aligned}
F_{\mu_1=0} &= \frac{\|\mathbf{X}(b_1,0,0)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{99.88}{4.8584} \approx 20.558;\ p = 0.003958 \\
F_{\mu_2=0} &= \frac{\|\mathbf{X}(0,b_2,0)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{276.48}{4.8584} \approx 56.908;\ p = 0.000281 \\
F_{\mu_3=0} &= \frac{\|\mathbf{X}(0,0,b_3)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{438.02}{4.8584} \approx 90.158;\ p = 0.000078
\end{aligned}$$

The results of these hypotheses tests suggest that we can reject the null hypotheses that any one of the factor level means is zero; in each case, the gain-score is significantly different from 0 because the p-values are smaller than 0.05. (Other common significance levels are 0.1 and 0.01—researcher judgment and consensus in an academic field guide the choice.)

3.1.3 Dummy variables and contrasts

In spite of the hypothesis tests we were able to preform with the simplest kind of dummy variables, we have not yet answered the most important question we sought to address with the tutoring experiment: Does private tutoring have a different effect on gain-scores than group tutoring? In addition, we might also like to answer the question: Are there differences in the effect on gain-scores between 1 hour and 2 hours of private tutoring? These questions can be answered by using a more clever strategy of constructing dummy variables so that they encode hypotheses of interest.

These more elaborate dummy variables are often called *contrasts*. Researchers will often select contrasts that are orthogonal to each other so the contrasts are independent and the

hypotheses they encode can be tested separately. Although not strictly necessary, most designs include the vector $\mathbf{1}$ in order to test the null hypothesis $H_0 : \mu_i = 0$ where i ranges over all factor levels. In the following discussion, we let $\mathbf{X}_1'$ indicate the vector $\mathbf{1}$.

First we write down the questions we seek to answer and their translation as null hypotheses.

Question	Null Hypothesis
1. Does private tutoring have a different effect on gain-scores than group tutoring?	$H_0 : \frac{\mu_2+\mu_3}{2} - \mu_1 = 0$
2. Are there differences in the effect on gain-scores between 1 hour and 2 hours of private tutoring?	$H_0 : \mu_2 - \mu_3 = 0$

The next step is to find a dummy variable for each question so that when the F-ratio is large, we have evidence to reject the associated null hypothesis. Geometrically, we want to construct a vector in the column space of $\mathbf{X}$ so that the squared length of the projection of $\mathbf{y}$ on this vector can be compared with the average squared length of the projection of $\mathbf{y}$ on arbitrary vectors spanning the error space. In particular, we want relatively large projections on this vector to be inconsistent with the hypothesis that average gain-score for private tutoring is the same as the gain-score for group tutoring. This can be accomplished with (any multiple of) the vector $\frac{1}{2}\mathbf{X}_2 + \frac{1}{2}\mathbf{X}_3 - \mathbf{X}_1$, where the $\mathbf{X}_i$ indicate the simple dummy variables from the previous section. Essentially, we are checking whether our hypothesis about the group means ($H_0 : \frac{\mu_2+\mu_3}{2} - \mu_1 = 0$) is a feasible description of the relationship between observed values in these groups, across the whole data set. Usually, it is convenient to pick dummy values that are integers to ease data-entry, so let $\mathbf{X}_2' = \mathbf{X}_2 + \mathbf{X}_3 - 2\mathbf{X}_1$. This vector works for testing the first hypothesis because if the squared length of the projection of $\mathbf{y}$ on this vector is large then the observed data are unlikely to have come from a population that is described by the null hypothesis (1): Private tutoring (in 1 or 2 hour sessions) has no different effect on gain-scores than group tutoring.

Similarly, we can construct a vector for the null hypothesis corresponding to the second question, Are there differences in the effect on gain-scores between 1 hour and 2 hours of private tutoring? We take $\mathbf{X}_3'$ to be $\mathbf{X}_2 - \mathbf{X}_3$, and reason that large squared lengths of the projection of $\mathbf{y}$ on this vector are inconsistent with the null hypothesis for the second question. Furthermore,

$\mathbf{X}'_1 \cdot \mathbf{X}'_2 = 0$, $\mathbf{X}'_1 \cdot \mathbf{X}'_3 = 0$, and $\mathbf{X}'_2 \cdot \mathbf{X}'_3 = 0$, so the new design matrix $\mathbf{X}' = [\ \mathbf{X}'_1\ \mathbf{X}'_2\ \mathbf{X}'_3\ ]$ is full-rank. The fitted model is

$$\mathbf{y} \;=\; \mathbf{X}'\mathbf{b}' + \mathbf{e},$$

$$\begin{bmatrix} 6.93 \\ 6.13 \\ 4.25 \\ 11.94 \\ 7.43 \\ 9.43 \\ 12.44 \\ 14.64 \\ 9.17 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 0 \\ 1 & -2 & 0 \\ 1 & -2 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} 9.15 \\ 1.69 \\ -1.24 \end{bmatrix} + \begin{bmatrix} 1.16 \\ 0.36 \\ -1.52 \\ 2.34 \\ -2.17 \\ -0.17 \\ 0.36 \\ 2.56 \\ -2.91 \end{bmatrix}.$$

Notice that the error vector is the same as in the previous fitted model; this makes sense because the column space of $\mathbf{X}$ is the same as the column space of $\mathbf{X}'$.

Comparing these two models, we find that only the values of the estimate $\mathbf{b}'$ are different and this is because they have different interpretations. The value $b'_1 \approx 9.15$ can be interpreted as the mean gain-score over all the students. A hypothesis test of this value can allow one to reject the null hypothesis that all three tutoring treatments have average gain-scores of 0.

The value b'_2 is related by a constant to an estimate $\hat{d}_1$ for the average difference in gain-scores between group and private tutoring treatments $d_1 = \frac{\mu_2+\mu_3}{2} - \mu_1$. This constant depends on k, the number of observations at each factor level (in this example $k = 3$) and on the particular dummy variable selected (we chose $2 \cdot \left(\frac{1}{2}\mathbf{X}_2 + \frac{1}{2}\mathbf{X}_3 - \mathbf{X}_1\right)$). When the number of observations at each factor level are not equal, the computation is more complicated but still possible—this case is not discussed here. We have

$$\begin{aligned} E(\mathbf{y}\cdot\mathbf{X}_2') &= -2E(y_{11}+y_{12}+\ldots+y_{1k})+1E(y_{21}+y_{22}+\ldots+y_{2k}) \\ &\quad +1E(y_{31}+y_{32}++\ldots+y_{3k}) \\ &= -2k\mu_1+k\mu_2+k\mu_3=2kd_1, \end{aligned}$$

and so if we assume

$$\mathbf{y}\cdot\mathbf{X}_2'=2k\hat{d_1},$$

then we can compute

$$\begin{aligned} \hat{d_1} &= \frac{\|\mathbf{X}_2'\|^2}{2k}b_2 \\ \hat{d_1} &= \frac{18}{6}b_2=3b_2=5.07. \end{aligned}$$

Thus, we estimate the difference in gain-scores between group and private tutoring to be about 5 points.

In the same way, b_3' is related by a constant to an estimate $\hat{d_2}$ for the average difference in gain-score between the 1 hour and 2 hours private tutoring treatments $d_2=\mu_2-\mu_3$. Following an argument similar to the one above, we compute:

$$\hat{d_2} = \frac{\|\mathbf{X}_3'\|^2}{k}b_3 \tag{3.1}$$

$$\hat{d_2} = \frac{6}{3}b_3=2b_3=-2.48 \tag{3.2}$$

Thus, we estimate that 1 hour of tutoring yields gain-scores that are a little more than 2 points lower than the gain-scores after 2 hours of tutoring.

Next we want to check to see if these values are significantly different from 0. As before, we can compute the F-ratio for the model overall (which tests the null hypothesis $H_0:\mathbf{b}'=\mathbf{0}$) and compute F-ratios for each estimate b_i'. An overall test of the null hypothesis that $\mathbf{b}'=\mathbf{0}$ can be

accomplished by calculating the F-ratio with 3 and 6 degrees of freedom:

$$F = \frac{\|\mathbf{X}'\mathbf{b}'\|^2/3}{\|\mathbf{e}\|^2/6} \approx \frac{271.45}{4.8584} \approx 55.875$$

This value is exactly the same as the F-ratio for the hypothesis test that $\mathbf{b} = \mathbf{0}$ because $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{b}'$. As before, the F value is so large that we would expect an F value this large or larger only 0.009 % of the time. We can conclude that at least one of the estimates in $\mathbf{b}'$ is not zero.

Whenever the vector $\mathbf{1}$ is included in the design matrix, a more sensitive test of the model fit is possible. The general model $\mathbf{y} = \mathbf{X}'\mathbf{b}' + \mathbf{e}$ can be written

$$\mathbf{y} = b_1'\mathbf{X}_1' + b_2'\mathbf{X}_2' + b_3'\mathbf{X}_3' + \mathbf{e},$$

which is equivalent to

$$\mathbf{y} - \overline{y}\mathbf{1} = b_2'\mathbf{X}_2' + b_3'\mathbf{X}_3' + \mathbf{e}.$$

This new model equates the centered vector $\mathbf{y}_c$ with the sum of two (orthogonal) vectors in the effect space and the error vector. All of these vectors (and those spanning the error space) are orthogonal to $\mathbf{1}$, and so analysis can be restricted to the 8-dimensional subspace of individual space orthogonal to $\mathbf{1}$. The null hypothesis becomes $H_0 : b_2' = b_3' = 0$ and the corresponding reduced model has a design matrix made up entirely of the vector $\mathbf{1}$. The corresponding F-ratio has only 2 degrees of freedom for the centered estimate $\hat{\mathbf{y}}_c$ but retains 6 degrees of freedom for the centered error vector $\mathbf{e}_c = \mathbf{e}$.

$$F = \frac{\|\hat{\mathbf{y}}_c\|^2/2}{\|\mathbf{e}_c\|^2/6} \approx \frac{30.347}{4.8584} \approx 6.246.$$

This result has a p-value of 0.03416, so we can reject the hypothesis that $b_2' = b_3' = 0$. This test is more sensitive because we will not reject the overall null hypothesis in those cases where only the intercept of the model b_1' is significantly different from 0.

It remains to see if each of the values estimated by $\mathbf{b}'$ are significantly different than zero.

As with the first set of dummy variables, we can accomplish this by means of three F-ratios:

$$\begin{aligned} F_{b'_1=0} &= \frac{\|\mathbf{X}'(b'_1,0,0)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{753.69}{4.8584} \approx 155.131;\ p = 0.00002 \\ F_{b'_2=0} &= \frac{\|\mathbf{X}'(0,b'_2,0)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{51.44}{4.8584} \approx 10.589;\ p = 0.01738 \\ F_{b'_3=0} &= \frac{\|\mathbf{X}'(0,0,b'_3)^T\|^2/1}{\|\mathbf{e}\|^2/6} \approx \frac{9.25}{4.8584} \approx 1.904;\ p = 0.21685 \end{aligned}$$

From these ratios we can reject the first two null hypotheses: the average overall gain score and the average difference between private and group tutoring are significantly different than 0 (both p-values are below the common threshold of 0.05). However, there is a relatively high chance (p-value = 22%) of obtaining the observed estimate for b'_3 under the null hypothesis $H_0 : b'_3 = 0$ and we cannot reject this hypothesis. We say that the difference in gain scores between the students who received 1 hour and 2 hours of tutoring is not significant.

In fact, the data for the 1 hour and 2 hour treatment gain scores was simulated from normal distributions with different means, but there are not enough scores to separate the pattern from chance. This is a problem of insufficient statistical *power*, and corresponds with the probability of failing to reject a false null hypothesis. In linear models, the primary determinant of power is the size of the sample. Techniques are available to find the minimum sample size for obtaining a sufficiently powerful test so that the probability of failing to reject a false null hypothesis is guaranteed to be less than some predetermined threshold. Further discussion of statistical power is beyond the scope of this thesis.

We conclude this section by observing that the number of independent hypotheses that can be simultaneously tested is constrained by the need to invert the matrix $\mathbf{X}^T\mathbf{X}$ in order to estimate $\boldsymbol{\beta}$ using least squares. When conducting one-way ANOVA, a model of a factor that has k levels can be constructed with $k-1$ dummy variables and the vector $\mathbf{1}$. Any more, and the matrix $\mathbf{X}^T\mathbf{X}$ will be singular and a different technique for finding something analogous to $\mathbf{X}^T\mathbf{X}^{-1}$ is required: the generalized inverse. This extension of the least-squares method is beyond the scope of this thesis.

3.2 Factorial designs

We discuss one other kind of ANOVA design, factorial designs. These models are quite flexible, and, although we only discuss an example of two-way ANOVA, the methods can easily be generalized to any number of factors.

Consider an extension to the tutoring study discussed in the previous section in which the researchers would like to discover if a short content lecture affects the gain-scores experienced by the students who are being tutored. In this new experiment, there are two factors: lecture and the tutoring treatment. Each factor has two levels: students are randomly assigned to attend (or not attend) the lectures, and students are randomly assigned to participate in group tutoring or in private tutoring for a one month period. Simulated data for this example are presented in Table 3.2.

Table 3.2: Data for a 2-factor experiment recording observed gain-scores for tutoring and lecture treatments.

	No lectures	Lectures
Group tutoring	5.00	10.78
	5.58	4.83
	2.60	6.05
Private tutoring	10.83	12.53
	5.17	9.20
	6.68	10.39

3.2.1 Dummy variables with factorial designs

One way we can think of this problem is as a one-way ANOVA of a factor with four levels. The effect or regression space is then the subspace of individual space spanned by the simple dummy variables seen in the last section, $\mathbf{X}_i$ where $\mathbf{X}_{ij} = 1$ whenever $\mathbf{y}_j$ is an observation of factor level i. However, to aid the reader, we instead use subscripts that denote the group: $\mathbf{X}_{gn}$ (group tutoring and no lectures), $\mathbf{X}_{gl}$ (group tutoring and lectures), $\mathbf{X}_{pn}$ (private tutoring and no lectures), and $\mathbf{X}_{pl}$ (private tutoring and lectures). The fitted model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

$$\begin{bmatrix} 5.00 \\ 5.58 \\ 2.60 \\ 10.78 \\ 4.83 \\ 6.05 \\ 10.83 \\ 5.17 \\ 6.68 \\ 12.53 \\ 9.20 \\ 10.39 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 4.393 \\ 7.220 \\ 7.560 \\ 10.707 \end{bmatrix} + \begin{bmatrix} 0.61 \\ 1.19 \\ -1.80 \\ 3.56 \\ -2.39 \\ -1.17 \\ 3.27 \\ -2.39 \\ -0.88 \\ 1.82 \\ -1.51 \\ -0.32 \end{bmatrix}.$$

As in the last example, the coefficients for these dummy variables can be interpreted as the means of each group in the population: group tutoring and no lectures ($\hat{\mu}_{gn} = 4.393$), group tutoring and lectures ($\hat{\mu}_{gl} = 7.220$), private tutoring and no lectures ($\hat{\mu}_{pn} = 7.560$), and private tutoring and lectures ($\hat{\mu}_{pl} = 10.707$). As before, F tests can be used to show that each one of these estimated means is significantly different from zero.

By using these simple dummy variables, each group mean is estimated using only 3 of the 12 data points; the rest of the data are ignored. We are not as interested in each of these four groups, however, as much as we are interested in the overall effect of each factor on the outcome measure. The real strength of factorial designs are the contrasts that can be constructed to make use of all of the data in the experiment, effectively increasing the sample size for the factors of interest. This is beneficial because it increases statistical power without the expense of collecting more data.

3.2.2 Constructing factorial contrasts

To be explicit, consider the question of the effect of lectures on gain-score. We would like to compare all of the individuals in the experiment who attended lectures with those who did not. The null hypothesis for this question might be written $H_0 : \frac{1}{2}(\mu_{gl} + \mu_{pl}) - \frac{1}{2}(\mu_{gn} + \mu_{pn}) = 0$. The appropriate contrast can be formed as a linear combination of the corresponding simple dummy variables: $\mathbf{X}'_2 = -\mathbf{X}_{gn} + \mathbf{X}_{gl} - \mathbf{X}_{pn} + \mathbf{X}_{pl}$. This contrast helps answer the question, Do lectures affect gain scores?

In a similar way, we form the contrast $\mathbf{X}'_3 = -\mathbf{X}_{gn} - \mathbf{X}_{gl} + \mathbf{X}_{pn} + \mathbf{X}_{pl}$ to test the null hypothesis $H_0 : \frac{1}{2}(\mu_{pn} + \mu_{pl}) - \frac{1}{2}(\mu_{gn} + \mu_{gl})$. This null hypothesis corresponds to the question, Does private tutoring affect gain scores?

A final kind of contrast important in factorial ANOVA models is called the *interaction contrast*. This contrast helps to answer the question, Is the increase of gain-scores due to lectures with group tutoring the same as the increase of gain-scores due to lectures with private tutoring? The null hypothesis for this question says there is no difference in the increase of gain-score due to lectures between the two tutoring conditions: $H_0 : (\mu_{gn} - \mu_{gl}) - (\mu_{pn} - \mu_{pl}) = 0$. Constructing the corresponding contrast is straightforward: $\mathbf{X}'_4 = \mathbf{X}_{gn} - \mathbf{X}_{gl} - \mathbf{X}_{pn} + \mathbf{X}_{pl}$. As in the one-way ANOVA example, the first column in the modified design matrix $\mathbf{X}'$ is the vector $\mathbf{1}$.

$$\mathbf{y} = \mathbf{X}'\mathbf{b}' + \mathbf{e},$$

$$\begin{bmatrix} 5.00 \\ 5.58 \\ 2.60 \\ 10.78 \\ 4.83 \\ 6.05 \\ 10.83 \\ 5.17 \\ 6.68 \\ 12.53 \\ 9.20 \\ 10.39 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 7.47 \\ 1.493 \\ 1.663 \\ 0.080 \end{bmatrix} + \begin{bmatrix} 0.61 \\ 1.19 \\ -1.80 \\ 3.56 \\ -2.39 \\ -1.17 \\ 3.27 \\ -2.39 \\ -0.88 \\ 1.82 \\ -1.51 \\ -0.32 \end{bmatrix}.$$

3.2.3 Interpreting hypothesis tests of factorial designs

Since these contrasts are orthogonal, they can each be tested independently for significance using F-ratios and the F-distribution to obtain the p-values for the corresponding null hypotheses.

$$\begin{aligned} F_{b'_1=0} &= \frac{\|\mathbf{X}'(b'_1,0,0,0)^T\|^2/1}{\|\mathbf{e}\|^2/8} \approx \frac{669.61}{5.953} \approx 112.490;\ p = 0.00000 \\ F_{b'_2=0} &= \frac{\|\mathbf{X}'(0,b'_2,0,0)^T\|^2/1}{\|\mathbf{e}\|^2/8} \approx \frac{26.76}{5.953} \approx 4.496;\ p = 0.06680 \\ F_{b'_3=0} &= \frac{\|\mathbf{X}'(0,0,b'_3,0)^T\|^2/1}{\|\mathbf{e}\|^2/8} \approx \frac{33.20}{5.953} \approx 5.577;\ p = 0.04584 \\ F_{b'_4=0} &= \frac{\|\mathbf{X}'(0,0,0,b'_4)^T\|^2/1}{\|\mathbf{e}\|^2/8} \approx \frac{0.08}{5.953} \approx 0.013;\ p = 0.91236 \end{aligned}$$

When studying the results of tests of factorial designs, the first thing to check is the interaction term. As indicated by the high p-value of 0.91, the estimated interaction parameter is quite

likely under the null hypothesis of no interaction. This is the desired result, because when there is a significantly non-zero interaction, we can no longer interpret the parameter estimates b_2' and b_3' as estimates of the main effect for lectures and private tutoring, respectively. The reason for this is that $\mathbf{X}_2'$ averages the increase due to lectures in the case of group and of private tutoring. If there is an interaction, then this average includes not just the main effect of lectures (the increase in gain score due to lectures) but also half of the interaction effect (the increase in gain score over and above the increase due to lectures and the increase due to private tutoring).

