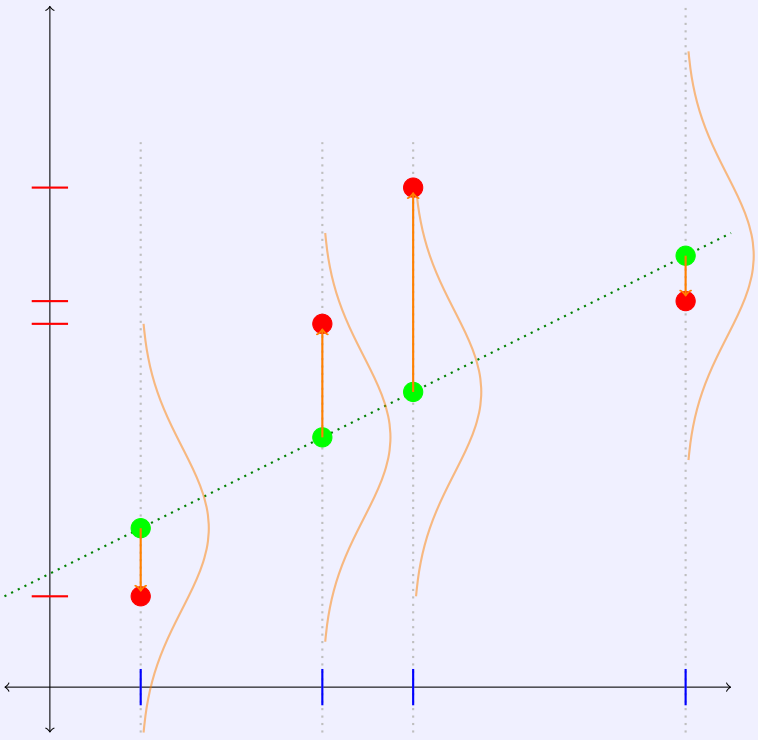
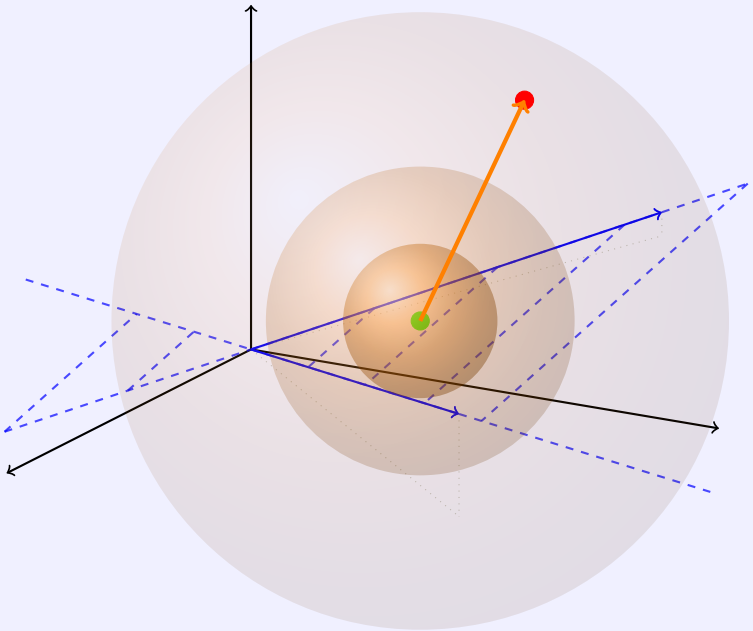
Visualizing Linear Models





W. D. Brinda, PhD

VISUALIZING
LINEAR MODELS

W. D. BRINDA

Contents

PREFACE

0.1. About the book

This book accompanies my Linear Models (S&DS 312/612) course at Yale University. It provides an intuitive and visual approach to the material that is (I hope) accessible to students who are comfortable with linear algebra and with the basics of statistical theory.

My purpose is to develop the student's understanding of the core aspects of linear model theory by practicing two invaluable and complementary ways of visualizing the data and model: the *observations* picture and the *variables* picture. The *observations* picture is natural. The *variables* picture, on the other hand, will seem challenging at first. Read the text carefully. Contemplate the figures. Work through the relevant exercises multiple times. Eventually you'll achieve a mental breakthrough and the pictures will all make sense.

This book has three chapters on linear model theory, each of which is preceded by a "review" chapter covering essential mathematical background. The review chapters are designed to help the reader develop a high level of fluency with the *ingredients* of linear model theory. Because the ingredients of linear model theory are also the ingredients of a broad spectrum of quantitative fields and applications, this aspect of the book is every bit as valuable as its coverage of linear models.

0.2. Chunking and purposeful practice

Imagine that you want to become as good as possible at tennis. How should you spend your time in order to achieve this? One option is to read book after book about playing tennis. You'll learn some useful things, but reading by itself will do very little for your tennis game. Alternatively, you could find a partner and spend all of your time actually playing tennis matches. This will surely make you better, but it won't come close to making you the best player you can be. You'll plateau pretty quickly. A combination of reading and playing would be better than either one by itself, but ultimately you're still missing what is by far the most important activity for improving performance in tennis or any other sport: *practice exercises*. Tennis performance involves numerous *ingredients*, such as forehand, backhand, serving, and so on, and those skills can be further broken down into component parts. For every little component, you can find practice exercises designed to perfect your execution of that action. In particular, you'll do the exact same exercise over and over rapidly, trying to notice and correct imperfections, ideally with the assistance of an expert coach. Here's the key: these practice exercises are literally

building neural circuitry in your brain that will later enable you to execute the action correctly automatically.

Let's return to the original suggestions of reading about tennis and playing tennis in order to clarify how they fit into the process of developing expert performance. You can and should read explanations of how to execute some action properly, but that isn't enough. You then have to spend hours practicing that action in order to actually be able to do it. You should also spend some time playing tennis in order to practice putting the ingredients together. But your overall performance can only be great if the ingredients are great, and playing the game isn't a very effective way to improve the ingredients.

This isn't just about sports, of course, or I wouldn't have bothered discussing it. Performance in intellectual domains works the same way. For instance, it's been shown that chess experts have developed specialized neural circuitry to quickly see important formations of pieces on the board. The mind can only deal with three or four "items" at a time. Experts overcome this limitation by "chunking" multiple ideas together into a single "item" so that they can be processed in a more automated way, leaving the mind extra RAM to work with. Over time, complex items continue to combine with each, and eventually the mind is capable of processing ideas that would be literally unthinkable without all that circuitry. Neurologically speaking, chunking means developing specialized neural circuitry not unlike what's needed for an athlete to automate certain actions. The most effective way to build this circuitry is with practice exercises resembling those in sports.

So what are the characteristics of a practice exercise that's well-designed for building this neural circuitry? Here are four, according to psychologist K. Anders Ericsson who has spent his career studying expert performance:

- well-defined task
- appropriate level of difficulty for the individual
- rapid and informative feedback
- opportunities for repetition and correction of errors.

Dr. Ericsson uses the term "purposeful practice" for exercises that use these principles. Engaging in purposeful practice exercises is the most effective way to restructure your brain for expert performance.

In the tennis discussion, I brought up three different types of activities that can be involved in improving performance in a domain: reading about it, doing it, and engaging in practice exercises. For any advanced math course, students will spend most of their study time reading the textbook and working out problems. The time-consuming problems are often too challenging, feedback comes much later if ever, and they're never repeated; they don't resemble the practice exercises of an athlete. They're much more like playing the sport than practicing it. So purposeful practice, the most effective activity for building the neural circuitry necessary for expert performance, is apparently absent from the process; frankly, it's not even on the radar.

When, as a doctoral student, I read a few books on the study of expert performance, I realized that my own learning process bore little resemblance to purposeful practice. So I decided to change how I spent my time. I identified key skills and knowledge that would likely be essential to doing effective research on my topic, then I made a sizable deck of notecards, each one prompting me to work through very short derivations that used some of my targeted skills or knowledge. I set

aside a block of several hours every day to go through a portion of those notecards. This practice accelerated my research progress more than I expected. In fact, my most important advances came spontaneously *during* those practice sessions. When you're working out a derivation for the tenth (or perhaps fiftieth) time, you've developed such good circuitry for processing it that you have plenty of RAM to spare for new variations, generalizations, or connections that might naturally come to mind.

There's much more I could say about optimizing the learning process, but I won't go into them here. Some of these additional aspects (including "spaced repetition" and "interleaving") are built into learning software such as the Anki flashcard program which you'll use as you go through this class.

Most of the results throughout this book are presented in the form of exercises. This format is designed to help you review and internalize the material as efficiently as possible: every exercise has a corresponding Anki card that can be imported from a file that I'll share with you. The most crucial cards appear in red boxes in the text; I've carefully designed them to emphasize key insights and skills that underlie expertise in linear modeling. The next most important cards are in orange boxes, and less important ones are in green. Finally, a handful of blue exercises are only included in order make the book more complete more self-contained; they can be used for reference but aren't so valuable for the reader to work through. The number of exercises is likely to seem overwhelming. Start with the red cards, and work through them several times. Then work on the orange cards. If you still have any time and energy remaining after that, you can try going through the green ones. A solution manual will be provided for your convenience.

Let me reiterate that this isn't about memorization. You aren't trying to store facts on your hard drive, you're trying to build new circuitry. It's also not about understanding. When a tennis player is satisfied that she knows how to swing backhand correctly, that's not when she stops practicing; on the contrary, that's when she's finally able to start practicing. Likewise, just because you're confident that you really understand a card, doesn't mean you should stop practicing it. Expertise requires more than just memorization and understanding. So just keep working through the cards; trust the process.

At the end of most of my Linear Models class meetings, I administer a very short quiz asking one of the book exercises. Some days I'll specify a subset of cards to focus on; other days, it could be any past red or orange card. At the beginning of the course, students can survive the quizzes by just cramming before class, but soon the number of cards becomes overwhelming for those two haven't been diligently and purposefully practicing!

Each week as homework, I require my students to work through a short coding task to put what they're learning into practice. I don't make these homeworks too time-consuming, because I want to leave plenty of time and energy for notecard practice sessions. All homework files are available at this book's website (quantitations.com/books).

CHAPTER

1

REVIEW: LINEAR ALGEBRA

We'll begin our course of study by reviewing the most relevant definitions and concepts of linear algebra. We'll also expand on various aspects of spectral decomposition and orthogonal projection that aren't necessarily covered in a first linear algebra course.

Vector spaces can be studied in great abstraction, but any vector mentioned in this chapter can be assumed to be in Euclidean space. Similarly, every matrix can be assumed to have real entries.