In this case, the interaction is very small and also statistically non-significant. This means that we are free to interpret the values b_2' and b_3' as functions only of the increase in gain-scores due to lectures and private tutoring, respectively, as long as these values are significant. To answer the question of statistical significance, we return to the results of the F-ratios. The p-value for the $F_{b_2'=0}$ ratio tells us there is a 6.7% chance of obtaining an estimate of the gain score due to lectures as large or larger, which is greater than the common threshold of 5%. However, depending on the consequences of rejecting a true null hypothesis, many researchers might still consider this a useful estimate of the effect. The p-value for the hypothesis test of main effect of private tutoring is 4.6% and below the common threshold. We can conclude that the increase in gain-scores is significantly different from 0.

However, we must be careful when interpreting the estimates obtained from b_2' and b_3'. As we saw in the last section, we must compute the constant that relates the estimated difference $\hat{d_1}$ in gain scores due to lectures to the value b_2', and find the constant relating b_3' to an estimate $\hat{d_2}$ of the difference in gain scores due to private tutoring. In both cases the squared length of the column vector is 12, we multiplied the contrasts by a constant of 2 to clear fractions, and the number of people in each groups is 3. We compute

$$\hat{d_1} = \frac{12}{3 \cdot 2} b_2' = 4b_2' = 2.97,$$

and

$$\hat{d_2} = \frac{12}{3 \cdot 2} b_3' = 4b_3' = 3.33.$$

We can conclude (since b_2' and b_3' are significantly different from 0) that private tutoring and

short lectures each independently increase gain-scores by about 3 points.

Chapter 4

The Geometry of Simple Regression and Correlation

Regression analysis seeks to characterize the relationship between two continuous variables in order to describe, predict, or explain phenomena. The nature of the data plays an important role in interpreting the results of regression analysis. For example, when regression analysis is applied to experimental data, only the dependent variable is a random variable. The independent variables are fixed by the researcher and therefore not random. In the context of medical research, this kind of data can be used to explain how independent variables such as dosage are related to continuous outcome variables such as blood pressure or the level of glucose in blood. The experimental design and the scientific theory explaining the mechanism by which the drug effects the dependent variable together support causal claims concerning the effect of changes in dosage.

With observational data, regression analysis supports predictions of unknown quantities but the assumption of causality may not be justified or causality may go in the opposite direction. For example, vocabulary and height among children are correlated but this is likely caused by another variable that causes both: the child's age. One can imagine using data from several observatories to estimate the trajectory of a comet, but in fact it is the actual trajectory that causes the data collected by the observatories, not vice-versa. Many businesses and other institutions rely on regression analyses to make predictions. For example, colleges and universities solicit students

scores on standardized tests such as the SAT in order to make enrollment decisions because these scores can partially predict student success.

In the social sciences and economics, experimental data are rare. Regression analysis can be applied to observational data sets as long as appropriate conditions are met (in particular, the independent variables must not be correlated with estimation errors), and regression analysis can be used to analyze data sets that are composed of random, independent variables. In these cases, care must be taken that the presumed regression makes sense from a logical and theoretical point of view. Regressing incidents of lung cancer on tobacco sales by congressional district makes sense because smoking may cause cancer. Interpreting the results of such a study would allow one to make statements such as, "a decrease in tobacco sales by x amount will result in a decrease in cancer incidence by y amount." Indeed, this reasoning might motivate tax policy. However, to speak of increasing the incidence of cancer does not make sense (and were it somehow possible to do so, it is still doubtful this would then cause more people to smoke). Regressing tobacco sales on cancer incidence does not make sense because cancer incidence is not the kind of variable that can be manipulated directly by researchers or society. In certain areas of the social sciences, a dearth of appropriate theoretical explanations may not allow researchers to make causal claims at all, although predictions and descriptions are well warranted and useful.

When regression analysis is applied to observational data for the purpose of prediction, the independent variables are called predictors and the dependent variable is called the criterion. The theoretical assumption of causality is relaxed in this case, but the regression equation still has a meaningful interpretation in the context of prediction. For example, economists regressing income on the number of books owned might concluded that it is reasonable to predict an increase of x dollars above average income for those who own y more books than average. However, it is likely in analyses like these that the number of books is a proxy for other factors that might be more difficult or costly to measure and that presumably cause both book ownership and income. No one would propose giving out books as policy to eradicate poverty (especially if the population was illiterate), but the information about books in the home can be used to adjust predictions about future income.

Regression analysis is very flexible. One flexibility is that the dependent variable in a re-

gression model can be transformed so that the relationship between independent and dependent variables is more nearly linear. For example, population growth is often an exponential phenomenon and if population is the dependent variable for a model that is linear in time, then using the logarithm of the population will likely provide a better-fitting model. Regression analysis is closely related to *correlation analysis* in which both the independent and dependent variables are assumed to be random. This variation is discussed in Section 4.2.

4.1 Simple regression

Before taking up an example of multiple regression, it is worthwhile to consider an example of a simple regression model, regression of a single independent variable on a dependent variable. We take as our example a variation of the tutoring study discussed in the last chapter. Suppose a tutoring company wanted to research the efficacy of particular tutors in order to generate data for hiring decisions. The analysis will use the scores of tutors on the standardized test as the independent variable and the average gain-scores of tutees on the same tests as the dependent variable. To illustrate this example we use the simulated data presented in Table 4.1.

Table 4.1: Simulated score data for 4 tutors employed by the tutoring company.

	Tutor score	Average tutee gain-score
Tutor A	620	7.33
Tutor B	690	8.16
Tutor C	720	11.07
Tutor D	770	11.94

To proceed with regression analysis, we must assume that the three conditions discussed in Section 2.1.4 hold for these data. First, we suppose the model of the true relationship between tutor scores and tutee gain-scores to be of the form

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E},$$

and we assume that all E_i are independent random variables with normal distributions, where $E(E_i) = 0$ and $\text{Var}(E_i) = \sigma^2$ for all i. This is equivalent to assuming that $\mathbf{Y}$ is a vector of random variables Y_i, each with common variance σ^2 and mean μ. Recall that $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is estimated by $\mathbf{b} = (b_0, b_1)^T$ using the method of least squares.

There are two conventions for defining the design matrix $\mathbf{X}$ for linear regression. One option is to set $\mathbf{X} = [\ \mathbf{1}\ \mathbf{x}\ ]$, where $\mathbf{x}$ is the vector of observations of the independent variable, in this case the vector of tutor scores, $\mathbf{x} = (620, 690, 720, 770)^T$. The other option is to use the centered vector $\mathbf{x}_c = \mathbf{x} - \overline{x}\mathbf{1}$ instead of $\mathbf{x}$; this yields the design matrix $\mathbf{X'} = [\ \mathbf{1}\ \mathbf{x}_c\ ]$. We saw in Section 2.1.2 that these matrices produce equivalent results in general, and so we are free to adopt the centered design matrix $\mathbf{X'}$ for the discussion of regression analyses. Using this design matrix is convenient for producing vector drawings of relevant subspaces of individual space because the subspace of individual space spanned by $\mathbf{x}_c$ (and by the centered columns $\mathbf{x}_{ci}$ in general) is orthogonal to $\mathcal{V}_{\mathbf{1}}$.

The fitted model for the tutoring study can now be expressed explicitly:

$$\mathbf{y} \;=\; \mathbf{X'b'} + \mathbf{e},$$

$$\begin{bmatrix} 7.33 \\ 8.16 \\ 11.07 \\ 11.94 \end{bmatrix} = \begin{bmatrix} 1 & -80 \\ 1 & -10 \\ 1 & 20 \\ 1 & 70 \end{bmatrix} \begin{bmatrix} 9.625 \\ 0.033 \end{bmatrix} + \begin{bmatrix} 0.344 \\ -1.135 \\ 0.785 \\ 0.006 \end{bmatrix}.$$

We can interpret this fitted model by providing meaning for the estimated parameters in the vector $\mathbf{b'} = (b'_0, b'_1)^T$ from the given context. In particular, this fitted model gives the overall mean gain-score of $b'_0 = 9.625$ and says that for every point increase in the score of tutors above the mean of 700, the gain-score of the tutees increases by $b'_1 = 0.033$ points. It remains to determine whether the results are statistically significant. Before answering this important question using the familiar F statistic, we briefly discuss the geometry of the fitted regression model.

Since there are only 4 observations of each variable, individual space for this study is $\mathbb{R}^4$. The model is fitted by projecting $\mathbf{y}$ onto the mean space $\mathcal{V}_{\mathbf{1}}$ and the model space $\mathcal{V}_{\mathbf{X}_c}$. Since each of these spaces is a line, there are 2 remaining dimensions in the error space $\mathcal{V}_{\mathbf{e}}$. Understood geometrically, the vector $\mathbf{b}'$ indicates that $\hat{\mathbf{y}}$, the projection of $\mathbf{y}$ into the column space of $\mathbf{X}'$, is the sum of the component in mean space (the vector $9.625(1,1,1,1)^T$) and the component in model space (the vector $0.033(-80,-10,20,70)^T$f). This is illustrated by the vector diagram in Figure 4.1.

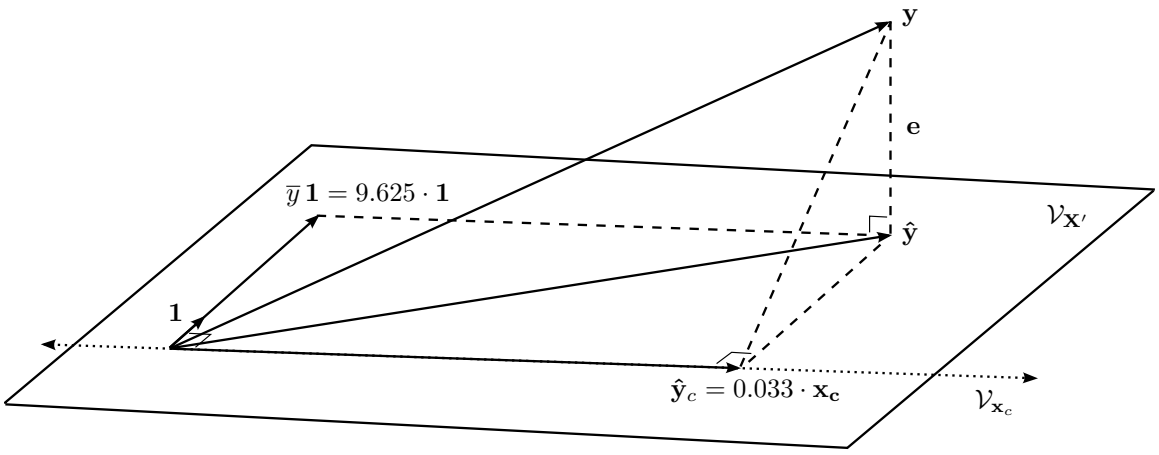


Figure 4.1: The vector $\hat{\mathbf{y}}$ is the sum of two components: $\overline{y}\,\mathbf{1}$ and $\hat{\mathbf{y}}_c$.

These figures give some indication that $\mathbf{y}$ and $\hat{\mathbf{y}}$ are close and thus it is plausible that the fitted model may provide useful information about the relative performance of the tutors. Recall that the least-squares method for obtaining $\mathbf{b}'$ guarantees the $\hat{\mathbf{y}}$ that minimizes the error vector $\mathbf{e}$. (Note that in regression contexts, the error vector is often described as the vector of *residuals*.) As in the case of ANOVA, we rely on the F statistic to provide a rigorous determination of closeness. The important hypothesis to test here is simply that b'_1 is not equal to zero ($H_0 : b'_1 \neq 0$). We wish to know if the average per-dimension squared length of $\hat{\mathbf{y}}$ is significantly greater than the average per-dimension squared length of $\mathbf{e}$, or alternatively if the average per-dimension squared length of $\hat{\mathbf{y}}_c$ is significantly greater than the average per-dimension squared length of $\mathbf{e}_c$. These are two different tests and will have different results. The first compares the error between the regression model and the model that assumes $E(Y_i) = 0$ for all i. The second is more sensitive test, analogous to the F test introduced in Section 1.4.2. In this second test we are comparing

the full regression model with the model that allows $\mathbf{Y}$ to have a non-zero mean. In the first test, $\hat{\mathbf{y}}$ has 2 degrees of freedom, but in the second test $\hat{\mathbf{y}}_c$ has only 1 degree of freedom. In both cases the error vector has 2 degrees of freedom. In either case, the general procedure is the same: if the F statistic is sufficiently large then we can reject the null hypothesis in favor of the alternative hypothesis ($H_\alpha : b'_1 = 0.033$). The results of these tests follow.

$$F_{b'_0=b'_1=0} = \frac{\|\hat{\mathbf{y}}\|^2/2}{\|\mathbf{e}\|^2/2} \approx \frac{191.6986}{1.0117} \approx 189.482,\ p = 0.00525$$

$$F_{b'_1=0} = \frac{\|\hat{\mathbf{y}}_c\|^2/1}{\|\mathbf{e}\|^2/2} \approx \frac{12.837}{1.0117} \approx 12.688,\ p = 0.07057$$

The results of these hypothesis tests are clearly different. The low p-value for the test of the fitted model against the null model provides support for rejecting the hypothesis that $\mathbf{b}' = \mathbf{0}$. However, the second test tells a slightly different story. The hypothesis that $b'_1 = 0$ cannot be rejected at the traditionally accepted level of risk for failing to reject a false null hypothesis (5%). However, another often-used level is 10% and the p-value for the second F test is below this threshold. In some cases, an analyst may decide to reject this null hypothesis in favor of the hypothesis that the coefficient b'_1 is not 0.

It is worth nothing that because the two columns of the design matrix are orthogonal ($\mathbf{1} \cdot \mathbf{x}_c = 0$), the coefficients b'_0 and b'_1 can be tested independently. Thus, we can independently test the hypothesis that μ_Y is zero by comparing the squared length of the other component of $\hat{\mathbf{y}}$, namely $\hat{\mathbf{y}} - \hat{\mathbf{y}}_c = \overline{y}\mathbf{1} = b'_0\mathbf{1}$ (see Figure 4.1), with the squared length of the error vector $\mathbf{e}$:

$$F_{b'_0=0} = \frac{\|\hat{\mathbf{y}} - \hat{\mathbf{y}}_c\|^2/1}{\|\mathbf{e}_c\|^2/2} \approx \frac{370.563}{1.0117} \approx 366.260,\ p = 0.00272.$$

We will see that in multiple regression analyses the columns of $\mathbf{X}$ are not always orthogonal and thus do not afford independent tests of each predictor. This is one of the primary differences between the columns of ANOVA design matrices, which almost always are orthogonal, and design matrices for multiple regression, which are rarely orthogonal.

To conclude our discussion of simple regression, we consider the contribution of the geometry of variable space and compare this with the geometry of individual space presented above.

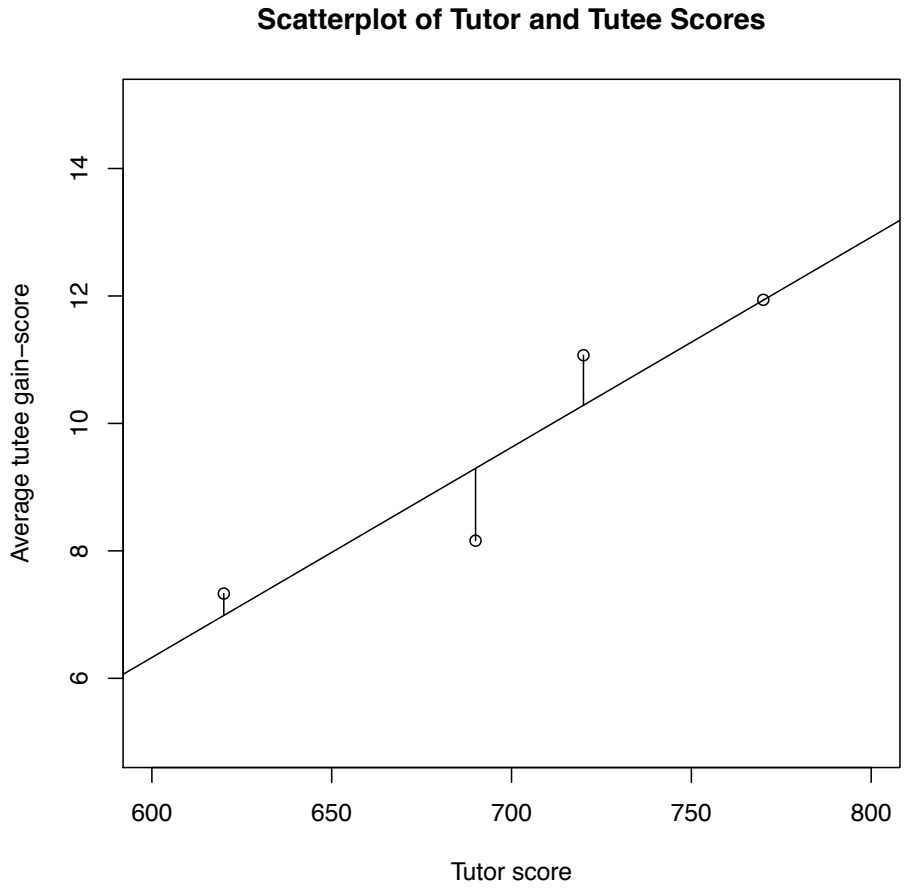


Figure 4.2: A scatterplot for the simple regression example showing the residuals, the difference between the observed and predicted values for each individual.