1.1. Subspaces

Let $\mathcal{S}$ be a subset of Euclidean space $\mathbb{R}^n$. If for every pair $\mathbf{v}_1, \mathbf{v}_2 \in \mathcal{S}$ and every $b_1, b_2 \in \mathbb{R}$, the linear combination $b_1\mathbf{v}_1 + b_2\mathbf{v}_2$ is also in $\mathcal{S}$, then $\mathcal{S}$ is called a **subspace** [of $\mathbb{R}^n$].

The **span** of a set of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\} \subset \mathbb{R}^n$ is the set of all their possible linear combinations[1] $\{b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m : (b_1, \ldots, b_m) \in \mathbb{R}^m\}$.

Notice that the set of possible coefficients of vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is itself the vector space $\mathbb{R}^m$. It can be convenient to realize that taking the $(b_1, \ldots, b_m)$ linear combination of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is the same as multiplying the matrix whose columns are $\mathbf{v}_1, \ldots, \mathbf{v}_m$ by the [column] vector $(b_1, \ldots, b_m) \in \mathbb{R}^m$.

$$\begin{bmatrix} | & & | \\ \mathbf{v}_1 & \ldots & \mathbf{v}_m \\ | & & | \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} = b_1 \begin{bmatrix} | \\ \mathbf{v}_1 \\ | \end{bmatrix} + \ldots + b_m \begin{bmatrix} | \\ \mathbf{v}_m \\ | \end{bmatrix}$$

Exercise 1.1

Show that the span of $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is a subspace.

[1] Although vectors are written out horizontally in this book's paragraph text, e.g. $(b_1, \ldots, b_m)$, they should always be interpreted as column vectors in any mathematical expression.

1.2. Bases

If no one of the vectors in $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is in the span of the others, then the set is called **linearly independent**.

Exercise 1.2

Show that if $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent then $(b_1, \ldots, b_m) = (0, \ldots, 0)$ is the only set of coefficients for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = \mathbf{0}$.

Exercise 1.3

Show that if $(b_1, \ldots, b_m) = (0, \ldots, 0)$ is the only set of coefficients for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = 0$, then $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent.

Given vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$, the subset of the coefficients $\{\mathbf{b} = (b_1, \ldots, b_m) : b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m = \mathbf{0}\} \subseteq \mathbb{R}^m$ that produce linear combinations equal to zero is called the **null space** of $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

Exercise 1.4

Let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be a set of vectors. Prove that their null space is a subspace of $\mathbb{R}^m$.

Exercise 1.5

Suppose that $\mathbf{w} = b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m$, and denote $\mathbf{b} = (b_1, \ldots, b_m)$. Show that for any $\mathbf{z}$ in the null space of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, the coefficient vector $\mathbf{b} + \mathbf{z}$ also produces $\mathbf{w}$.

Exercise 1.6

Suppose that $\mathbf{w} = b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m$, and denote $\mathbf{b} = (b_1, \ldots, b_m)$. Show that every coefficient vector in $\mathbb{R}^m$ that produces a linear combination equal to $\mathbf{w}$ must equal $\mathbf{b} + \mathbf{z}$ for some $\mathbf{z}$ in the null space of $\mathbf{v}_1, \ldots, \mathbf{v}_m$.

Exercise 1.7

Show that $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are linearly independent if and only if every $\mathbf{w}$ in their span has a unique set of coefficients $(b_1, \ldots, b_m)$ for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = \mathbf{w}$.

Let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be a set of vectors whose span is $\mathcal{S}$. Then $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ is called a **basis** for $\mathcal{S}$ if any (and therefore all) of the following equivalent conditions are true.[2]

- They are linearly independent.
- $(0, \ldots, 0) \in \mathbb{R}^m$ is the only vector in their null space.

[2] This equivalence is established by Exercises 1.2, 1.3, and 1.7.

- Every $\mathbf{w} \in \mathcal{S}$ has a unique set of coefficients $(b_1, \ldots, b_m)$ for which $b_1\mathbf{v}_1 + \ldots + b_m\mathbf{v}_m = \mathbf{w}$.

Exercise 1.8

Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ be a basis for $\mathcal{S}$. How do you know that $\{\mathbf{v}_2, \ldots, \mathbf{v}_m\}$ is not a basis for $\mathcal{S}$?

Exercise 1.9

Let $\mathbf{v}_1, \ldots, \mathbf{v}_m$ be a basis for $\mathcal{S}$, and let $\mathcal{T}$ be a *proper* subspace of $\mathcal{S}$ (meaning that there exists at least one vector in $\mathcal{S}$ that isn't in $\mathcal{T}$). Show that at least one of $\mathbf{v}_1, \ldots, \mathbf{v}_m$ is not in $\mathcal{T}$.

❧ ❋ ❧

Every basis for $\mathcal{S}$ has the same number of vectors (Exercise 1.10); that number is called the **dimension** of $\mathcal{S}$.

Exercise 1.10

Prove that every basis for $\mathcal{S}$ has the same number of vectors.

Exercise 1.11

Find the dimension of $\mathbb{R}^n$.

Exercise 1.12

Let $\mathcal{S}$ be an m-dimensional subspace. Prove that any set of m linearly independent vectors in $\mathcal{S}$ must be basis for $\mathcal{S}$.

Exercise 1.13

Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. Prove that a basis for $\mathcal{S}$ exists.

Exercise 1.14

Suppose two matrices $\mathbb{M}_1$ and $\mathbb{M}_2$ have the exact same behavior on a basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$, that is, $\mathbb{M}_1\mathbf{u}_j = \mathbb{M}_2\mathbf{u}_j$ for every $j \in \{1, \ldots, n\}$. Show that $\mathbb{M}_1$ and $\mathbb{M}_2$ must be the same matrix.

❧ ❋ ❧

1.3. Inner product

The **inner product** of two vectors $\mathbf{v} = (v_1, \ldots, v_n) \in \mathbb{R}^n$ and $\mathbf{w} = (w_1, \ldots, w_n) \in \mathbb{R}^n$ is the sum of the products of their coordinates. It's the same as the matrix multiplication $\mathbf{v}^T\mathbf{w}$, but we'll also denote it with angular brackets as

$$\langle \mathbf{v}, \mathbf{w} \rangle := v_1w_1 + \ldots + v_nw_n.$$

The inner product is symmetric in its arguments, meaning that for any vectors $\mathbf{v}$, and $\mathbf{w}$, the order of the arguments doesn't matter: $\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle$. The inner product is also *linear* in its

arguments, meaning that for any vectors $\mathbf{v}$, $\mathbf{w}$, and $\mathbf{y}$ and real numbers a_1 and a_2,

$$\langle a_1\mathbf{v} + a_2\mathbf{w}, \mathbf{y}\rangle = a_1\langle \mathbf{v}, \mathbf{y}\rangle + a_2\langle \mathbf{w}, \mathbf{y}\rangle.$$

Two vectors are called **orthogonal** if their inner product is zero.

Exercise 1.15

Show that if $\mathbf{y}$ is orthogonal to every one of $\mathbf{v}_1, \ldots, \mathbf{v}_m$, then it is orthogonal to every vector in their span.

⋄

Solution: Any $\mathbf{w}$ in the span can be represented as some linear combination of the vectors $\mathbf{w} = b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m$. Using this representation, we can work out the inner product of $\mathbf{y}$ and $\mathbf{w}$,

$$\begin{aligned}\langle \mathbf{y}, \mathbf{w}\rangle &= \langle \mathbf{y}, b_1\mathbf{v}_1 + \ldots b_m\mathbf{v}_m\rangle \\ &= b_1 \underbrace{\langle \mathbf{y}, \mathbf{v}_1\rangle}_{0} + \ldots + b_m \underbrace{\langle \mathbf{y}, \mathbf{v}_m\rangle}_{0} \\ &= 0\end{aligned}$$

because $\mathbf{y}$ is orthogonal to each of the basis vectors.

Exercise 1.16

If $\mathbf{v}_1, \ldots, \mathbf{v}_m$ are nonzero vectors that are all orthogonal to each other, show that they are linearly independent.

❧ ❋ ❧

The **norm** of a vector $\mathbf{v}$ is the square root of the inner product of $\mathbf{v}$ with itself:

$$\|\mathbf{v}\| := \sqrt{\langle \mathbf{v}, \mathbf{v}\rangle}.$$

The *Pythagorean identity* is the statement that if $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal to each other, then

$$\|\mathbf{v}_1 + \mathbf{v}_2\|^2 = \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2$$

which is proven in Exercise 1.17. It is equivalent to think about the vertices at $\mathbf{0}$, $\mathbf{v}_1$, and $\mathbf{v}_1 + \mathbf{v}_2$ forming three sides of a right triangle whose sides are translated versions of $\mathbf{v}_1$, $\mathbf{v}_2$, and $\mathbf{v}_1 + \mathbf{v}_2$. Then the Pythagorean identity is the familiar statement that the squared length of the hypotenuse of a right triangle equals the sum of the squared lengths of the other two sides – see Figure 1.1.

Exercise 1.17

Prove the Pythagorean identity.

⋄

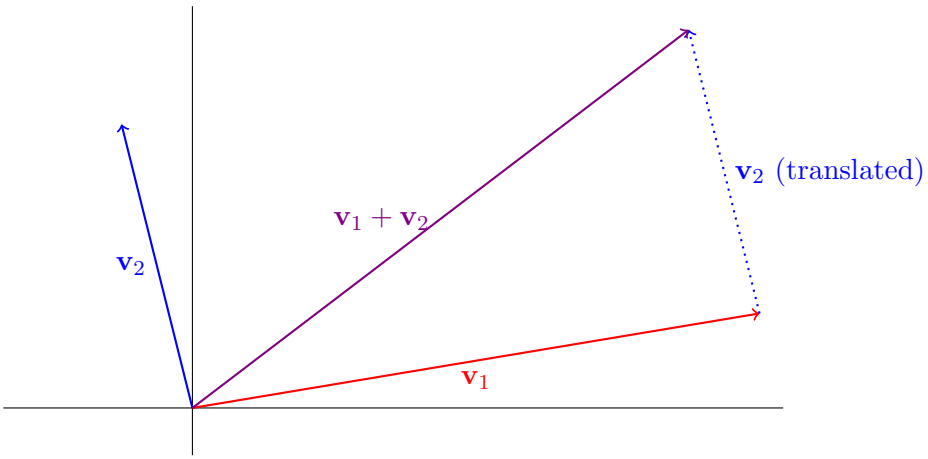


Figure 1.1: If $\mathbf{v}_1$ and $\mathbf{v}_2$ are orthogonal, then they can be arranged as sides of a right triangle whose hypotenuse is $\mathbf{v}_1 + \mathbf{v}_2$.