Recall that scatterplots can be used to represent data sets in variable space; in the case of simple regression, there are just two variables and so scatterplots often provide convenient representations of the data. The scatterplot corresponding to Table 4.1 is provided in Figure 4.2, and shows the *residuals*, the (vertical) differences between each data point and the line representing the least-squares estimates of gain-scores for all tutor scores. The error vector in Figure 4.1 under the natural coordinate system of individual space is the vector of residuals, $(0.344, -1.135, 0.785, 0.006)^T$.

The simple regression example also provides an opportunity to discuss how linear models can

Table 4.2: Modified data for 4 tutors and log-transformed data.

	Tutor score	Average tutee gain-score	Log-transformed ave. tutee gain-score
Tutor A	620	7.33	1.992
Tutor B	690	8.16	2.099
Tutor C	720	9.77	2.280
Tutor D	770	11.94	2.480

be extended with *link functions*. For example, suppose Tutor C had an average tutee gain-score of 9.77 instead of 11.07. In this case, transforming the dependent variable by taking the natural logarithm produces a data set that is more nearly linear than the untransformed data (see Table 4.2). The scatterplot with the least-squares estimation line (as well as the corresponding, untransformed analog) is shown in Figure 4.3.

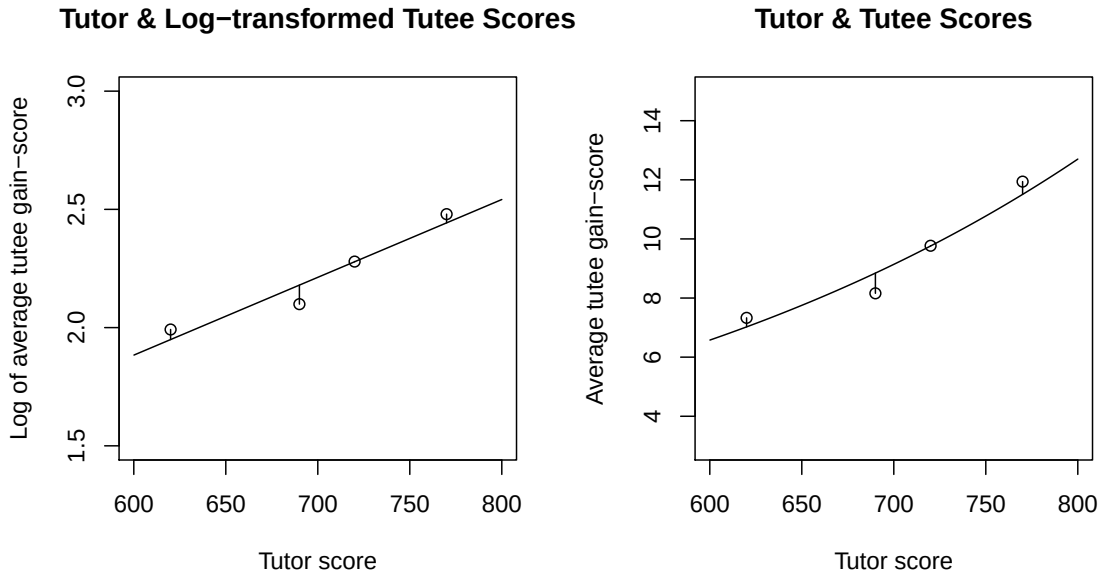


Figure 4.3: Least-squares estimate and residuals for the transformed and untransformed data.

4.2 Correlation analysis

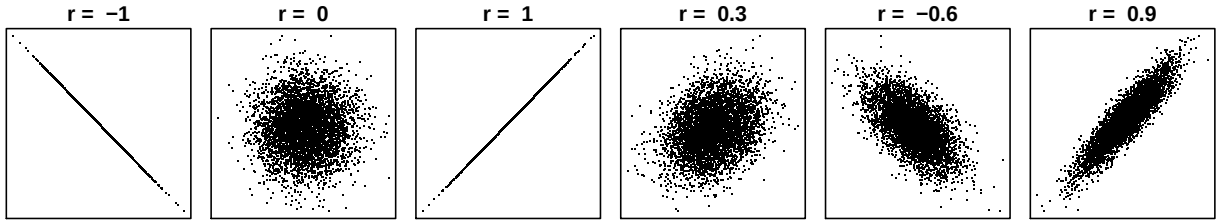
In the most basic sense, correlation is a measure of the linear association of two variables. The conceptual distinction between cause-effect relationships and mere association did not long precede the development of a statistical measure of this association. In the middle of the nineteenth century, the philosopher John Stuart Mill recognized the associated occurrence of events as a necessary but insufficient criterion for causation (Cook and Campbell, 1979). His philosophical work set the stage for Sir Francis Galton, who defined correlation conceptually, worked out a mathematical theory for the bivariate normal distribution by 1885, and also observed that correlation must always be less than 1. In 1895, Karl Pearson built on Galton's work, developing the most common formula for the correlation coefficient used today. Called the product-moment correlation coefficient or Pearson's r, this index of correlation can be understood as the dot product of the normalized, centered variables:

$$r_{xy} = \frac{\sum(x_i - \overline{x})(x_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}} = \frac{\sum(x_{ci})(y_{ci})}{\sqrt{\sum(x_{ci})^2 \sum(y_{ci})^2}} = \frac{\mathbf{x}_c}{\|\mathbf{x}_c\|} \cdot \frac{\mathbf{y}_c}{\|\mathbf{y}_c\|} \tag{4.1}$$

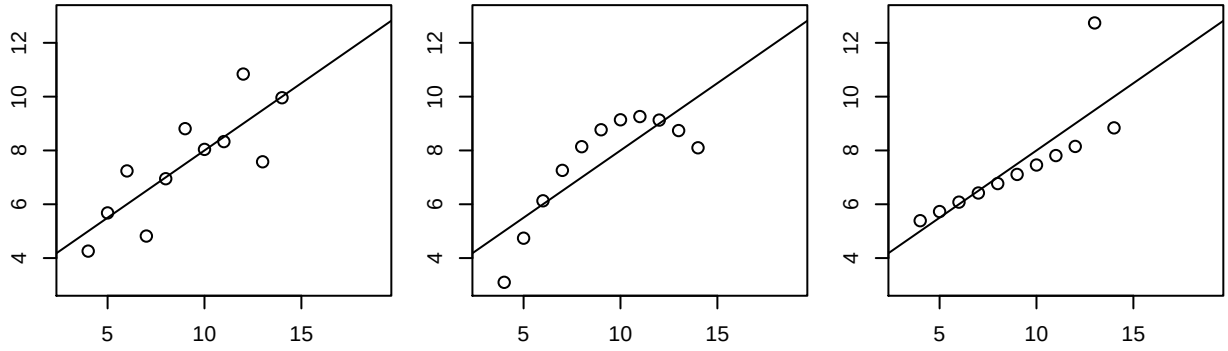
It is straightforward to establish that $|r| \leq 1$ using the Cauchy-Schwarz Inequality.

In variable space, the correlation coefficient can be understood as an indication of the linearity of the relationship between two variables. It answers the question: How well can one variable be be approximated as a linear function of the other? In introductory statistics texts, panels of scatterplots are frequently presented to illustrate what various values of correlation might look like (Figure 4.4). These diagrams may lead to the misconception that the correlation tells more about a bivariate relationship than in fact it does (for example, see Anscombe, 1973). In practice, it is often difficult to estimate the correlation by looking at a scatterplot or given the correlation, to obtain a clear sense of what the scatter plot might look like.

In individual space, however, the correlation between two variables has the simple interpretation of being the cosine of the angle between the centered variable vectors. Much of the power of the vector approach is derived from this straightforward geometric interpretation of correlation. This relationship is clear when we consider the 2-dimensional subspace spanned by the centered variable vectors (Figure 4.5). Given the centered vectors $\mathbf{x}_c$ and $\mathbf{y}_c$, a right triangle is formed



(a) Typical correlation examples.



(b) All have correlation of 0.816 (Anscombe, 1973).

Figure 4.4: (a) Panels of scatter plots give an idealized image of correlation, but in practice, (b) plots with the same correlation can vary quite widely.

by $\mathbf{y}_c$, the projection of $\mathbf{y}_c$ onto $\mathbf{x}_c$ (called $\hat{\mathbf{y}}_c$), and the difference between these vectors, $\mathbf{y}_c - \hat{\mathbf{y}}_c$. The cosine of an angle in a right triangle is the ratio of the adjacent side and the hypotenuse. When the lengths of $\mathbf{x}_c$ and $\mathbf{y}_c$ are 1, the cosine of the angle between them is simply the length of the projection $\hat{\mathbf{y}}_c$, the quantity $\mathbf{x}_c \cdot \mathbf{y}_c$. More generally, the cosine of the angle between two centered vectors is the dot product of the centered, normalized vectors:

$$\cos(\theta_{\mathbf{x}_c\mathbf{y}_c}) = \frac{\|\text{Proj}_{\mathbf{y}_c}\mathbf{x}_c\|}{\|\mathbf{x}_c\|} = \frac{\frac{\mathbf{x}_c\cdot\mathbf{y}_c}{\|\mathbf{y}_c\|^2}\|\mathbf{y}_c\|}{\|\mathbf{x}_c\|} = \frac{\mathbf{x}_c \cdot \mathbf{y}_c}{\|\mathbf{x}_c\|\ \|\mathbf{y}_c\|} = r \tag{4.2}$$

When $\hat{\mathbf{y}}_c$ lies in the opposite direction as $\mathbf{x}_c$, the correlation is negative.

Correlation analysis usually involves using r to estimate the parameter called Pearson's ρ, which is defined by $\rho = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}$, where σ_{XY} is the covariation of the random variables X and Y. Both variables are assumed to be normal and to follow a bivariate normal distribution. The

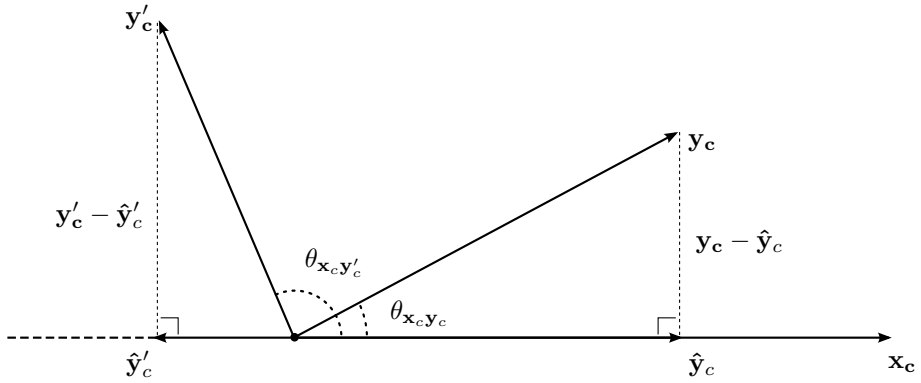


Figure 4.5: The vector diagram illustrates that $r_{\mathbf{x}_c\mathbf{y}_c} = \cos(\theta_{\mathbf{x}_c\mathbf{y}_c})$ and that $r_{\mathbf{x}_c\mathbf{y}'_c} = -\cos(\pi - \theta_{\mathbf{x}_c\mathbf{y}'_c})$.

most common hypothesis test is whether or not the correlation of two variables is significantly different from zero. The test statistic used for this test follows *Student's* t distribution (see Section 6.4) which when squared is equivalent to the F-ratio for the same hypothesis test with 1 degree of freedom in the numerator (see Section 1.4.2 and 6.3).

For example, consider the variable S defined as the average total scores on state assessments per district of eighth graders in Massachusetts at the end of the 1997-1998 school year and the variable I defined as the per capita income in the same school districts during the tax year 1997. A random sample of 21 districts yields a sample correlation between S and I of $r_{SI} = 0.7834$. This suggests there may be a moderate positive linear association between the variables. We want to test whether this correlation is significantly different from 0 because a correlation of 0 means there is no association. As in the last section, first we establish the null hypothesis, $H_0 : \rho = 0$, and the alternative hypothesis, $H_1 : \rho \neq 0$. We assume the null hypothesis is true and compute the probability of obtaining the observed correlation. If it is low, we reject the null hypothesis and conclude that the correlation of the variables is positive.

The test statistic can be computed:

$$\frac{r_{XY}\sqrt{n-2}}{\sqrt{1-r_{XY}^2}} = \frac{0.7834\sqrt{21-2}}{\sqrt{1-0.7834^2}} = 5.4942.$$

This number is the t value corresponding to the observed correlation. The probability of obtain-

ing a test statistic at least this high can be computed $1 - \int_{5.4942}^{\infty} f(t)dt \approx 10^{-5}$ where $f(t)$ is the distribution function of Student's t-distribution with $21 - 2 = 19$ degrees of freedom. Because the probability is well below the standard threshold of 0.05, we reject the null hypothesis in favor of the alternative hypothesis, concluding that there is a positive correlation between districts' per capita income and the average total score on the eighth grade state exam.

Although correlation plays an important role in research, it frequently does not give the most useful information about a data set. Fisher (1958) wrote, "The regression coefficients are of interest and scientific importance in many classes of data where the correlation coefficient, if used at all, is an artificial concept of no real utility" (p. 129). Correlations are easy to compute but often hard to interpret. Even correct interpretations might not answer the instrumental question at the heart of most scientific research, How do manipulations of one variable affect another? In many cases, social scientists and market analysts are content to avoid addressing causality and instead answer the different question: How do variations in some variables predict the variation in others?

Chapter 5

The Geometry of Multiple Regression

In the last chapter, we saw that one flexibility of regression analysis is that variables can be transformed via (not necessarily linear) functions and in this way used to model non-linear phenomena. Another flexibility is the option to use more than one continuous predictor. Regression models that include more than one independent variable are called *multiple regression* models. We begin with a two-predictor multiple regression model using data from the Massachusetts Comprehensive Assessment System and the 1990 U.S. Census on Massachusetts school districts in the academic year 1997-1998. (This data set is included with the statistical software package R.)

Given data from all 220 Massachusetts school districts, suppose Y denotes the per-district average total score of fourth graders on the Massachusetts state achievement test, and suppose X_1 and X_2 denote the per-capita income and the student-teacher ratio, respectively. We want to predict the district average total score (Y) given the per-capita income and the student-teacher ratio (X_1 and X_2). We hypothesize that higher student-teacher ratio will be predictive of a lower average total score and that greater per-capita income will be predictive of higher average total scores. The first 5 data points are shown in Table 5.1.

The model for multiple regression with two predictors is similar to a two-way ANOVA model; the only difference is that the vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ contain measurement data instead of numerical

Table 5.1: Sample data for Massachusetts school districts in the 1997-1998 school year. Source: Massachusetts Comprehensive Assessment System and the 1990 U.S. Census.

District name	Average total achievement score (fourth grade)	Ratio of students per teacher	Per capita income (thousands)	Percent free or reduced lunch	Percent English learners
Abington	714	19.0	16.379	11.8	0
Acton	731	22.6	25.792	2.5	1.246
Acushnet	704	19.3	14.040	14.1	0
Agawam	704	17.9	16.111	12.1	0.323
Amesbury	701	17.5	15.423	17.4	0
⋮	⋮	⋮	⋮	⋮	⋮
N=220	$\overline{\mathbf{y}} = 709.827$ $s_{\mathbf{y}} = 15.126$	$\overline{\mathbf{x}}_1 = 17.344$ $s_{\mathbf{x}_1} = 2.277$	$\overline{\mathbf{x}}_2 = 18.747$ $s_{\mathbf{x}_2} = 5.808$	$\overline{\mathbf{x}}_3 = 15.316$ $s_{\mathbf{x}_3} = 15.060$	$\overline{\mathbf{x}}_4 = 1.118$ $s_{\mathbf{x}_4} = 2.901$

tags for specifying factor levels and contrasts. The design matrix has three columns: $\mathbf{X} = [\mathbf{1}\ \mathbf{x}_1\ \mathbf{x}_2]$. The first 5 data points from Table 5.1 are shown below.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e},$$

$$\begin{bmatrix} 714 \\ 731 \\ 704 \\ 704 \\ 701 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & 19.0 & 16.379 \\ 1 & 22.6 & 25.792 \\ 1 & 19.3 & 14.040 \\ 1 & 17.9 & 16.111 \\ 1 & 17.5 & 15.423 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 699.657 \\ -1.096 \\ 1.557 \end{bmatrix} + \begin{bmatrix} 9.673 \\ 15.967 \\ 3.642 \\ -1.116 \\ -3.483 \\ \vdots \end{bmatrix}.$$

As always, the goal of the analysis is to find the $\mathbf{b} = (b_0, b_1, b_2)^T$ so that the squared length of the vector $\mathbf{e} = [e_1, e_2, \cdots]^T$ is minimized. One geometric interpretation of the values e_i is as

the vertical distances between the i^{th} data point and the *regression plane* in a three-dimensional scatterplot. This interpretation corresponds to the two-dimensional interpretation of residuals (see Figure 4.2) and is illustrated in Figure 5.1. The measure of closeness is the sum of the squared lengths of the vertical error components, and the scatterplot provides a rough sense of how well the model fits the data.

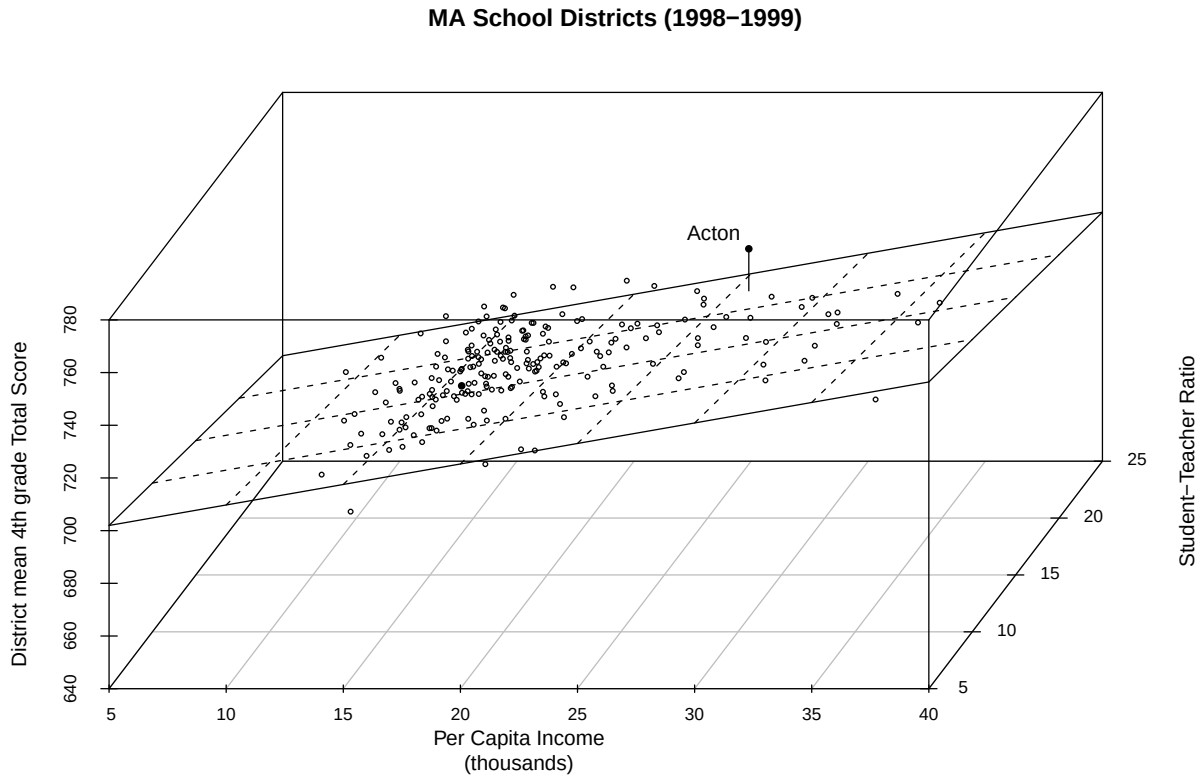


Figure 5.1: The data are illustrated with a 3D scatter plot that also shows the regression plane and the error component for the prediction of district mean total fourth grade achievement score (e_2) in the Acton, MA.