Solution: The desired result follows readily after distributing the terms.

$$\begin{aligned}\|\mathbf{v}_1 + \mathbf{v}_2\|^2 &= \langle \mathbf{v}_1 + \mathbf{v}_2, \mathbf{v}_1 + \mathbf{v}_2 \rangle \\ &= \langle \mathbf{v}_1, \mathbf{v}_1 \rangle + \underbrace{\langle \mathbf{v}_1, \mathbf{v}_2 \rangle}_{0} + \underbrace{\langle \mathbf{v}_2, \mathbf{v}_1 \rangle}_{0} + \langle \mathbf{v}_2, \mathbf{v}_2 \rangle \\ &= \|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2\end{aligned}$$

Exercise 1.18

Justify the Pythagorean identity extended to m orthogonal vectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$:

$$\|\mathbf{v}_1 + \ldots + \mathbf{v}_m\|^2 = \|\mathbf{v}_1\|^2 + \ldots + \|\mathbf{v}_m\|^2.$$

The term *orthogonal* is often applied more generally than our original definition indicates. A vector is called *orthogonal* to a subspace if it is orthogonal to every vector in that subspace. Two subspaces are called *orthogonal* if every vector in the first subspace is orthogonal to every vector in the second subspace.

1.4. Orthonormal bases

Given any two vectors $\mathbf{v}$ and $\mathbf{y}$ in $\mathbb{R}^n$, there exists a unique representation of $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{v}$ and a vector orthogonal to the span of $\mathbf{v}$. (This fact will be generalized in Section 1.9.)

Exercise 1.19

Given a non-zero vector $\mathbf{v}$, find the unique representation of the vector $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{v}$ and a vector orthogonal to the span of $\mathbf{v}$.

◇

Solution: We'll explicitly construct the desired vector in the span of $\mathbf{v}$. The vector we seek must equal $\hat{b}\mathbf{v}$ for some real $\hat{b}$. Based on the trivial identity $\mathbf{y} = \hat{b}\mathbf{v} + (\mathbf{y} - \hat{b}\mathbf{v})$, we see that we need the other vector $\mathbf{y} - \hat{b}\mathbf{v}$ to be orthogonal to $\mathbf{v}$.

$$\begin{aligned} \langle \mathbf{y} - \hat{b}\mathbf{v}, \mathbf{v} \rangle &= 0 \\ \Updownarrow & \\ \langle \mathbf{y}, \mathbf{v} \rangle - \hat{b}\langle \mathbf{v}, \mathbf{v} \rangle &= 0 \\ \Updownarrow & \\ b &= \frac{\langle \mathbf{y}, \mathbf{v} \rangle}{\|\mathbf{v}\|^2} \end{aligned}$$

Therefore, $\mathbf{y}$ can be represented as the sum of $\frac{\langle \mathbf{y},\mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v}$ which in the span of $\mathbf{v}$ and $(\mathbf{y} - \frac{\langle \mathbf{y},\mathbf{v} \rangle}{\|\mathbf{v}\|^2}\mathbf{v})$ which is orthogonal to the span of $\mathbf{v}$.

Exercise 1.20

Given a vector $\mathbf{v}$, we've derived in Exercise 1.19 the unique representation of $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{v}$ and a vector orthogonal to the span of $\mathbf{v}$. How does the expression simplify when $\mathbf{v}$ is a unit vector?

◇

Solution: When $\mathbf{v}$ has length 1, our expression simplifies to $\langle \mathbf{y}, \mathbf{v} \rangle \mathbf{v} + (\mathbf{y} - \langle \mathbf{y}, \mathbf{v} \rangle \mathbf{v})$.

❧ ❋ ❧

A set of unit vectors that are all orthogonal to each other are called **orthonormal**. If an orthonormal set of vectors is also a basis for $\mathcal{S}$ it is called, naturally, an *orthonormal basis* for $\mathcal{S}$.

Exercise 1.21

If the columns of $\mathbb{U}$ are orthonormal, show that $\mathbb{U}^T\mathbb{U}$ equals the identity matrix $\mathbb{I}$.

Exercise 1.22

If $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are an orthonormal basis for $\mathcal{S}$, find a unique representation of $\mathbf{y} \in \mathcal{S}$ as a linear combination of the basis vectors.

◇

Solution: Let $\mathbf{y} = b_1\mathbf{u}_1 + \ldots + b_m\mathbf{u}_m$; we'll work out the correct coefficients $b_1, \ldots, b_m$. We have already observed that there is a unique representation of $\mathbf{y}$ as the sum of a vector in the span of $\mathbf{u}_1$ plus a vector orthogonal to $\mathbf{u}_1$. This fits our current scenario because the vector $b_1\mathbf{u}_1$ is in the span of $\mathbf{u}_1$ while the vector $b_2\mathbf{u}_2 + \ldots + b_m\mathbf{u}_m$ is orthogonal to the span of $\mathbf{u}_1$. So according to Exercise 1.20, the part of the unique representation in the span of $\mathbf{u}_1$ must be exactly $\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1$, i.e. $b_1 = \langle \mathbf{y}, \mathbf{u}_1 \rangle$. By reasoning similarly for each of the basis vectors, we conclude that $\mathbf{v}$ must have the unique representation

$$\mathbf{y} = \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m.$$

❧ ❋ ❧

Based on Exercise 1.22, any orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ for $\mathbb{R}^n$ can naturally be understood to provide its own coordinate system for $\mathbb{R}^n$ that is an alternative to the standard basis. The coordinates of $\mathbf{y}$ with respect to $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are $(\langle \mathbf{y}, \mathbf{u}_1 \rangle, \ldots, \langle \mathbf{y}, \mathbf{u}_n \rangle)$. More generally, given any unit vector $\mathbf{u}$, we may also refer to $\langle \mathbf{y}, \mathbf{u} \rangle$ as the *coefficient* or *coordinate* of $\mathbf{y}$ with respect to $\mathbf{u}$.

The squared norm of $\mathbf{y}$ is by definition the sum of its squared coordinates with respect to the standard basis; in fact, as Exercise 1.23 verifies, it's also equal to the sum of the squared coordinates with respect to any basis, a result known as *Parseval's identity*.

Exercise 1.23

If $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are an orthonormal basis for $\mathcal{S}$, and $\mathbf{y} \in \mathcal{S}$, use the Pythagorean identity to derive an expression for the squared norm of $\mathbf{y}$ that makes use of the representation from Exercise 1.22:

$$\mathbf{y} = \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m.$$

◇

Solution: By the (extended) Pythagorean identity,

$$\begin{aligned} \|\mathbf{y}\|^2 &= \|\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m\|^2 \\ &= \|\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1\|^2 + \ldots + \|\langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m\|^2 \\ &= \langle \mathbf{y}, \mathbf{u}_1 \rangle^2 \underbrace{\|\mathbf{u}_1\|^2}_{1} + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2 \underbrace{\|\mathbf{u}_m\|^2}_{1} \\ &= \langle \mathbf{y}, \mathbf{u}_1 \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2. \end{aligned}$$

Exercise 1.24

Let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be an orthonormal basis for $\mathcal{S}$, and let $\mathbf{y} \in \mathcal{S}$. Consider the *approximation* $\hat{\mathbf{y}} := \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k$ with $k \leq m$. Find a simple formula for the squared norm of $\mathbf{y} - \hat{\mathbf{y}}$, which we might call the *squared approximation error*.

◇

Solution: One approach is to apply the Pythagorean identity to the right triangle to get $\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$. Solving for $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ then using Parseval's identity,

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 &= \|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2 \\ &= (\langle \mathbf{y}, \mathbf{u}_1 \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2) - (\langle \mathbf{y}, \mathbf{u}_1 \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_k \rangle^2) \\ &= \langle \mathbf{y}, \mathbf{u}_{k+1} \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2. \end{aligned}$$

Alternatively, by first representing $\mathbf{y}$ and $\mathbf{y}_0$ in terms of the orthonormal basis then applying the extended Pythagorean identity,

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 &= \|(\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m) - (\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_k \rangle \mathbf{u}_k)\|^2 \\ &= \|\langle \mathbf{y}, \mathbf{u}_{k+1} \rangle \mathbf{u}_{k+1} + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m\|^2 \\ &= \|\langle \mathbf{y}, \mathbf{u}_{k+1} \rangle \mathbf{u}_{k+1}\|^2 + \ldots + \|\langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m\|^2 \\ &= \langle \mathbf{y}, \mathbf{u}_{k+1} \rangle^2 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle^2. \end{aligned}$$

Exercise 1.25

If $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are an orthonormal basis for $\mathcal{S}$, and $\mathbf{y} \in \mathcal{S}$ explain which term in the representation $\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m$ best approximates $\mathbf{y}$ in the sense that it results in the smallest approximation error $\|\mathbf{y} - \langle \mathbf{y}, \mathbf{u}_j \rangle \mathbf{u}_j\|$.

$\diamond$

Solution: Based on Exercise 1.24, the squared approximation error $\|\mathbf{y} - \langle \mathbf{y}, \mathbf{u}_j \rangle \mathbf{u}_j\|^2$ is equal to the sum of the squares of the other coefficients $\sum_{i \neq j} \langle \mathbf{y}, \mathbf{u}_i \rangle^2$. Therefore, the approximation error is minimized if we use the term with the largest squared coefficient.

Exercise 1.26

If $\mathbf{u}_1, \ldots, \mathbf{u}_m$ are an orthonormal basis for $\mathcal{S}$, and $\mathbf{y}$ is not in $\mathcal{S}$, show that $\mathbf{y} - (\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m)$ is orthogonal to $\mathcal{S}$.

$\diamond$

Solution: By Exercise 1.15, it suffices to show that the vector in question is orthogonal to every basis vector. Consider its inner product with $\mathbf{u}_1$:

$$\begin{aligned}
\langle \mathbf{u}_1, \mathbf{y} - (\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m) \rangle &= \langle \mathbf{u}_1, \mathbf{y} \rangle - (\langle \mathbf{u}_1, \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 \rangle + \ldots + \langle \mathbf{u}_1, \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m \rangle) \\
&= \langle \mathbf{u}_1, \mathbf{y} \rangle - (\langle \mathbf{y}, \mathbf{u}_1 \rangle \underbrace{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}_{1} + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \underbrace{\langle \mathbf{u}_1, \mathbf{u}_m \rangle}_{0}) \\
&= \langle \mathbf{u}_1, \mathbf{y} \rangle - \langle \mathbf{u}_1, \mathbf{y} \rangle \\
&= 0
\end{aligned}$$

and likewise for each of $\mathbf{u}_2, \ldots, \mathbf{u}_m$.