To make use of the second geometric interpretation (a diagram of vectors in individual space), it is necessary to center the data vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The new design matrix is $\mathbf{X'} = [\ \mathbf{1}\ \mathbf{x}_{c1}\ \mathbf{x}_{c2}]$, and the corresponding model is $\mathbf{y} = \mathbf{X'b'} + \mathbf{e}$. As noted before, this does not change the model space because $C(\mathbf{X})$ is equal to $C(\mathbf{X'})$. In general, $b'_i = b_i$ when $i \neq 0$, but the first coefficients in the two models are not necessarily equal. The value b_0 denotes the intercept or the expected value of Y when $X_1 = X_2 = 0$. The value of b'_0 instead denotes the mean value $\overline{y}$ which is also

the expected value of Y when $X_1 = \overline{x}_1$ and $X_2 = \overline{x}_2$.

$$\begin{aligned} \mathbf{y} &= \mathbf{X'b'} + \mathbf{e}, \\ \begin{bmatrix} 714 \\ 731 \\ 704 \\ 704 \\ 701 \\ \vdots \end{bmatrix} &= \begin{bmatrix} 1.000 & 1.656 & -2.368 \\ 1.000 & 5.256 & 7.045 \\ 1.000 & 1.956 & -4.707 \\ 1.000 & 0.556 & -2.636 \\ 1.000 & 0.156 & -3.324 \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 709.827 \\ -1.096 \\ 1.557 \end{bmatrix} + \begin{bmatrix} 9.673 \\ 15.967 \\ 3.642 \\ -1.116 \\ -3.483 \\ \vdots \end{bmatrix}. \end{aligned}$$

We can check by inspection that the error vectors in both models are the same (recall that this follows from the orthogonality of mean space and error space). By considering the geometric representation of the models in individual space, we can see that the models are simply two different ways to write the explanatory portion of the model, $\mathbf{Xb} = \mathbf{X'b'}$. These two ways of writing the model essentially correspond to different choice of basis vectors for model space. In fact, we see that:

$$\begin{aligned} \mathbf{y} &= \mathbf{1}\overline{\mathbf{y}} + \mathbf{x}_{c1}b_1 + \mathbf{x}_{c2}b_2 + \mathbf{e} \\ \mathbf{y} - \mathbf{1}\overline{\mathbf{y}} = \mathbf{y}_c &= \mathbf{x}_{c1}b_1 + \mathbf{x}_{c2}b_2 + \mathbf{e} \end{aligned}$$

Because the vectors $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ are both orthogonal to $\mathbf{1}$, by choosing the centered representation, it is possible to view only the vectors orthogonal to $\mathbf{1}$ in the vector diagram, leaving enough dimensions to show the relationship among $\mathbf{y}_c$ and the vectors $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$ under the constraints of a two-dimensional figure (see Figure 5.3). We can create similar diagrams schematically with a greater number of predictors by representing hyper planes using planes or lines. In vector diagrams, the measure of fit is the squared length of the error vector. The representation in a single entity, the error vector $\mathbf{e}$, of all of the error across the entire data set is one of the

strengths of vector diagrams and the geometric interpretation afforded by individual space.

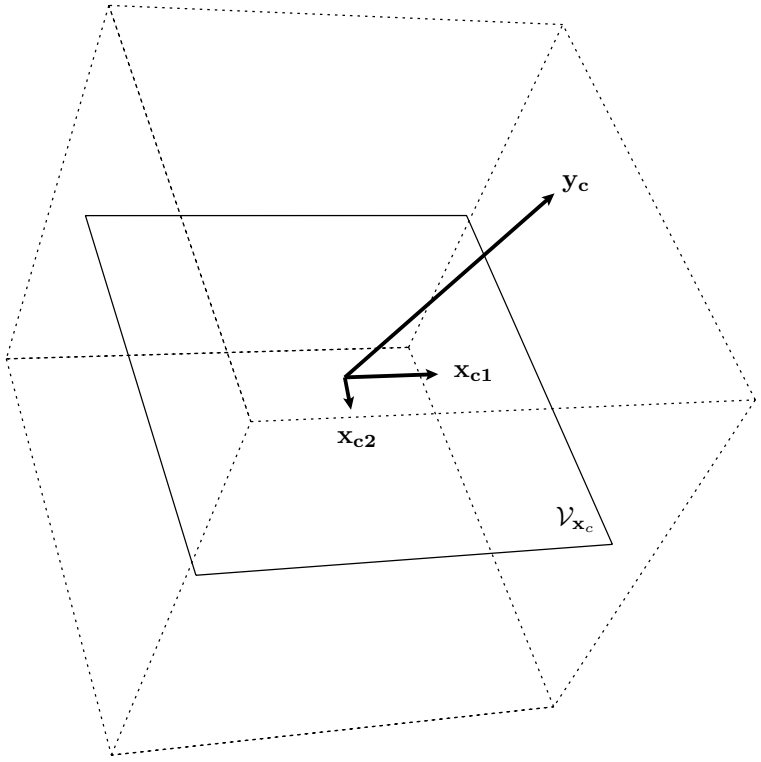


Figure 5.2: The geometric relationships among the vectors $\mathbf{y}_c$, $\mathbf{x}_{c1}$, and $\mathbf{x}_{c2}$.

After representing the multiple regression solution in these two ways, it remains to determine if the vector $\hat{\mathbf{y}} = \mathbf{X}'\mathbf{b}'$ actually provides a better prediction of $\mathbf{y}$ than chance. As before, there are two competing hypotheses to consider. On the one hand, the null hypothesis states that the population parameters $\boldsymbol{\beta}'_1 = \boldsymbol{\beta}'_2 = 0$. The alternative hypothesis is that $\hat{\mathbf{y}}$ is indeed close to $E(\mathbf{Y})$ and that it is consequently acceptable to use $\mathbf{X}_1$ and $\mathbf{X}_2$ to predict $\mathbf{Y}$.

As before, we use the F-ratio to compare the estimate of the variance of $\mathbf{Y}$ obtained from the average per-dimension squared length of the error vector with the estimate of the variance of $\mathbf{Y}$ obtained from the average per-dimension squared length of $\hat{\mathbf{y}}$. Under the null hypothesis, these estimates should be fairly close and the F-ratio should be small. On the other hand, we reject the null hypothesis if the F-ratio is large—if the estimates of the variance of $\mathbf{Y}$ by projecting $\mathbf{y}$ into model space and into error space are significantly different. In the case of this model, we

have

$$F_{b'_1=b'_2=0} = \frac{\|\hat{\mathbf{y}}_c\|^2/1}{\|\mathbf{e}_c\|^2/217} \approx \frac{10402.5}{135.042} \approx 77.03,\ p = 0.00000.$$

Based on this analysis, we can reject the null hypothesis and tentatively conclude that $\hat{\mathbf{y}}$ is close to $\mathbf{Y}$. Note that the very small (and likely non-zero) p-value is due to the high degrees of freedom in the denominator (see Section 6.3). It is important to stress, however, that one cannot rule out the possibility that other models provide even better predictions of $\mathbf{Y}$. In the following sections, we will see extensions of this example that illustrate this point.

5.1 Multiple correlation

One way to measure overall model fit is the generalization of the correlation coefficient, r, called the *multiple correlation coefficient*; it is denoted R. In terms of the geometry of individual space, we saw that the notation r_{xy} indicates the cosine of the angle between the (centered) vectors $\mathbf{x}$ and $\mathbf{y}$. Multiple correlation is the correlation between the criterion variable and the projection of this variable onto the span of the predictor variables. We thus define

$$R_{y.x_1 \ldots x_n} = r_{\mathbf{y}_c \hat{\mathbf{y}}_c},$$

where $\hat{\mathbf{y}}_c$ is the projection of $\mathbf{y}_c$ onto the space $C([\mathbf{x}_{c1}, \ldots, \mathbf{x}_{cn}])$.

A simple geometric argument is sufficient to justify the role of the *squared multiple correlation coefficient*, R^2, as the measure of the proportion of the variance of the criterion $\mathbf{y}_c$ that is explained by the model in question. Consider the triangle formed by $\mathbf{y}_c$, $\hat{\mathbf{y}}_c$, and the error vector $\mathbf{e}$. Since $\hat{\mathbf{y}}_c$ and $\mathbf{e}$ are orthogonal, the Pythagorean theorem provides the following equation:

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

The squared correlation is the squared cosine and must therefore be $\frac{||\hat{\mathbf{y}}_c||^2}{||\mathbf{y}_c||^2}$. Now a frequently-used estimator for the variance of a random variable vector $\mathbf{Y}$ is the per-dimension squared length of $\mathbf{y}_c$ (see equation 1.3.4). Thus, the ratio of the variance explained by the model ($\|\hat{\mathbf{y}}_c\|^2/(n-1)$) to the sample variance ($\|\mathbf{y}_c\|^2/(n-1)$) is simply the ratio $\frac{||\hat{\mathbf{y}}_c||^2}{||\mathbf{y}_c||^2}$. This equality demonstrates

that R^2 can be used to assess model fit. For example, if two models are proposed, then the one with the greater R^2 value is said to be the model with the better fit.

In the present example, the squared length of the vector $\hat{\mathbf{y}}_c$ is 20805.1 and the squared length of $\mathbf{y}_c$ is 50109.44. Therefore, we have that $R^2_{y.x_1x_2} = 0.4152$, and we can say that this model, which uses the student-teacher ratio and the district average per-capita income to predict district average total MCAS scores in the fourth grade, explains a little more than 40% of the observed variance among these district average scores.

5.2 Regression coefficients

We turn now to the task of interpreting the model coefficient vector $\mathbf{b}$. The first coefficient is the easiest to deal with; it is simply the value of $\mathbf{y}$ when the predictors are both $\mathbf{0}$. The meaning of the first coefficient is slightly different for the centered and non-centered models. One must decide if the data and the meaning of the variables allow a prediction of $\mathbf{y}$ when both predictors are zero (i.e., when $\mathbf{x}_1 = \mathbf{x}_2 = \mathbf{0}$). However, in centered models, $\mathbf{x}_{c1} = \mathbf{x}_{c2} = \mathbf{0}$ when $\mathbf{x}_1 = \overline{x}_1\mathbf{1}$ and $\mathbf{x}_2 = \overline{x}_2\mathbf{1}$. It follows that b_0 is the model prediction of $\mathbf{Y}$ in the case that both predictors attain their mean values.

One might be tempted to interpret the regression coefficients $\mathbf{b}'_1$, and $\mathbf{b}'_2$ in the last example as the respective changes produced in the mean total score per unit change in the student-teacher ratio and per capita income. However, it is quite possible that other variables are actually responsible for the change in average test score and only happen to be associated with the variables in the model as well. For example, the cock's crow precedes and is highly correlated with sunrise but the spinning earth causes both phenomena.

In ANOVA designs with 2 or more factors, the orthogonality of the factors means we can interpret each model coefficient independently. In regression analyses, however, the predictors are often correlated. This means that regression coefficients cannot be interpreted without considering the whole model. The relationship can certainly be expressed $\mathbf{x}_2 = \mathbf{x}_1 + \mathbf{z}$, for some vector $\mathbf{z}$. Suppose, for example, that the predictors $\mathbf{x}_1$ and $\mathbf{x}_2$ are correlated. Then in the model

$$\mathbf{y} = b_1\mathbf{x}_1 + b_2\mathbf{x}_2,$$

and if $\mathbf{x}_1$ is increased by $\mathbf{1}$, then we have

$$\mathbf{y}' = b_1(\mathbf{x}_1 + \mathbf{1}) + b_2(\mathbf{x}_1 + \mathbf{1} + \mathbf{z}) = \mathbf{y} + (b_1 + b_2)\mathbf{1}.$$

If the predictors were independent, then one would expect

$$\mathbf{y}' = \mathbf{y} + b_1\mathbf{1},$$

instead.

The best interpretation of the regression coefficient b_i when predictors are correlated is as the increase in the criterion variable per unit increase in the predictor variable, *holding all other variables constant*. In applied contexts, however, this might not actually make much sense. For example, it might not be feasible for a school district to hire more teachers (increasing the student-teacher ratio) without a larger per capita income and concomitantly higher tax revenue because of budgetary constraints.

Without careful design and theoretical support for the variables included in the model and reasonable confidence that these variables are the only relevant ones, multiple regression offers at best prediction with regression coefficients that have little value for generating explanations or identifying causal effects.

5.3 Orthogonality, multicollinearity, and suppressor variables

Multiple regression analyses can have features that seem counter-intuitive or paradoxical on the first take and the explanation of many of these is greatly facilitated by vector diagrams of individual space and the associated geometric orientation. This section elucidates four of these features: orthogonality, multicollinearity, partial correlation, and suppressor variables.

We have already discussed the caution required when interpreting regression coefficients in a model that has non-orthogonal predictors. When predictors in a model are orthogonal, the estimates of corresponding regression coefficients can be interpreted independently of one another. Moreover, as we saw in the section on factorial contrasts (see Section 3.2.2), the

associated regression coefficients can be tested independently for significance. Experimental design allows the researcher to ensure that predictors are orthogonal, but this is rarely the case in observational research. What is more important than whether or not two predictors are orthogonal is *how close* the predictors are to being orthogonal. Correlation, especially understood in relation to the measure of the angle formed between two centered vectors in individual space, is especially helpful for quantifying this relationship.

To examine the impact of near orthogonality and its absence, consider three models of the mean total fourth grade scores for Massachusetts districts, summarized in Table 5.1. The first model is

$$\mathbf{y} = [\ \mathbf{1}\ \mathbf{x}_{c1}\ \mathbf{x}_{c2}\ ]\mathbf{b}' + \mathbf{e}$$

and was presented above: the variables of student-teacher ratio and per capita income predict mean total fourth grade achievement. The second model is simpler, using only per capita income variable to predict achievement and can be expressed

$$\mathbf{y} = [\ \mathbf{1}\ \mathbf{x}_{c2}\ ]\mathbf{b}' + \mathbf{e}.$$

The third model is like the first, but uses the percentage of students eligible for free or reduced lunch, $\mathbf{x}_3$, instead of the student-teacher ratio. It can be expressed

$$\mathbf{y} = [\ \mathbf{1}\ \mathbf{x}_{c2}\ \mathbf{x}_{c3}\ ]\mathbf{b}' + \mathbf{e}.$$

The estimates of the regression coefficient b_2' from each model are presented in Table 5.2 along with the squared length of the corresponding projection of $\mathbf{y}$ onto $\mathbf{x}_{c2}$, the associated degrees of freedom, and the average per dimension length.

One important observation to be made about the regression coefficient for the per capita income variable is that it is seems quite similar (but yet different) in the first two models and very different in the third model. In is outside the scope of this thesis to describe how one decides whether differences in these estimates are significant. However, calculating the p-values assures us that all three F-ratios are significantly different than 0 and, using statistical methods

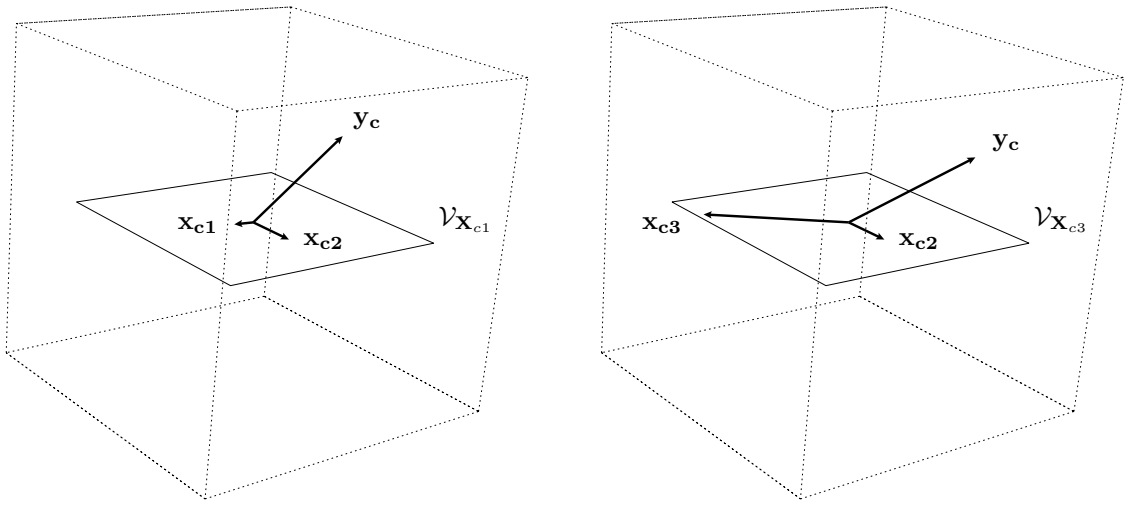
Table 5.2: The value of the regression coefficient for per capita income and corresponding F-ratios in three different models of mean total fourth grade achievement.

Design matrix	Source of variation	Squared length	Degrees of freedom	Ave. length per dim.	F-ratio	Estimate of b_2'
$\mathbf{X}_1' = [\ \mathbf{1}\ \mathbf{x}_{c1}\ \mathbf{x}_{c2}\ ]$	$b_2'\mathbf{x}_{c2}$	19475.2	1	19475.2	144.215	1.557
	$\mathbf{e}$	29304.3	217	135.0		
$\mathbf{X}_2' = [\ \mathbf{1}\ \mathbf{x}_{c2}\ ]$	$b_2'\mathbf{x}_{c2}$	19475.2	1	19475.2	138.59	1.624
	$\mathbf{e}$	30634	218	140.5		
$\mathbf{X}_3' = [\ \mathbf{1}\ \mathbf{x}_{c2}\ \mathbf{x}_{c3}\ ]$	$b_2'\mathbf{x}_{c2}$	19475	1	19475.2	250.64	0.694
	$\mathbf{e}$	16861	217	77.7		

outside the scope of the present discussion, the estimates from the first and second model are not significantly different but the third estimate is significantly different from the first two. Recall that the independence of predictors in an orthogonal design implies that the inclusion of other variables in the model would not affect the estimation of the model. In three different orthogonal models containing $\mathbf{x}_2$ and different other variables as predictors, all three of the estimates for b_2' would be identical.