Given any basis $\mathbf{v}_1, \ldots, \mathbf{v}_m$ for a subspace $\mathcal{S} \subseteq \mathbb{R}^n$, the *Gram-Schmidt algorithm* is a process for constructing a an orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_m$ from the original basis vectors. First, normalize the first basis vector $\mathbf{u}_1 := \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$. Then $\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1$ is orthogonal to $\mathbf{u}_1$; dividing by its length creates a vector $\mathbf{u}_2 := \frac{\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1}{\|\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1\|}$ which remains orthogonal to $\mathbf{u}_1$ but has length 1. (Notice that $\mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{u}_1 \rangle \mathbf{u}_1$ is guaranteed to have positive length because it can be represented as a non-zero linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ and therefore cannot be the zero vector.) Repeat this process to define $\mathbf{u}_3 := \frac{\mathbf{v}_3 - (\langle \mathbf{v}_3, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{v}_3, \mathbf{u}_2 \rangle \mathbf{u}_2)}{\|\mathbf{v}_3 - (\langle \mathbf{v}_3, \mathbf{u}_1 \rangle \mathbf{u}_1 + \langle \mathbf{v}_3, \mathbf{u}_2 \rangle \mathbf{u}_2)\|}$ through $\mathbf{u}_m := \frac{\mathbf{v}_m - (\langle \mathbf{v}_m, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{v}_m, \mathbf{u}_{m-1} \rangle \mathbf{u}_{m-1})}{\|\mathbf{v}_m - (\langle \mathbf{v}_m, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{v}_m, \mathbf{u}_{m-1} \rangle \mathbf{u}_{m-1})\|}$ in turn. These m orthonormal vectors must be a basis for $\mathcal{S}$ because it has dimension m.

We know from Exercise 1.13 that a basis for any subspace $\mathcal{S} \subseteq \mathbb{R}^n$ exists; the Gram-Schmidt algorithm provides the stronger fact that an orthonormal basis for $\mathcal{S}$ exists.

1.5. Eigenvalues and eigenvectors

Let $\mathbb{V}$ be an $n \times n$ real matrix. If $\mathbb{V}\mathbf{w} = \lambda \mathbf{w}$ for some $\lambda \in \mathbb{R}$ and some $\mathbf{w} \in \mathbb{R}^n$, then λ is called an **eigenvalue** for $\mathbf{v}$ and $\mathbf{w}$ is called an **eigenvector** for $\mathbf{v}$. The set of all eigenvectors corresponding to the eigenvalue λ forms a subspace called the **eigenspace** for λ.

Exercise 1.27

Let λ be an eigenvalue for $\mathbb{V}$. Show that the set of all eigenvectors corresponding to the eigenvalue λ must form a subspace.

Exercise 1.28

Explain why any matrix that has 0 as an eigenvalue is not invertible.

Exercise 1.29

If a matrix $\mathbb{V}$ has eigenvalues $\lambda_1, \ldots, \lambda_m$ with corresponding eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_m$, identify eigenvalues and eigenvectors of $a\mathbb{V}$ for a non-zero $a \in \mathbb{R}$.

1.6. Spectral decomposition

The **transpose** of a matrix $\mathbb{M}$, denoted $\mathbb{M}^T$ is the matrix whose columns are the rows of $\mathbb{M}$; likewise the rows of $\mathbb{M}^T$ are the columns of $\mathbb{M}$. A matrix is called **symmetric** if it is equal to its transpose. If $\mathbb{V}$ is a symmetric $n \times n$ matrix, then it can be represented by a so-called **spectral decomposition**

$$\mathbb{V} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T$$

where $\lambda_1, \ldots, \lambda_n$ are real numbers, and $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are an orthonormal basis for $\mathbb{R}^n$. Every spectral decomposition of $\mathbb{V}$ has the same values $\lambda_1, \ldots, \lambda_n$ as its n coefficients. (We'll not include a proof for spectral decompositions, but the interested reader can find justifications elsewhere.)

Exercise 1.30

Let $\mathbf{q}_1, \ldots, \mathbf{q}_n$ be orthonormal. Show that if

$$\mathbb{V} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T$$

then $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are eigenvectors of $\mathbb{V}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.

◇

Solution: Without loss of generality, we'll verify that $\mathbb{V}\mathbf{q}_1 = \lambda_1 \mathbf{q}_1$.

$$\begin{aligned}
\mathbb{V}\mathbf{q}_1 &= (\lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T)\mathbf{q}_1 \\
&= \lambda_1 \mathbf{q}_1 \underbrace{\mathbf{q}_1^T \mathbf{q}_1}_{1} + \ldots + \lambda_n \mathbf{q}_n \underbrace{\mathbf{q}_n^T \mathbf{q}_1}_{0} \\
&= \lambda_1 \mathbf{q}_1
\end{aligned}$$

Exercise 1.31

Prove that a matrix $\mathbb{M}$ is symmetric if and only if $\langle \mathbf{v}, \mathbb{M}\mathbf{w} \rangle = \langle \mathbb{M}\mathbf{v}, \mathbf{w} \rangle$ for every $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

The behavior of $\mathbb{V}$ is very easily described with respect to the eigenvector basis: it simply multiplies the jth coordinate by λ_j, as a spectral decomposition reveals.

$$\begin{aligned}
\mathbb{V}\mathbf{y} &= (\lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T)\mathbf{y} \\
&= \lambda_1 \langle \mathbf{y}, \mathbf{q}_1 \rangle \mathbf{q}_1 + \ldots + \lambda_n \langle \mathbf{y}, \mathbf{q}_n \rangle \mathbf{q}_n
\end{aligned} \tag{1.1}$$

We see from this expression that the coordinates of $\mathbb{V}\mathbf{y}$ with respect the eigenvector basis are simply $\lambda_1, \ldots, \lambda_n$ times the corresponding coordinates of $\mathbf{y}$ with respect to that basis.