The correlations among the variables $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ cause differences in the estimates for b_2' reported in Table 5.2. The reason that the first two estimates are similar is that student-teacher ratio and per capita income are not highly correlated ($r = -0.157$). This implies that the corresponding centered vectors for these variables, $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$, form a 99° angle in individual space. Geometrically, we can see that they are fairly close to orthogonal. On the other hand, the per capita income and the percentage of students eligible for free or reduced lunch are more highly correlated variables with $r = -0.563$, as one might expect. The angle formed between the (centered) vectors $\mathbf{x}_{c2}$ and $\mathbf{x}_{c3}$ in individual space is 124°.

The vector diagrams of the three model subspaces in Figure 5.3 illustrate these ideas clearly. The Pythagorean theorem guarantees that whenever the predictors are mutually orthogonal, the sum of the squared lengths of the error vector and each projection of $\mathbf{y}$ onto the subspaces



(a) The vector $\mathbf{y}_c$ and model space $\mathcal{V}_{\mathbf{X}_{c1}}$.

(b) The vector $\mathbf{y}_c$ and model space $\mathcal{V}_{\mathbf{X}_{c3}}$.

Figure 5.3: The vector diagrams of $\mathcal{V}_{\mathbf{X}_{c1}}$ and $\mathcal{V}_{\mathbf{X}_{c3}}$ suggest why the value of the coefficient b_2 varies between models.

corresponding to individual predictors is equal to the squared length of the observation vector. When predictors are nearly orthogonal, then this additivity and independence are to some extent maintained. The inclusion of the teacher-student ratio variable did not significantly change the estimate of b_2'. Whenever the predictors are not orthogonal, the additivity property fails and individual projections are no longer the same as the contribution of a predictor to the overall model. For this reason, interpreting the regression coefficients for variables in models of observational data are difficult; the inclusion or exclusion of a correlated variable can have large consequences for the estimation of these coefficients. The safer approach is to use the whole model rather than attempting to interpret the regression coefficients. This is especially appropriate when the model is being used for prediction rather than explanation.

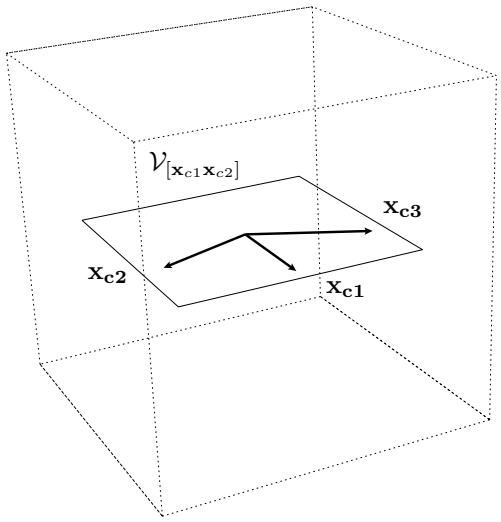
At the other extreme from orthogonality lies *collinearity* or *multicollinearity*, the feature of linear dependence among a set of predictors. Multicollinearity is easy to define (there exists a linear combination of $\mathbf{x}_i s$ equal to zero) and collinearity is the 2-dimensional analog. It is often easy to identify and fix. In studies using observational data, multicollinearity often means that

the analyst has inadvertently included redundant variables such as the sub-scores and the total score on an instrument. More problematic is *near multicollinearity* in which the design matrix $\mathbf{X}$ is not singular, yet the analysis results in unstable (and therefore likely misleading) conclusions.

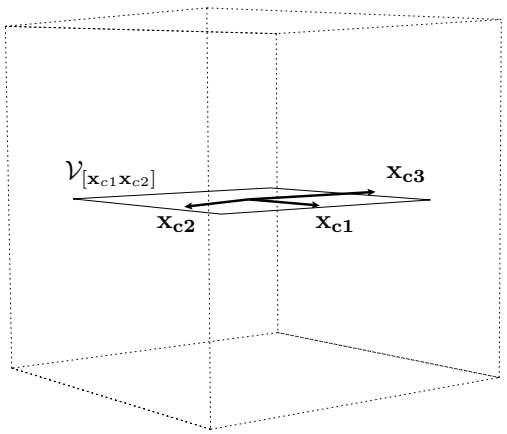
Multicollinearity is fundamentally a geometric feature of the set of predictors and is best understood through representations in individual space. Considering each predictor as a vector in individual space, we know from linear algebra that a linearly independent set of p-vectors spans a p-dimensional space. In a set of predictors that is nearly collinear, there is at least one vector $\mathbf{x}_i$ that is close to the subspace spanned by the remaining predictors in the sense that the angle between $\mathbf{x}_i$ and the projection of $\mathbf{x}_i$ into the space spanned by the rest of the vectors is small. In practice, this suggests that we can detect near multicollinearity by regressing $\mathbf{x}_i$ onto the set of the rest of the predictors for each i and checking for good fit. If the fit is good, then including $\mathbf{x}_i$ in the set of predictors $\mathbf{x}_j$, $j \neq i$, may not be justified because it likely adds little new information.

It is important to note that high pairwise correlation is indicative of near multicollinearity but that multicollinearity is possible even if all of the variables are only moderately pairwise correlated. As we will see, the MCAS variables we examine in this chapter are not related in this way, but it is not hard to create a hypothetical data set that has high multicollinearity but in which no pair of predictors are highly correlated. Let $\mathbf{x}_{c1} = (1, 0, -0.5, -0.5)^T$, $\mathbf{x}_{c2} = (0, 1, -0.5, -0.5)^T$, and $\mathbf{x}_{c3} = (1, -1, 0.05, -.05)^T$. Then all three of pairwise correlations are moderate: $r_{1,2} = 0.333$, $r_{2,3} = -0.577$, and $r_{1,3} = 0.577$. However, $\mathbf{x}_{c3}$ is very close to the span of $\mathbf{x}_{c1}$ and $\mathbf{x}_{c2}$. For example, the vector $\mathbf{v} = \mathbf{x}_{c1} - \mathbf{x}_{c2} = (1, -1, 0, 0)^T$ is very close to $\mathbf{x}_{c3}$; their correlation is almost 1: $r_{\mathbf{v},\mathbf{x}_{c3}} = 0.99875$. (See Figure 5.4.)

Another geometric way to think about near multicollinearity is to consider the parallelepiped defined by the set of normalized and centered predictor vectors. Let $\mathbf{u}_i = \frac{\mathbf{x}_{ci}}{|\mathbf{x}_{ci}|}$; then the generalized volume of the parallelepiped defined by the set of vectors $\{\mathbf{u}_i : 0 < i \leq n\}$ is given



(a) The vectors $\mathbf{x_{c1}}$, $\mathbf{x_{c2}}$, and $\mathbf{x_{c3}}$ are moderately pairwise correlated.



(b) The vector $\mathbf{x}_{c3}$ is very close to $\mathcal{V}_{[\mathbf{x}_{c1}\mathbf{x}_{c2}]}$.

Figure 5.4: The vectors $\mathbf{x_{c1}}$, $\mathbf{x_{c2}}$, and $\mathbf{x_{c3}}$ are moderately pairwise correlated but nearly collinear.

by

$$\left(\begin{vmatrix} \mathbf{u}_1 \cdot \mathbf{u}_1 & \mathbf{u}_1 \cdot \mathbf{u}_2 & \cdots & \mathbf{u}_1 \cdot \mathbf{u}_n \\ \mathbf{u}_2 \cdot \mathbf{u}_1 & \mathbf{u}_2 \cdot \mathbf{u}_2 & \cdots & \mathbf{u}_2 \cdot \mathbf{u}_n \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{u}_n \cdot \mathbf{u}_1 & \mathbf{u}_n \cdot \mathbf{u}_2 & \cdots & \mathbf{u}_n \cdot \mathbf{u}_n \end{vmatrix} \right)^{1/2}.$$

If $n = 1$, then this is just the length of $\mathbf{u}_1$ (which is always 1) and when $n = 2$, the generalized volume is simply the area of a parallelogram with vertices at the the origin and $\mathbf{u}_1,\ \mathbf{u}_2,\ \mathbf{u}_1 + \mathbf{u}_2$ (see Figure 5.5). When $\mathbf{u}_1$ and $\mathbf{u}_2$ are orthogonal, the area of the parallelogram is 1. As these vectors approach multicollinearity, the area approaches 0. The same relationship holds when we extend to higher dimensions. If the predictors are orthogonal, then the generalized volume of the associated parallelepiped is 1. A generalized volume that is close to 0 is evidence of near multicollinearity.

If one suspects near multicollinearity in a set of p predictors, multiple regression can be used to identify which vectors in the set are problematic. One merely uses regression p times

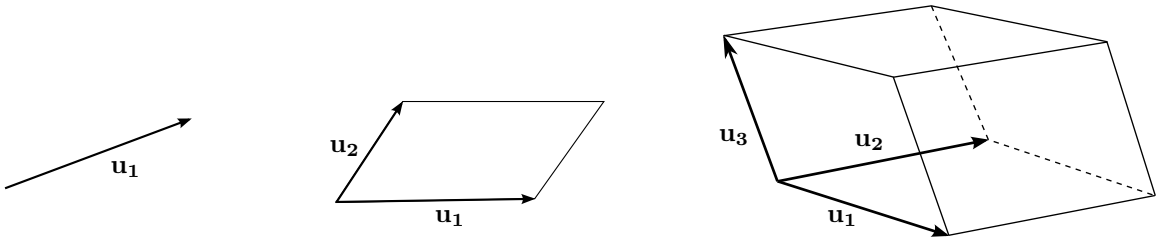


Figure 5.5: The generalized volume of the parallelepiped formed by the set of vectors $\{\mathbf{u}_i : 0 < i \leq n\}$ is equivalent to length in one dimension, area in two dimensions, and volume in three dimensions.

and in each regression analysis, the set $\{\mathbf{x}_i : i \neq k, 1 \leq k \leq p\}$ is used to predict $\mathbf{x}_k$. If the angle between $\mathbf{x}_k$ and $\hat{\mathbf{x}}_k$ (which is in the span of $\{\mathbf{x}_i : i \neq k\}$) is small, then we know that $\mathbf{x}_k$ likely adds little new information to the set of predictors that excludes $\mathbf{x}_k$. Considerations about theoretical importance of variables should be used in choosing an appropriate subset of the original vectors to be used for predicting the dependent variable.

The spending per pupil and the spending per 'regular' pupil (not including occupational, bilingual, or special needs students) are two variables in the MCAS data set that provide a good example of near multicollinearity. As we might expect, these variables are highly correlated ($r = 0.966$) and comparing the models for the student-teacher ratio using each predictor alone and the model with both predictors illustrates why multicollinearity is problematic.

We consider the student-teacher ratio $\mathbf{x}_2$ as an independent variable and denote it $\mathbf{y}'$. Then let $\mathbf{x}_{reg}$ indicate the observed spending per regular pupil in each district and $\mathbf{x}_{all}$ the observed spending per pupil. The three models can then be expressed $\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{reg}] + \mathbf{e}$, $\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{all}] + \mathbf{e}$, and $\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{all}\ \mathbf{x}_{reg}] + \mathbf{e}$. The estimates for b_{reg} and b_{all} in these models is summarized in Table 5.3. All three models explain about a quarter of the variation in the student-teacher ratio (R^2) and the overall fit is significant (p-value > 0.05) . However, there is not much increase in explanatory power in the model with both predictors over the models with just one predictor, especially over the model using spending per regular student. The changes in the regression coefficients are also noteworthy. Both spending per pupil and spending per regular pupil on their own contribute negatively to the student-teacher ratio. (This makes sense because low

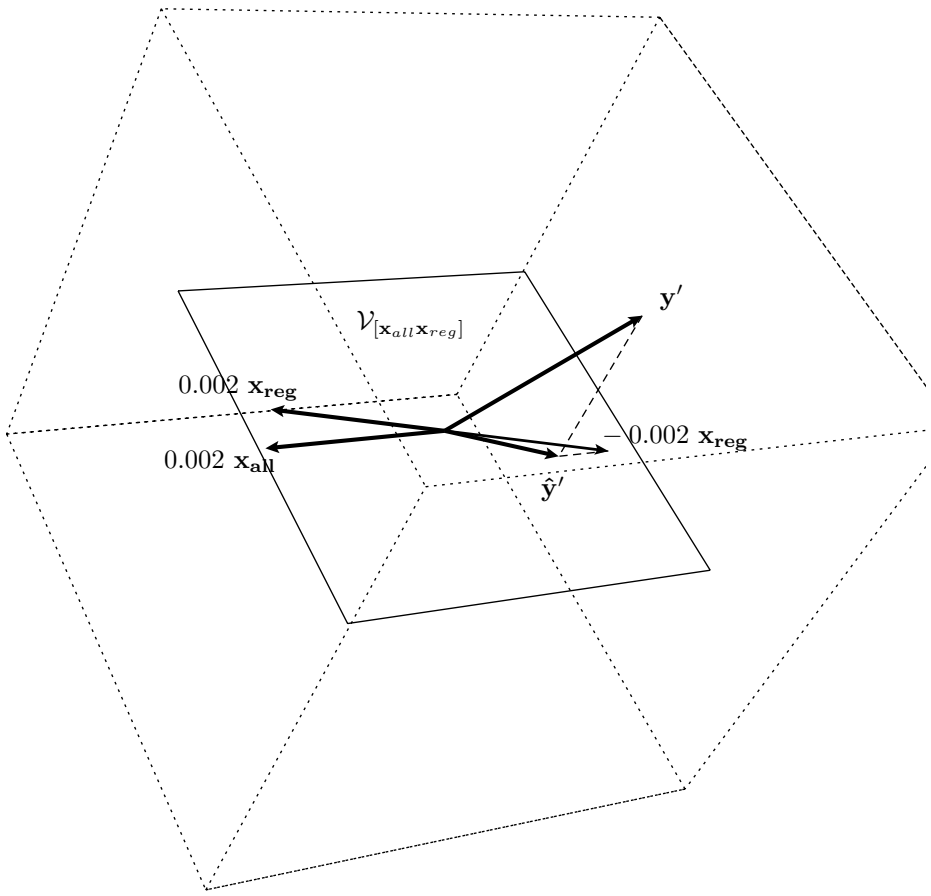


Figure 5.6: The linear combination of $\mathbf{x_{reg}}$ and $\mathbf{x_{all}}$ equal to $\hat{\mathbf{y}}'$ (the projection of $\mathbf{y}'$ into $\mathcal{V}_{[\mathbf{x}_{all}\mathbf{x}_{reg}]}$) must include a term with a positive sign.

student-teacher ratios cost more per pupil.) However, in the third model the magnitude of the coefficient of pupil spending is halved and that of regular student spending is doubled. Even more unexpected is the change in the sign for b_{all}: in the third model this coefficient is *positive*, implying that as spending per pupil increases the student-teacher ratio increases. This coefficient is not statistically significant; it has a p-value greater than 0.05.

These results might seem paradoxical—Why would two good predictors of the student-teacher ratio not be even better when used together? The geometry of individual space makes the reason obvious. The vectors $\mathbf{x}_{all}$ and $\mathbf{x}_{reg}$ point in almost the same direction. Since the

Table 5.3: The effect of multicollinearity on the stability of regression coefficients.

Model	b_{all} (p-value)	b_{reg} (p-value)	R^2 (p-value)
$\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{all}] + \mathbf{e}$	-0.0011 (0.000)	n/a	0.2289 (0.000)
$\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{reg}] + \mathbf{e}$	n/a	-0.0013 (0.000)	0.2643 (0.000)
$\mathbf{y}' = [\mathbf{1}\ \mathbf{x}_{all}\ \mathbf{x}_{reg}] + \mathbf{e}$	0.0006 (0.241)	-0.0020 (0.001)	0.2656 (0.000)

correlation between these vectors is 0.966, the angle between them is only 15° . To move sufficiently in the direction orthogonal to this (so as to reach the projection of $\mathbf{y}$ in the model space) requires a multiple of one of the predictors that goes so far in the first direction that it must be corrected with a predictor with the wrong sign (see Figure 5.6).

We saw that including predictors that are highly correlated with each other is counterproductive even when each is a good predictor of the criterion variable. It is perhaps paradoxical that including predictors that are nearly orthogonal to the criterion variable (and hence very poor predictors of criterion variables) can actually improve the prediction considerably. Such a variable is called a *suppressor variable* and the reason that suppressor variables function as they do is easily explained using the geometry of individual space.

Let $\mathbf{y}$ denote the observed percentages of students eligible for free or reduced lunch in each school district and $\mathbf{x}_{percap}$ denote the observed average per capita income in each district. The total spending per pupil $\mathbf{x}_{all}$ has a very low correlation with the percentage of students eligible for free or reduced lunch ($r = 0.07$) and explains only 0.04% of the variation. However, when it is added to the model using $\mathbf{x}_{percap}$ to predict $\mathbf{y}$, a much better prediction is achieved. The two models $\mathbf{y} = [\mathbf{1}\ \mathbf{x}_{percap}] + \mathbf{e}$ is contrasted with the model $\mathbf{y} = [\mathbf{1}\ \mathbf{x}_{percap}\ \mathbf{x}_{all}] + \mathbf{e}$ in Table 5.4.

What is striking about this example is that $\mathbf{x}_{all}$ on its own predicts essentially nothing of the percentage of students eligible for free or reduced lunch. However, adding it to the model that uses $\mathbf{x}_{percap}$ significantly improves the prediction. From the geometry, we can see that the plane spanned by $\mathbf{x}_{percap}$ and $\mathbf{x}_{all}$ is much closer to $\mathbf{y}$ than the line generated by $\mathbf{x}_{percap}$ alone. Given such a plane, any vector in the plane together with the first vector provides an improved

Table 5.4: Suppressor variables increase the predictive power of a model although they themselves are uncorrelated with the criterion.

Model	b_{percap} (p-value)	b_{all} (p-value)	R^2 (p-value)
$\mathbf{y} = [\mathbf{1}\ \mathbf{x}_{percap}] + \mathbf{e}$	-1.4591 (0.000)	n/a	0.3135 (0.000)
$\mathbf{y} = [\mathbf{1}\ \mathbf{x}_{all}] + \mathbf{e}$	n/a	0.0011 (0.298)	0.0004 (0.298)
$\mathbf{y} = [\mathbf{1}\ \mathbf{x}_{percap}\ \mathbf{x}_{all}] + \mathbf{e}$	-1.8038 (0.000)	0.0053 (0.000)	0.4101 (0.000)

prediction of the criterion, even when the new vector happens to be orthogonal to the criterion variable.

5.4 Partial and semi-partial correlation

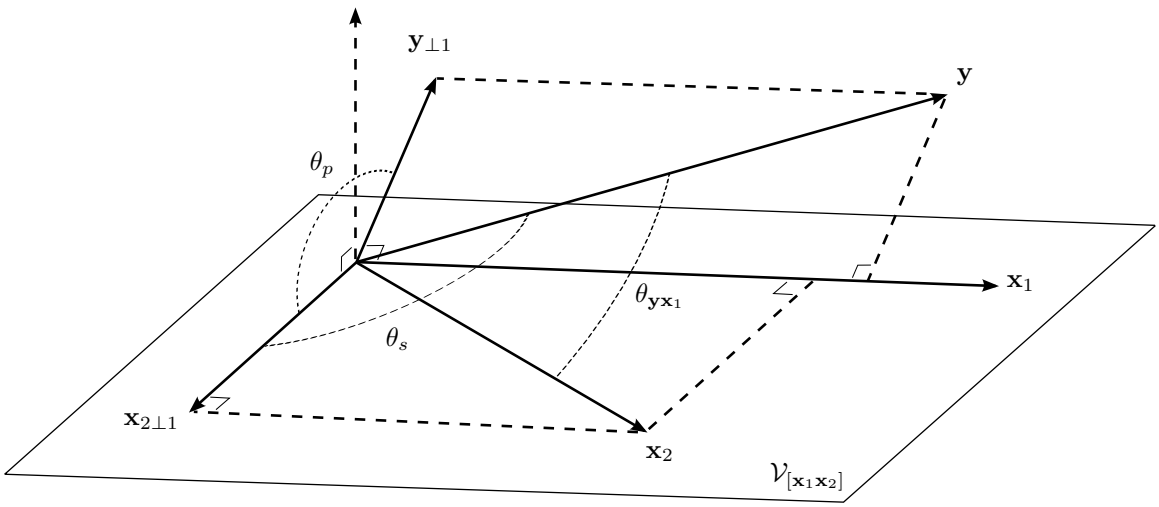


Figure 5.7: The arcs in the vector diagram indicate angles for three kinds of correlation between $\mathbf{y}$ and $\mathbf{x}_2$: the angle θ_p corresponds to the partial correlation conditioning for $\mathbf{x}_1$; the angle θ_s corresponds to the semi-partial correlation with $\mathbf{x}_1$ after controlling for $\mathbf{x}_2$, and the angle $\theta_{\mathbf{yx}_1}$ corresponds to Pearson's correlation, $r_{\mathbf{yx}_1}$.

In regression analyses involving two or more predictors, it is often useful to examine the relationship between a criterion variable and one of the predictors after taking into account the relationship of these variables with the other predictors in the model. For example, suppose the correlation between children's height and intelligence is found to be quite high and one

is tempted by the dubious hypothesis that tall people are more intelligent. By using a third variable, age, which is also known to be correlated with both height and intelligence, we would like to examine the correlation between height and intelligence after taking into consideration the age of the participants. This is a simple example of *statistical control* and is described as *conditioning* for the effect of some variable(s). It is motivated by the idea of experimental control in which randomness removes all of the differences between the treatment and control groups except the variables being studied. The statistic that encodes the correlation between two variables while conditioning for others is called the *partial correlation coefficient*.

In order for conditioning to be a valid procedure, we must check the implicit assumptions about the causal relationship among the variables involved. In particular, by conditioning we assume that the correlation between the conditioning variable(s) and each of the two variables to be correlated is entirely due to a presumed causal relationship by which the conditioning variable(s) affects each of the variables to be correlated. Returning to our example, we assume that the correlations between age and height and between age and performance are entirely due to the causal process of maturation; one expects that as children get older they also grow taller and become more intelligent. We would likely find in this hypothetical example that after accounting for the ages of children the remaining association between height and intelligence would be quite small and most probably due to chance rather than any true relationship. In this way, conditioning can be used to identify cases of so-called *spurious correlation* in which two variables are highly correlated only because they are both correlated to a third variable, often a common cause. However, care must be taken with analyses of partial correlation because in cases where the assumption of causality is not warranted, the partial correlation coefficients have little if any meaning.

The partial correlation between the variables $\mathbf{x}_1$ and $\mathbf{x}_2$, conditioning for the variable $\mathbf{x}_3$ is written $r_{\mathbf{x}_1\mathbf{x}_2.\mathbf{x}_3}$ or simply $r_{12.3}$. Partial correlation is best understood using the geometry of individual space. Given the predictors $\mathbf{x}_1$ and $\mathbf{x}_2$, and the predictor $\mathbf{x}_3$ (whose effects we are controlling for), partial correlation of $\mathbf{x}_1$ and $\mathbf{x}_2$ controlled for $\mathbf{x}_3$ is simply the correlation of $\mathbf{x}_{1\perp 3} = \mathbf{x}_1 - \text{proj}_{\mathbf{x}_3}\mathbf{x}_2$ and $\mathbf{x}_{2\perp 3} = \mathbf{x}_2 - \text{proj}_{\mathbf{x}_3}\mathbf{x}_1$. (Note that we extend this notation for centered vectors in the following way: $\mathbf{x}_{1c\perp 3} = \mathbf{x}_{1c} - \text{proj}_{\mathbf{x}_3}\mathbf{x}_{1c}$.)

Thus we have the following definition which depends on intuition from individual-space geometry:

$$r_{12.3} = \cos(\theta_{\mathbf{x}_{1\perp 3c}\mathbf{x}_{2\perp 3c}}) = \frac{\mathbf{x}_{1\perp 3c} \cdot \mathbf{x}_{2\perp 3c}}{|\mathbf{x}_{1\perp 3c}||\mathbf{x}_{2\perp 3c}|}$$

that has the same relationship to geometry as correlation: the cosine of the angle between two centered, normalized vectors. Partial correlation is more often defined

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}.$$

It is straightforward to show that these definitions are equivalent. To simplify notation, in the following computations we take all vectors as the corresponding centered vector. Using the first definition, it follows that

$$\begin{aligned}
r_{12.3} &= \frac{(\mathbf{x}_1 - \text{proj}_{\mathbf{x}_3}\mathbf{x}_1) \cdot (\mathbf{x}_2 - \text{proj}_{\mathbf{x}_3}\mathbf{x}_2)}{|\mathbf{x}_{1\perp 3}||\mathbf{x}_{2\perp 3}|} \\
&= \frac{\mathbf{x}_1 \cdot \mathbf{x}_2 - \frac{\mathbf{x}_3 \cdot \mathbf{x}_1}{\|\mathbf{x}_3\|}(\mathbf{x}_2 \cdot \mathbf{x}_3) - \frac{\mathbf{x}_3 \cdot \mathbf{x}_2}{\|\mathbf{x}_3\|}(\mathbf{x}_1 \cdot \mathbf{x}_3) + \frac{\mathbf{x}_3 \cdot \mathbf{x}_1}{\|\mathbf{x}_3\|}\frac{\mathbf{x}_3 \cdot \mathbf{x}_2}{\|\mathbf{x}_3\|}(\mathbf{x}_3 \cdot \mathbf{x}_3)}{|\mathbf{x}_{1\perp 3}||\mathbf{x}_{2\perp 3}|} \\
&= \frac{\mathbf{x}_1 \cdot \mathbf{x}_2 - \frac{(\mathbf{x}_1 \cdot \mathbf{x}_3)(\mathbf{x}_2 \cdot \mathbf{x}_3)}{\|\mathbf{x}_3\|^2}}{|\mathbf{x}_{1\perp 3}||\mathbf{x}_{2\perp 3}|} \\
&= \frac{r_{12} - r_{13}r_{23}}{\sqrt{1 - r_{13}^2}\sqrt{1 - r_{23}^2}}
\end{aligned}$$

since we have $\mathbf{x}_i \cdot \mathbf{x}_j = \|\mathbf{x}_i\|\|\mathbf{x}_j\|r_{ij}$ from the definition of correlation and $\|\mathbf{x}_{i\perp j}\|^2 = (1 - r_{ij}^2)\|\mathbf{x}_i\|^2$ by the Pythagorean theorem. We conclude that the standard definition for partial correlation can be derived from the definition inspired by geometric intuition of individual space.

The vector diagram in Figure 5.7 illustrates the geometric interpretation of correlation and partial correlation in individual space. We also observe that correlation, the angle between two unconditioned vectors, can be substantially different from partial correlation, the angle between vectors conditioned on a common set of variables.

The third notable angle in Figure 5.7 is θ_s, which is the angle between $\mathbf{y}$ and the projection of $\mathbf{y}$ on the component of $\mathbf{x}_2$ that is perpendicular to $\mathbf{x}_1$. The corresponding correlation is called the *semipartial correlation* and is used for measuring the unique contribution of $\mathbf{x}_2$ to the prediction $\hat{\mathbf{y}}$ over and above that contribution of $\mathbf{x}_1$.

Semipartial correlations play an important role in statistical decisions about whether or not to include a predictor or set of predictors in a model. In general, if there are two sets of predictors (say the column vectors of the matrices $\mathbf{X}_1$ and $\mathbf{X}_2$) for a criterion variable $\mathbf{y}$, then we can interpret the squared semipartial correlation of the first set ($\mathbf{X}_1$) and $\mathbf{y}$ as the amount of variation in $\mathbf{y}$ that is explained by the subspace of the model space orthogonal to the conditioning space, $C(\mathbf{X}_2)$. Figure 5.7 shows the geometry when these spaces are each spanned by a single vector.

The ideas here are similar to the orthogonal decomposition of the model space we saw in the case of ANOVA analyses. However, it is rarely the case that $C(\mathbf{X}_1)$, for example, is orthogonal to the space $C(\mathbf{X}_2)$, so instead we consider the orthogonal complement of the conditioning space *in the model space*, $C(\mathbf{X}_2)^{\perp} \cap C(\mathbf{X}_1) \oplus C(\mathbf{X}_2)$. With partial correlation, $C(\mathbf{X}_2)^{\perp}$ is compared to only the portion of $\mathbf{y}$ that is perpendicular to the conditioning set, whereas with semipartial correlation, a comparison is made to the whole of $\mathbf{y}$, including any portion within the span of the conditioning set.

Because of the additivity provided by orthogonality of the subspaces of the model space, it is clear that semipartial correlation corresponds to a decomposition of the variability of $\mathbf{y}$. Consider the model

$$\mathbf{y} = [\ \mathbf{X}_1\ \mathbf{X}_2\ ] \cdot \mathbf{b} + \mathbf{e}.$$

Taking the dot product of each side of the equation with itself and applying the fact that the error space and the model space are orthogonal, we obtain

$$|\mathbf{y}|^2 = \left|[\ \mathbf{X}_1\ \mathbf{X}_2\ ]\mathbf{b}\right|^2 + |\mathbf{e}|^2.$$

or

$$|\mathbf{y}|^2 = |\hat{\mathbf{y}}_{\mathbf{X}_1\mathbf{X}_2}|^2 + |\mathbf{e}|^2. \tag{5.1}$$

It can be written

$$\mathbf{y} = [\ \mathbf{X}_{1\perp 2}\ \mathbf{X}_2\ ]\mathbf{b} + \mathbf{e},$$

where $\mathbf{X}_{1\perp 2}$ is a matrix such that $C(\mathbf{X}_{1\perp 2}) = C(\mathbf{X}_2)^{\perp} \cap C(\mathbf{X}_1) \oplus C(\mathbf{X}_2)$. Then we can write

$$\mathbf{y} = \mathbf{X}_{1\perp 2}\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 + \mathbf{e}.$$

Taking the dot product of each side of the equation with itself (and invoking orthogonality) we get that

$$|\mathbf{y}|^2 = |\mathbf{X}_{1\perp 2}\mathbf{b}_1|^2 + |\mathbf{X}_2\mathbf{b}_2|^2 + |\mathbf{e}|^2,$$

which can also be written

$$|\mathbf{y}|^2 = |\hat{\mathbf{y}}_{\mathbf{X}_{1\perp 2}}|^2 + |\hat{\mathbf{y}}_{\mathbf{X}_2}|^2 + |\mathbf{e}|^2, \tag{5.2}$$

where $\hat{\mathbf{y}}_{\mathbf{X}_2}$, for example, indicates the projection of $\mathbf{y}$ onto $C(\mathbf{X}_2)$. Comparing equation 5.1 and equation 5.2 shows that

$$|\hat{\mathbf{y}}_{\mathbf{X}_{1\perp 2}}|^2 = |\hat{\mathbf{y}}_{\mathbf{X}_1\mathbf{X}_2}|^2 - |\hat{\mathbf{y}}_{\mathbf{X}_2}|^2.$$

It follows easily that the squared semipartial correlation can be written in terms of the squared multiple correlation coefficient of the full regression model, $\mathbf{y} = [\ \mathbf{X}_{1\perp 2}\ \mathbf{X}_2\ ]\mathbf{b}+\mathbf{e}$ and the reduced regression model, $\mathbf{y} = \mathbf{X}_2\mathbf{b} + \mathbf{e}$. We have

$$R^2_{y.1(2)} = \frac{|\hat{\mathbf{y}}_{\mathbf{X}_{1\perp 2}}|^2}{|\mathbf{y}|^2} = \frac{|\hat{\mathbf{y}}_{\mathbf{X}_1\mathbf{X}_2}|^2}{|\mathbf{y}|^2} - \frac{|\hat{\mathbf{y}}_{\mathbf{X}_2}|^2}{|\mathbf{y}|^2} = R_{y.12} - R_{y.2}.$$

This equation explains why squared semipartial correlation is often interpreted as the importance of a predictor or set of predictors; it is the increase in the explanatory power of the new model over and above the conditioning model.

Chapter 6

Probability Distributions

The statistical techniques of the preceding chapters rely on the four probability distributions discussed in this chapter. Probability distributions are families of functions indexed by parameters. These parameters specify a distribution function when they are fixed to particular values. As we have seen, it is common to assume a distribution family (in most examples we have assumed that variables follow a normal distribution), and then use what is known about the distribution to estimate the putative parameter(s) which will fix the distribution function that so it agrees with the sample data. For example, we use observation vector to estimate the mean and standard deviation of the dependent variable in ANOVA and regression models.

We saw that Chebyshev's inequality (see Section 1.3.4) is useful because it makes no assumptions about the distribution of a random variable. However, if something is known *a priori* about the distribution of a variable, then bounds for the probability of extreme events can often be significantly improved over the estimates provided by Chebyshev's inequality. Knowing (or assuming) the distribution function of a random variable gives a great deal of information about the expected values of the variable.

6.1 The normal distribution

All of the methods of analysis subsumed by the general linear model assume that the random variables Y_i and the random variables of error E_i have distributions that can be roughly approx-

imated by the *normal distribution*. The normal distribution is denoted $N(\mu, \sigma^2)$ because it is completely determined by the parameters μ and σ^2. If the random variable Y follows a normal distribution with mean μ_Y and variance σ_Y^2, then we write $Y \sim N(\mu_Y, \sigma_Y^2)$, and the probability density function for Y is given by

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2}. \tag{6.1}$$

The other distributions used in analyzing the general linear model (such as the F-distribution used in hypothesis testing) can be derived from the normal distribution. A more detailed description of the relationships among the distributions discussed in this section can be found in many standard statistical texts (e.g., Casella & Berger, 2002; Searle, 1971).

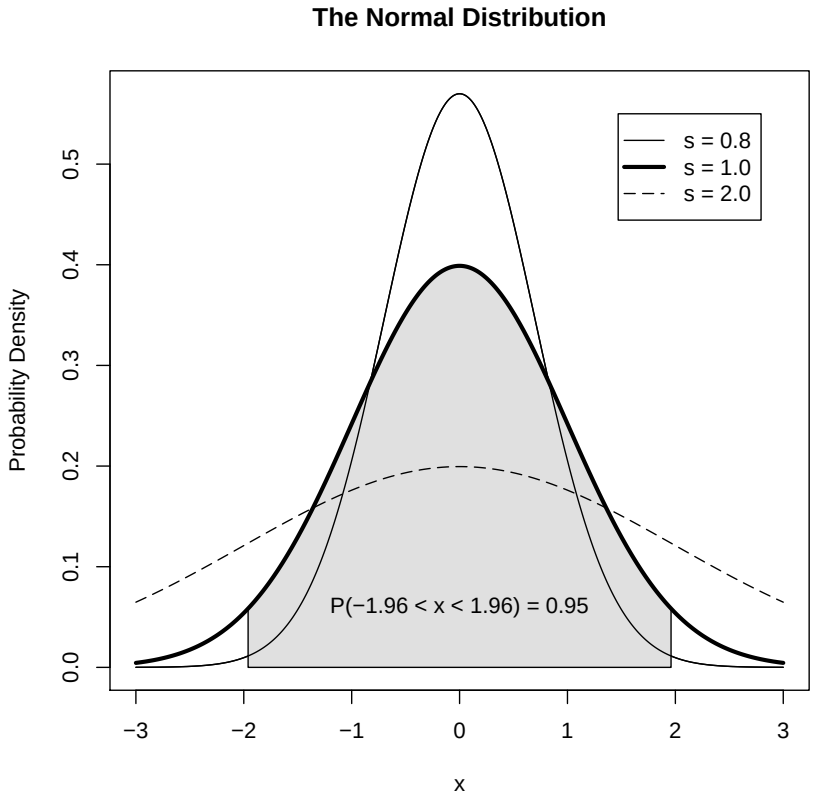


Figure 6.1: The normal distribution with three different standard deviations.

The normal distribution is often used in statistical analyses for a number of reasons. The

best motivation, perhaps, is provided by the Central Limit Theorem, which states that the distribution of the sample mean $\overline{Y}$, approaches a normal distribution as the size of the sample increases, no matter what the distribution of the random variable Y. This is very useful because it allows one to estimate distribution of the sample mean although very little is known about the distribution of the underlying random variables. The normal distribution is useful as the limit of other probability distributions such as the binomial distribution and the Student's t-distribution. The binomial probability distribution gives the probability of obtaining $0 \leq k \leq n$ successful outcomes in n trials when the probability of a success on a single trial is $0 \leq p \leq 1$. The mean of this distribution is np and the variance is $np(1-p)$. When both np and $n(1-p)$ are sufficiently large (in general, it is recommended that both be at least 5), the normal distribution with $\mu = np$ and $\sigma^2 = np(1-p)$ provides a very good continuous approximation for the discrete binomial distribution.

Knowing that a variable follows the normal probability distribution function provides a much stronger result than Chebyshev's inequality (see Section 1.3.4). The key point is that the information about the distribution of Y allows us to make much better approximations concerning the probability of extreme observations of Y. To be explicit, an observation of Y falls within 1 standard deviation of the mean with an approximate probability of 68% because $\int_{-\sigma}^{\sigma} f_Y(y)dy \approx 0.68$, and within 2 standard deviations of the mean with an approximate probability of 95% because $\int_{-2\sigma}^{2\sigma} f_Y(y)dy \approx 0.95$ (see Figure 6.1). (If the goal is to find bounds containing 95% of the area, then ± 1.96 standard deviations provide more accuracy.) Thus, the probability that an observation is more extreme than 2 standard deviations from the mean is only 5%, a significantly tighter bound than that of 25% provided by Chebyshev's inequality (which is the best bound available for an unknown distribution).

6.2 The χ^2 distribution

The χ^2 *distribution* is not used directly in the statistical analyses discussed in this text; however, it is an important distribution for many other kinds of statistical analyses because it can often be used to find the probability of the observed deviation of data from expectation. The χ^2

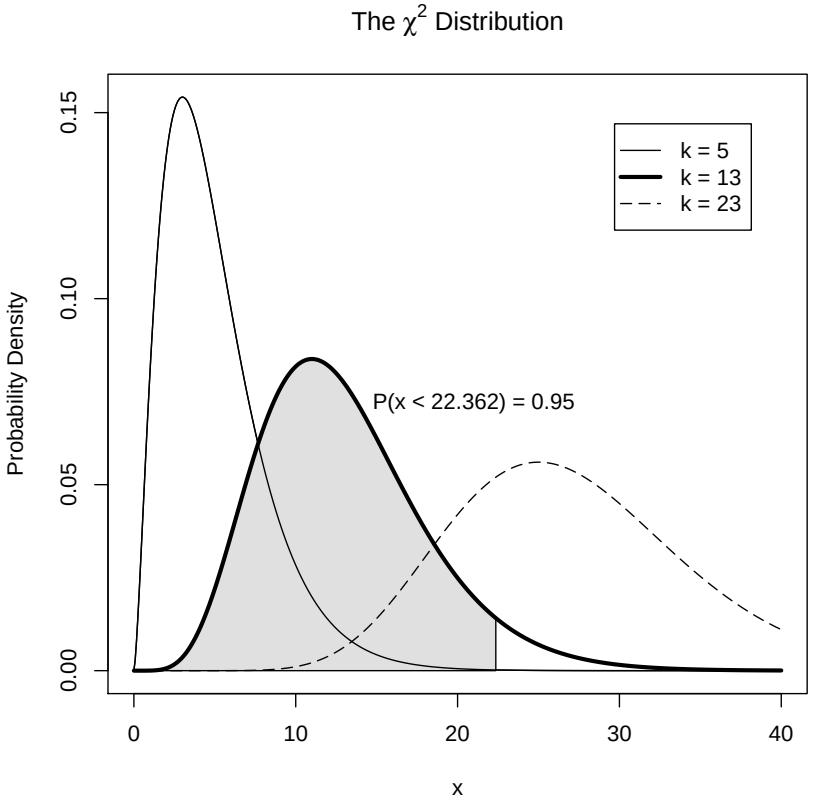


Figure 6.2: The χ^2 distribution with three different degrees of freedom.

distribution is denoted χ^2_k where k is the only parameter of the distribution, a posititive integer called the *degrees of freedom*. If a random variable V follows a χ^2 distribution with k degrees of freedom, then we write $V \sim \chi^2_k$, and the probability density function for V is given by

$$f_V(v) = \frac{1}{2^{k/2}\Gamma(k/2)} v^{k/2-1} e^{-v/2}, \tag{6.2}$$

where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$. You may notice that the mean of the χ^2 distribution for k degrees of freedom is k (see Figure 6.2).

The primary reason for mentioning the χ^2 distribution is that the squared length of a random vector $\mathbf{Y}$ in $\mathbb{R}^k$ (appropriately scaled and with 0 expectation for each coordinate) will follow a χ^2 distribution with k degrees of freedom. The proof follows from two facts:

Lemma 6.2.1. [1]

- *If* $Y \sim N(0,1)$ *is a random variable, then* $Y^2 \sim \chi^2_1$.
- *If* $W_1, \cdots, W_n$ *are independent random variables and* $W_i \sim \chi^2_{k_i}$ *for all* i*, then* $\sum W_i \sim \chi^2_{\Sigma k_i}$.

Because $\mathbf{Y}$ is a random variable in $\mathbb{R}^k$, we can write $\mathbf{Y} = (Y_1, \cdots, Y_k)^T$. Suppose that the coordinate variables Y_i are independent and that each $Y_i \sim N(0, \sigma^2)$. It follows that if we scale each coordinate variable by $\frac{1}{\sigma}$, denoting the result Y_i', then we have $Y_i' = \frac{1}{\sigma} Y_i \sim N(0,1)$. Applying the first part of Lemma 6.2.1, we conclude $(Y_i')^2 \sim \chi^2_1$. Applying the second part of Lemma 6.2.1, we obtain $\sum (Y_i')^2 \sim \chi^2_k$. Moreover,

$$\begin{aligned} \sum (Y_i')^2 &= \frac{1}{\sigma^2} \sum Y_i^2 \\ &= \frac{1}{\sigma^2} \mathbf{Y} \cdot \mathbf{Y} \\ &= \frac{1}{\sigma^2} \|\mathbf{Y}\|^2. \quad \square \end{aligned}$$

This proves

Theorem 6.2.2. *Let* $\mathbf{Y}$ *be a random vector in* $\mathbb{R}^k$ *with independent coordinate variables each distributed* $N(0, \sigma^2)$*. Then* $\frac{1}{\sigma^2} \|\mathbf{Y}\|^2 \sim \chi^2_k$.

Theorem 6.2.2 shows that the degrees-of-freedom parameter corresponds to the dimension of the space containing the vector $\mathbf{Y}$. As we will see, this theorem justifies the claim that the F-ratio (see equation 2.15) follows an F-distribution.

6.3 The F-distribution

Snedecor's F-distribution has a central role in testing hypotheses related to the general linear model, and it is named after the statistician R. A. Fisher. The F-distribution has two param-

[1] For proof see Casella & Berger (2002).

eters, often denoted p and q and called the numerator and denominator degrees of freedom, and is denoted $F(p, q)$. Both parameters affect the shape of the distribution, but in general, as the maximum degrees-of-freedom parameter increases (there are two), the distribution becomes more centered around 1 (see Figure 6.3). If a random variable W follows an F-distribution with p and q degrees of freedom, we write $W \sim F(p, q)$, and the probability density function for W is:

$$f_W(w) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)w}\left(\sqrt{\frac{(pw)^p q^q}{(pw+q)^{p+q}}}\right). \tag{6.3}$$

Variables that follow an F-distribution have a close relationship to variables that follow the χ^2 distribution. Whenever the variables U and V follow χ^2 distributions with with p and q degrees of freedom, respectively, then the (adjusted) ratio of these variables follows an F-distribution with p and q degrees of freedom. That is, if $U \sim \chi^2_p$ and $V \sim \chi^2_q$, then

$$F = \frac{U}{V} \cdot \frac{q}{p} = \frac{U/p}{V/q} \sim F(p, q). \tag{6.4}$$

This equation provides a more general description of the F-ratio (see equation 2.15). Notice that the adjustment factor $\frac{q}{p}$ corrects for the relative degrees of freedom in the two variables that follow χ^2 distributions. If U and V are the squared length of random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively, then we can understand the adjustment factor geometrically as a correction for the dimensions of these vectors. In this way, the F-ratio can be understood as a ratio of *per-dimension squared lengths.*

In Section 2.3, we saw that the vector $\hat{\mathbf{y}}$ is in the $(p+1)$-dimensional subspace of $\mathbb{R}^n$ that is spanned by the $p+1$ independent column vectors of the design matrix $\mathbf{X}$. In a similar way, the vector $\mathbf{e}$ is in the $(n-p-1)$-dimensional space which is the orthogonal complement of $C(\mathbf{X})$. The discussion above uses Theorem 6.2.2 to link the more general statement of F-ratio provided in equation (6.4) with the F-ratio we will use for testing hypotheses about the general linear model.

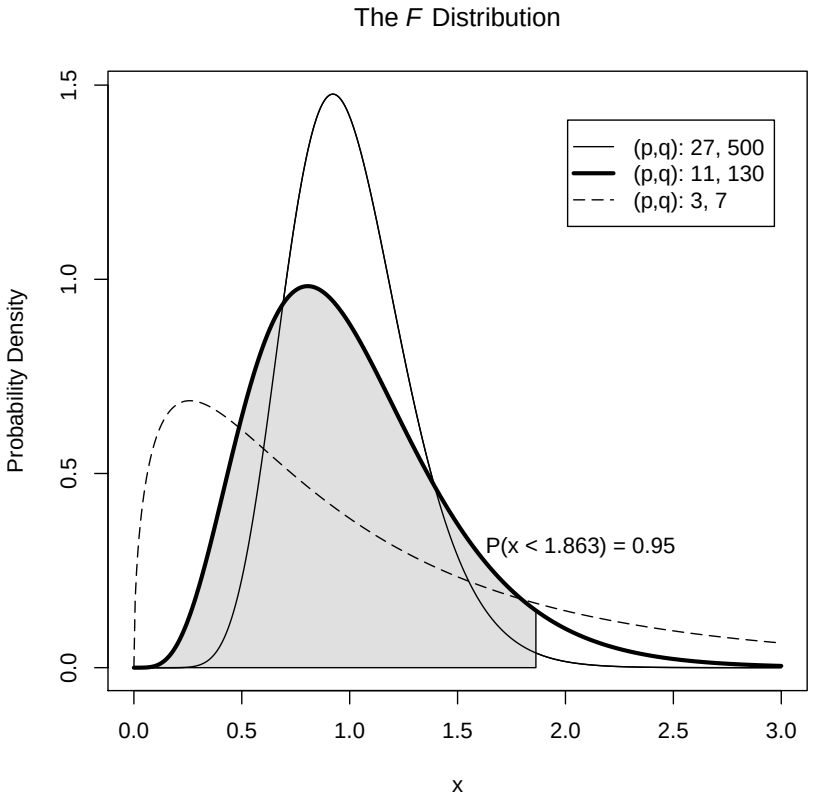


Figure 6.3: The F-distribution centers around 1 as the maximum degrees-of-freedom parameter becomes large.

6.4 Student's t-distribution

We briefly consider one other distribution that is often used in analyses based on the general linear model. This eponymous distribution is called the *Student's t-distribution* after the pen name of William Gosset, a statistician employed by the Guinness Brewery early in the twentieth century. The t-distribution is often applied to hypothesis tests involving small samples. In the context of the general linear model, the t-distribution is helpful for obtaining *confidence intervals* for the parameter estimates in the vector $\mathbf{b}$. Confidence intervals are an alternative to hypothesis testing, and provide a range of likely values for parameter estimates.

The t-distribution has one parameter k, which is the number of degrees of freedom, and

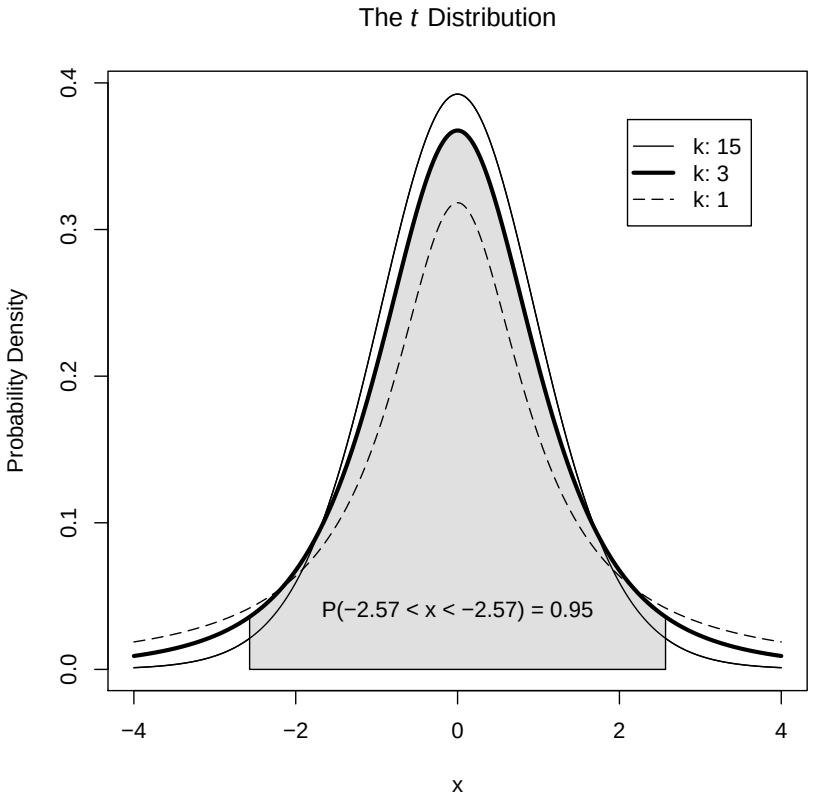


Figure 6.4: The t-distribution approaches the normal distribution as the degrees-of-freedom parameter increases.

the distribution is denoted $t(k)$. If U is a random variable that follows a t-distribution with k degrees of freedom, then we write $U \sim t(k)$ and the probability density function for U is:

$$f_U(u) = \frac{\Gamma(\frac{k+1}{2})}{\sqrt{k\pi}\Gamma(\frac{k}{2})} \left(1 + \frac{u^2}{k}\right)^{-(k+1)/2} \tag{6.5}$$

As k increases, the t-distribution approaches the normal distribution with a mean of 0 and variance of 1 (see Figure 6.4). In addition, there is a close relationship between the t-distribution and the F-distribution. This relationship is straightforward: a squared t statistic follows an F-distribution with 1 degree of freedom in the numerator. That is, if $U \sim t(k)$ then $U^2 \sim F(1, k)$.

This relationship has an intriguing geometric implication for the t-distribution. We can think

of U^2 as a ratio of the (per-dimension) lengths of two vectors. Since the numerator has 1 degree of freedom, the vector in the numerator, V_n, lives in a 1-dimensional subspace. The vector in the denominator, V_d, of the ratio lives in a k-dimensional subspace. We can choose an orthonormal basis $\{\mathbf{u}_i\}$ for the combined $k+1$ dimensional space so that $V_n = c\mathbf{u}_1$ and $V_d = \sum_{i=1}^{k} c_i \mathbf{u}_i$. The variable U^2 can then be expressed

$$U^2 = \frac{\|c\|^2}{\sum_{i=1}^{k} \|c_i\|^2/k},$$

and it follows that U can be expressed

$$U = \frac{\|c\|}{\sum_{i=1}^{k} \|c_i\|/\sqrt{k}} = \frac{\|V_n\|}{\|V_d\|/\sqrt{k}}. \tag{6.6}$$

We saw this expression of a variable that follows the t-distribution in the Section 1.4.

In summary, we can say that the sum of squares of normally distributed variables follows the χ^2 distribution. (Geometrically, this is the squared length of a vector.) In addition, a ratio of variables that follow the χ^2 distribution itself follows an F-distribution. Finally, the square of a variable that follows the t-distribution, itself follows an F-distribution.

Chapter 7

Manipulating Space: Homogeneous Coordinates, Perspective Projections, and the Graphics Pipeline

This chapter describes the mathematical basis of the DataVectors program. Points in $\mathbb{R}^3$ can be represented using *homogeneous coordinates* which facilitate affine translations of these points via matrix multiplication. Perspective projections of $\mathbb{R}^3$ to an arbitrary plane can also be realized via matrix multiplication directly from Euclidean or homogeneous coordinates. However, it is more convenient for computer systems producing computer-generated perspective drawings of 3-dimensional objects to instead transform the *viewing frustum* in $\mathbb{R}^3$ to an appropriately scaled parallelepiped in *perspective space*, retaining some information about relative depth.

In order to illustrate and explore more concretely the ideas discussed in this manuscript and to generate precise figures from data, I wrote the DataVectors program in the R language. This programming language has built-in support for statistical analysis (including matrix arithmetic routines) and celebrated graphical capabilities. Because it is open source and free, the language is used widely by academics and research scientists for developing new statistical techniques.

The DataVectors program accepts up to three data vectors of any length and displays a 3-dimensional model of the (centered) model space that can be rotated in any direction using a mouse. This chapter describes the mathematical basis of the program, including homogeneous coordinates, perspective projections, and the graphical pipeline.

7.1 Homogeneous coordinates for points in $\mathbb{R}^n$

The set of $n \times n$ matrices is denoted M_n. Linear transformations of $\mathbb{R}^n$ can be realized as left-multiplication by invertible elements of M_n. This subset of M_n forms a group under matrix multiplication of called the general linear group which is denoted GL_n. Affine transformations of $\mathbb{R}^n$ cannot be represented by matrix multiplication since $\mathbf{0}$ is a fixed by matrix multiplication. However, the *homogenization* of inhomogeneous equations by introducing another variable with the constant term as its coefficient provides a solution. The group GL_{n+1} includes every linear transformation that fixes one dimension (i.e., those matrices that correspond to linear transformations of $\mathbb{R}^n$). Moreover, the group GL_{n+1} contains matrices corresponding to shear transformations of $\mathbb{R}^{n+1}$ that can be used to achieve affine transformations of $\mathbb{R}^n$ viewed as $\mathbb{R}^{n+1}$ under an appropriate equivalence class. We begin with a definition.

Definition 1 (Homogeneous Coordinates)**.** Given $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$, the vector $\mathbf{X} = (X_1, \ldots, X_n, X_{n+1}) \in \mathbb{R}^{n+1}$ contains homogeneous coordinates for $\mathbf{x}$ if $x_i = \frac{X_i}{X_{n+1}}$ for all i.

Whenever $X_{n+1} = 0$, the coordinates represent a point at infinity but this case is not needed for the present discussion. When $X_{n+1} \neq 0$, the definition establishes an equivalence relation: $\mathbf{X} \equiv \mathbf{Y}$ if and only if there exists a non-zero c in $\mathbb{R}$ such that $\mathbf{Y} = c(X_1, \ldots, X_{n+1})$. We take $(X_1, \ldots, X_n, 1)$ as the canonical representative of the equivalence class of $\mathbf{X}$.

A shear transformation of $\mathbb{R}^{n+1}$ is a linear transformation that fixes a subspace $W \subset \mathbb{R}^{n+1}$, translating each points along a vector $\mathbf{a}$ parallel to W and proportionally to the distance between the point and W. To obtain a translation of R^{n+1}, pick W to be the hyper-plane $(x_{n+1} = 0)$ and $\mathbf{a}$ to be the desired vector of translation in R^n. Then $x_{n+1} = 1$ is the natural embedding of $\mathbb{R}^n$ in $\mathbb{R}^{n+1}$, where each point is mapped to its canonical homogeneous coordinate. This hyperplane is effectively translated by the vector $\mathbf{a}$ by means of the shear transformation of $\mathbb{R}^{n+1}$. Translated

coordinates in $\mathbb{R}^n$ can be recovered by means of the inverse embedding restricted to the range of the embedding. In other words, if $\mathbf{v} \in \mathbb{R}^n$ and $\mathbf{V} = (\mathbf{v}, 1)^T$ is the corresponding homogeneous embedding, we can recover the translated vector $\mathbf{v}'$ from the translated homogeneous vector $\mathbf{V}' \in \mathbb{R}^n$ by stripping off the last coordinate of $\mathbf{V}'$. The matrix representation of this translation is

$$ {}_hT_a = \left[\begin{array}{ccc|c} & & & \\ & I_n & & \mathbf{a} \\ & & & \\ \hline 0 & \ldots & 0 & 1 \end{array}\right], \qquad {}_hT_a \in M(n+1), \tag{7.1} $$

where I_n is the identity in GL_n. The subscript ${}_h$ denotes transformations of homogeneous coordinates and distinguishes these from transformations of $\mathbb{R}^n$ described in the rest of this chapter.

The extension of the linear transformations of rotations and dilations from Euclidean coordinates to homogeneous coordinates is straightforward. Although these results hold in greater generality, we restrict the discussion to transformations of homogeneous coordinates for points in $\mathbb{R}^3$. One common approach to rotations in $\mathbb{R}^3$ is the *yaw-pitch-roll* system. An arbitrary rotation is described (non-uniquely) as a succession of rotations by the angles γ, ϕ, and θ around the z-axis, y-axis, and x-axis, respectively. A rotation is denoted by the matrix $R_{\theta\phi\gamma}$, and the rotation of a vector $\mathbf{v}$ in a right-handed coordinate system is obtained by left-multiplying by the rotation matrix: $\mathbf{v}_{\text{rot}} = R_{\theta\phi\gamma}\mathbf{v}$. We have:

$$ R_{\theta\phi\gamma} = R_x(\theta)R_y(\phi)R_z(\gamma), \text{where} $$

$$ R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{bmatrix}, $$

$$R_y(\phi) = \begin{bmatrix} \cos(\theta) & 0 & \sin(\theta) \\ 0 & 1 & 0 \\ -\sin(\theta) & 0 & \cos(\theta) \end{bmatrix},$$

$$R_z(\gamma) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Rotation matrices can be easily realized in $\mathbb{R}^4$ by adjoining an extra row and column. Thus, $R_{\theta\phi\gamma}$ corresponds to the following matrix in $GL(4)$.

$${}_hR_{\theta\phi\gamma} = \left[\begin{array}{ccc|c} & & & 0 \\ & R_{\theta\phi\gamma} & & \vdots \\ & & & 0 \\ \hline 0 & \dots & 0 & 1 \end{array}\right] \tag{7.2}$$

The scaling transformation, where each vector coordinate is transformed to some multiple of itself, is similarly handled. Let $S_{k_xk_yk_z}$ denote the scaling transformation that multiplies the x, y, and z coordinates by the constants k_x, k_y, and k_z, respectively. We have

$$S_{k_xk_yk_z} = S_{k_x}S_{k_y}S_{k_z}, \text{where}$$

$$S_{k_x} = \begin{bmatrix} k_x & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$$S_{k_y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & k_y & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

$$S_{k_z} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & k_z \end{bmatrix}.$$

Moreover, S corresponds to the following matrix in $GL(4)$.

$$_hS_{k_xk_yk_z} = \begin{bmatrix} k_x & 0 & 0 & 0 \\ 0 & k_y & 0 & 0 \\ 0 & 0 & k_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{7.3}$$

By left-multiplying the product of translation, rotation, and scaling matrices to the homogeneous coordinates of points in $\mathbb{R}^3$ we can achieve all the transformations of interest for representing perspective drawings of 3-dimensional objects.

7.2 The perspective projection

The major challenge in modeling vector space is creating representations of vectors in 3-dimensional space on a 2-dimensional computer screen. This is made possible by the *perspective projection*, defined with a view plane and a viewpoint $\mathbf{v} = (v_1, v_2, v_3) \in \mathbb{R}^3$ not in the view plane. The possibly affine view plane is defined by $\mathbf{n} \cdot \mathbf{x} = c$ or equivalently as $\mathbf{N} \cdot \mathbf{X} = 0$ where $\mathbf{X}$ is the homogeneous coordinates for $\mathbf{x} \in \mathbb{R}^3$ and $\mathbf{N} = (n_1, n_2, n_3, -c)$. The perspective projection sends a point in space $\mathbf{x}$ to the intersection of the view plane with the line passing through $\mathbf{v}$ and $\mathbf{x}$.

Theorem 7.2.1. *Given the viewpoint* $\mathbf{V}$ *and the (possibly affine) view plane with normal* $\mathbf{N}$*, the transformation matrix corresponding to the perspective projection is given by* $P_{\mathbf{V},\mathbf{N}} = \mathbf{V}\mathbf{N}^T - (\mathbf{N} \cdot \mathbf{V})I_4$.[1]

Proof. Let $\mathbf{X}$ denote the homogeneous coordinates of a point $\mathbf{x} \in \mathbb{R}^3$ and let $k_1\mathbf{V} + k_2\mathbf{X}$ denote the image of $\mathbf{X}$ under the perspective projection $P_{\mathbf{V},\mathbf{N}}$, for constants $k_1,\ k_2 \in \mathbb{R}$; then k_1 and k_2

[1] This theorem and its proof follow Marsh (1999).

satisfy $k_1(\mathbf{N}\cdot\mathbf{V}) + k_2(\mathbf{N}\cdot\mathbf{X}) = 0$. If $\mathbf{N}\cdot\mathbf{X} = 0$, then

$$\begin{aligned}(\mathbf{V}\mathbf{N}^T - (\mathbf{N}\cdot\mathbf{V})I_4)X &= (\mathbf{N}\cdot\mathbf{X})\mathbf{V} - (\mathbf{N}\cdot\mathbf{V})I_4\mathbf{X} \\ &= -(\mathbf{N}\cdot\mathbf{V})\mathbf{X}\end{aligned}$$

Therefore, $(\mathbf{V}\mathbf{N}^T - (\mathbf{N}\cdot\mathbf{V})I_4)\mathbf{X}$ is a multiple of $\mathbf{X}$, which is precisely what we would expect given the equivalence relation on points expressed with homogeneous coordinates, and we conclude that $P_{\mathbf{V},\mathbf{N}} = \mathbf{V}\mathbf{N}^T - (\mathbf{N}\cdot\mathbf{V})I_4$ in this case. On the other hand, whenever $\mathbf{N}\cdot\mathbf{X} \neq 0$ (and $k_1 \neq 0$), we have $k_2 = -k_1(\mathbf{N}\cdot\mathbf{V})/(\mathbf{N}\cdot\mathbf{X})$ and by substitution obtain

$$\begin{aligned}P_{\mathbf{V},\mathbf{N}}\mathbf{X} &= k_1\mathbf{V} - (k_1(\mathbf{N}\cdot\mathbf{V})/(\mathbf{N}\cdot\mathbf{X}))\mathbf{X} \\ &= (\mathbf{N}\cdot\mathbf{X})\mathbf{V} - (\mathbf{N}\cdot\mathbf{V})\mathbf{X} \\ &= (\mathbf{V}\mathbf{N}^T - (\mathbf{N}\cdot\mathbf{V})I)\mathbf{X}\end{aligned}$$

as required. □

To apply this to the problem of computer graphics, we make the simplifying assumptions that the view plane is the xy-plane (i.e., $\mathbf{n} = (0,0,1,0)^T$) and that the view point is on the z-axis (i.e., $\mathbf{v} = (0,0,k,1)^T$). Under these assumptions, the matrix for the perspective transformation is:

$$P_{\mathbf{v},\mathbf{n}} = \begin{bmatrix} -k & 0 & 0 & 0 \\ 0 & -k & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -k \end{bmatrix} \tag{7.4}$$

Checking, we have:

$$P_{\mathbf{v},\mathbf{n}} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} -kx \\ -ky \\ 0 \\ z-k \end{bmatrix} \equiv \begin{bmatrix} kx/(k-z) \\ ky/(k-z) \\ 0 \\ 1 \end{bmatrix} \tag{7.5}$$

As we expect, the x- and y-coordinates of the image of this projection is proportional to the ratio of the distance from the viewpoint to the viewing plane and the distance of the pre-image from the view point along the z-axis (see Figure 7.1).

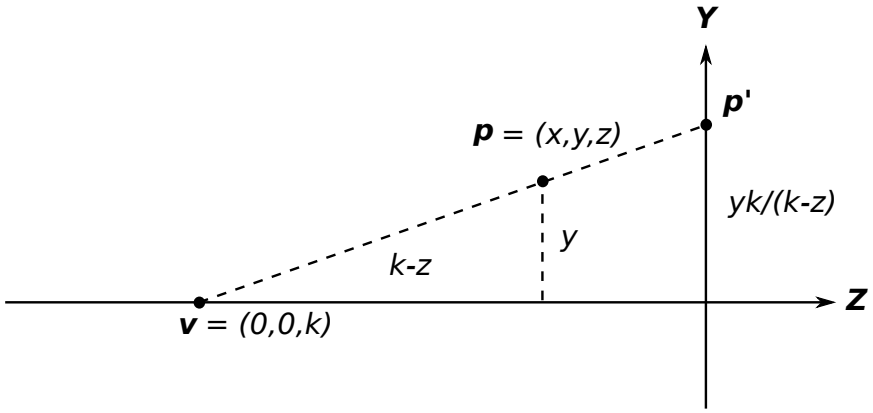


Figure 7.1: The x- and y-coordinates of the perspective projection are proportional to $k/(k-z)$.

7.3 The graphics pipeline

Although it is clear that the perspective projection from a view point on the z-axis to the xy-plane is quite simple, it is perhaps not yet clear why it is always possible to make these simplifying assumptions. As long as the line of sight is orthogonal to the view plane, a solution follows from the transformations developed in the first section. If the view point is not on the z-axis, we left multiply everything by the appropriate rotation matrix to ensure that the view point is on the z-axis. Let $\mathbf{v}'$ be a view point not on the z-axis and R_v a rotation taking $\mathbf{v}'$ to $\mathbf{v} = (0, 0, \|\mathbf{v}'\|)^T$. It follows that the matrix product $P_{\mathbf{v},\mathbf{n}} R_v$ takes any point $\mathbf{p}$ to some $\mathbf{p}'$ on the xy-plane. To recover the image of $P_{\mathbf{v}',\mathbf{n}}$ we simply compute $R_v^{-1}\mathbf{p}'$, where R_v^{-1} is the inverse rotation operator. The *graphics pipeline* refers to a sequence of mathematical operations such

as this that transform a point in $\mathbb{R}^3$ into a pixel on the computer monitor. All transformations are represented using matrix multiplication and the entire pipeline can be conceived as the composition of all the matrix transformations.

The perspective transformation is problematic because it is not of full rank and therefore singular. Information regarding the distance of a point to the center of projection is lost and cannot be recovered from the information that remains in the image. We will find that it is useful to decompose the perspective transformation into translation and dilation transformations followed by an orthogonal projection. We refrain from projecting from 3 dimensions into 2 dimensions until the last step of the pipeline just before the pixel is displayed. In the penultimate step, points have been distorted to achieve perspective but still retain information about relative depth. This space is called *perspective space*. The advantage is that the transformation of Euclidean space into perspective space has an inverse so the depth information used to resolve issues such as object collisions can be recovered. In addition, the depth information aides in drawing realistic effects such as simulated fog, in which the transparency (a color aspect) of a point is proportional to its distance from the center of projection.

Ideally, we want the x- and y-coordinates in perspective space would have the same values as the x and y screen coordinates. This allows the final projection from perspective space to the screen coordinates to be an orthogonal projection in the z direction, and obtaining the screen coordinates would require no further calculations. After rotating the view point to the z-axis and translating it to the origin, we associate the truncated pyramid called the viewing frustum (see Figure 7.2) with the region $[-b_x, b_x] \times [-b_y, b_y] \times [-1, 1]$. The viewing frustum is defined by the near and far clipping planes, $\mathbf{n} = (0, 0, n, 0)^T$ and $\mathbf{f} = (0, 0, f, 0)^T$, and the dimensions of the visible view plane. For a centered, square screen, $b_x = w/2$ and $b_y = h/2$, and so these dimensions are $[-w/2, w/2] \times [-h/2, h/2]$, where w and h are the screen width and height in screen coordinates. This association is achieved via the perspective space transformation:

$$S_{\mathbf{v},\mathbf{n}} = \begin{bmatrix} \frac{2n}{w} & 0 & 0 & 0 \\ 0 & \frac{2n}{h} & 0 & 0 \\ 0 & 0 & \frac{-(f+n)}{f-n} & \frac{-2fn}{f-n} \\ 0 & 0 & -1 & 0 \end{bmatrix} \tag{7.6}$$

The perspective space coordinates of $\mathbf{P} = (x, y, z, 1)^T$ are given by

$$S_{\mathbf{v},\mathbf{n}} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} = \begin{bmatrix} 2n/w \\ 2n/h \\ \frac{-(f+n)}{f-n} z - \frac{2fn}{f-n} \\ -z \end{bmatrix} = \begin{bmatrix} 2n/(-zw) \\ 2n/(-zh) \\ \frac{(f+n)}{f-n} + \frac{2fn}{z(f-n)} \\ 1 \end{bmatrix}$$

By using the 4^{th} row of $S_{\mathbf{v},\mathbf{n}}$ to record the z-coordinate, dividing by this coordinate to obtain the canonical homogeneous coordinates for the point applies the correct perspective scaling to the x- and y-coordinates. It follows that $S_{\mathbf{v},\mathbf{n}}$ can be decomposed further as the product of dilation transformations (in the x- and y-coordinates) that associate the original x and y coordinates in the viewing plane with the desired screen coordinates and an even simpler perspective transformation. The 3^{rd} coordinate of $\mathbf{p}'$ is not suppressed to 0 as under the perspective projection, but retains depth information via the invertible function $z' = \frac{(f+n)}{f-n} + \frac{2fn}{z(f-n)}$.

It is straightforward to verify that this formula maps the viewing frustum to the appropriate parallelepiped in perspective space. Once mapped to perspective space, the screen coordinates can be read off the first two coordinates of $\mathbf{p}'$. If needed, an inverse function can be used to determine the original z-coordinate of any point in perspective space.

In sum, beginning with an arbitrary viewpoint $\mathbf{v}$ and an orthogonal view plane with normal $\mathbf{v}$, the graphics pipeline (1) rotates space around the origin so that $\mathbf{v}$ lies on the z-axis (say, via $R_{\theta\phi\gamma}$). Then, (2) space is translated along the z-axis so that $\mathbf{v}$ lies at the origin, and the image of the origin is $(0, 0, -\|v\|)$, (via T). Next, (3) the space is dilated in order to identify the viewing plane with the computer screen (via D) and (4) the viewing frustum is transformed into a parallelepiped in perspective space (via S). Finally (5), the screen pixel can be drawn

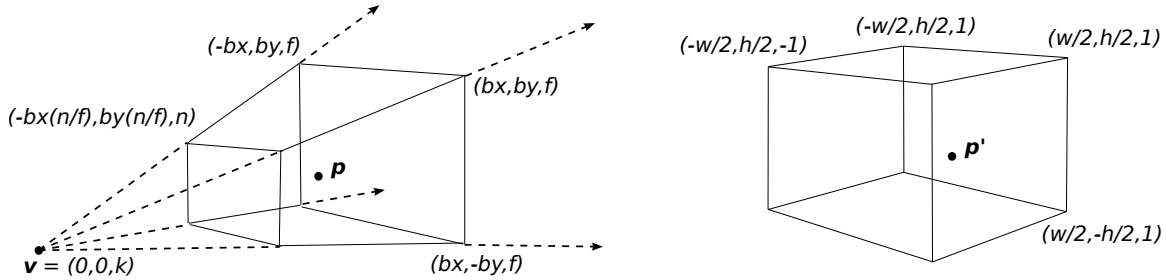


Figure 7.2: The perspective space transformation takes the viewing frustum to the parallelepiped $[-w/2, w/2] \times [-h/2, h/2] \times [-1, 1]$ in perspective space.

using the x-and y-coordinates of each point in perspective space, an orthogonal projection onto the xy-plane in perspective space:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Then the whole pipeline taking $\mathbf{p}$ in the original space to $\mathbf{p}'$ on the screen can be written as a product of matrices:

$$\mathbf{p}' = MSDTR\mathbf{p}.$$

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27* , 17–21.

Bryant, P. (1984). Geometry, statistics, probability: Variations on a common theme. *The American Statistician, 38* , 38–48.

Casella, G. & Berger, R. L. (2002). *Statistical Inference (2nd Ed.).* South Melbourne, Australia: Thompson Learning.

Christensen, R. (1996). *Plane Answers to Complex Questions.* New York, NY: Springer-Verlag.

Cook, T.D. & Campbell, D.T. (1979). *Quasi-Experimentation: Design and Analysis for Field Settings.* Rand McNally, Chicago, Illinois.

Elazar J. Pedhazur (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction (3rd Ed.).* South Melbourne, Australia: Thompson Learning.

Herr, D. G. (1980). On the history of the use of geometry in the general linear model. *The American Statistician, 34*, 43–47.

Faraway, J. J. (2004). *Linear Models with R.* Boca Raton, FL: Chapman & Hall.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models.* New York, NY: McGraw-Hill.

Marsh, D. (1999). *Applied Geometry for Computer Graphics and CAD.* New York, NY: Springer-Verlag.

Pitman, J. (1992). *Probability.* New York, NY: Springer-Verlag.

Rogers, J. L. & Nicewander, W. A., (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*, 59–66.

Saville, D. J. & Wood, G. R. (1986). A method for teaching statistics using n-dimensional geometry. *The American Statistician, 40*, 205–214.

Saville, D. J. & Wood, G. R. (1991). *Statistical Methods: The Geometric Approach.* New York, NY: Springer-Verlag.

Searle, S. R. (1971). *Linear Models.* New York, NY: Wiley.

Shifrin, T., & Adams, M. R. (2002). *Linear Algebra: A Geometric Approach.* New York, NY: Freeman.

Shoemake, K. (1992). ARCBALL: A User Interface for Specifying Three-Dimensional Orientation Using a Mouse. Paper presented at the annual proceedings of Graphics Interface in Vancouver, Canada.

Shoemake, K. (no date). *Quaternions.* Retrieved on January 11, 2011 from www.cs.caltech.edu/courses/cs171/quatut.pdf

Wickens, T. D. (1995). *The Geometry of Multivariate Statistics.* Mahwah, NJ: Erlbaum.