Exercise 1.32

Let $\mathbb{V} \in \mathbb{R}^{n\times n}$ be a symmetric matrix with non-negative eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding unit eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Show that the symmetric matrix that has eigenvalues $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is the *square root* of $\mathbb{V}$ in the sense that this matrix times itself equals $\mathbb{V}$. (This matrix is often denoted $\mathbb{V}^{1/2}$.)

Using the spectral decomposition, we can also express the product of $\mathbb{V}$ and $\mathbf{y}$ as

$$\begin{aligned}
\mathbb{V}\mathbf{y} &= (\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{y} \\
&= \lambda_1\mathbf{q}_1\mathbf{q}_1^T\mathbf{y} + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T\mathbf{y} \\
&= \begin{bmatrix} | & & | \\ \lambda_1\mathbf{q}_1 & \ldots & \lambda_n\mathbf{q}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \mathbf{q}_1^T\mathbf{y} \\ \vdots \\ \mathbf{q}_n^T\mathbf{y} \end{bmatrix} \\
&= \begin{bmatrix} | & & | \\ \mathbf{q}_1 & \ldots & \mathbf{q}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} - & \mathbf{q}_1 & - \\ & \vdots & \\ - & \mathbf{q}_n & - \end{bmatrix} \mathbf{y}.
\end{aligned}$$

Thus $\mathbb{V}$ can be written as $\mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T$ where $\mathbb{Q}$ is a matrix whose columns are orthonormal eigenvectors and $\mathbb{\Lambda}$ is a diagonal matrix of eigenvalues. We can interpret the actions of these three matrices in turn. The first matrix multiplication results in the vector of coordinates of $\mathbf{y}$ with respect to the eigenvector basis.

$$\begin{aligned}
\mathbb{Q}^T\mathbf{y} &= \begin{bmatrix} - & \mathbf{q}_1 & - \\ & \vdots & \\ - & \mathbf{q}_n & - \end{bmatrix} \mathbf{y} \\
&= \begin{bmatrix} \langle \mathbf{y}, \mathbf{q}_1 \rangle \\ \vdots \\ \langle \mathbf{y}, \mathbf{q}_n \rangle \end{bmatrix}
\end{aligned}$$

Next, the $\mathbb{\Lambda}$ matrix multiplies each coordinate by the appropriate eigenvalue.

$$\begin{aligned}
\mathbb{\Lambda}\mathbb{Q}^T\mathbf{y} &= \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix} \begin{bmatrix} \langle \mathbf{y}, \mathbf{q}_1 \rangle \\ \vdots \\ \langle \mathbf{y}, \mathbf{q}_n \rangle \end{bmatrix} \\
&= \begin{bmatrix} \lambda_1\langle \mathbf{y}, \mathbf{q}_1 \rangle \\ \vdots \\ \lambda_n\langle \mathbf{y}, \mathbf{q}_n \rangle \end{bmatrix}
\end{aligned}$$

Finally, multiplication by $\mathbb{Q}$ provides the linear combination of the eigenvector basis using those

coordinates as the coefficients.

$$\begin{aligned}\mathbb{Q}\mathbb{\Lambda}\mathbb{Q}^T\mathbf{y} &= \begin{bmatrix} | & & | \\ \mathbf{q}_1 & \dots & \mathbf{q}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \lambda_1\langle \mathbf{y}, \mathbf{q}_1\rangle \\ \vdots \\ \lambda_n\langle \mathbf{y}, \mathbf{q}_n\rangle \end{bmatrix} \\ &= \lambda_1\langle \mathbf{y}, \mathbf{q}_1\rangle \begin{bmatrix} | \\ \mathbf{q}_1 \\ | \end{bmatrix} + \ldots + \lambda_n\langle \mathbf{y}, \mathbf{q}_n\rangle \begin{bmatrix} | \\ \mathbf{q}_n \\ | \end{bmatrix}\end{aligned}$$

This is exactly what we observed in Equation 1.1.

Exercise 1.33

Let $\mathbb{M}$ be an $n \times m$ matrix and $\mathbb{V}$ be an $m \times n$ matrix. Show that the trace (sum of diagonals) of $\mathbb{M}\mathbb{V}$ is equal to the trace of $\mathbb{V}\mathbb{M}$.

Exercise 1.34

If $\mathbb{V}$ is a symmetric matrix, show that the trace of $\mathbb{V}$ equals the sum of its eigenvalues $\lambda_1, \ldots, \lambda_n$. (This fact holds for every square matrix, but we won't worry about the more general proof.)

Exercise 1.35

Let $\mathbb{V}$ be a symmetric $n \times n$ matrix that has non-zero eigenvalues $\lambda_1, \ldots, \lambda_n$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$. Show that the matrix that has eigenvalues $\frac{1}{\lambda_1}, \ldots, \frac{1}{\lambda_n}$ with eigenvectors $\mathbf{q}_1, \ldots, \mathbf{q}_n$ is the inverse of $\mathbb{V}$.

Exercise 1.36

Let $\mathbb{M}$ and $\mathbb{V}$ be matrices such that the product $\mathbb{M}\mathbb{V}$ is well-defined. Show that $(\mathbb{M}\mathbb{V})^T = \mathbb{V}^T\mathbb{M}^T$.

Exercise 1.37

Given a matrix $\mathbb{V}$, show that $\mathbb{V}^T\mathbb{V}$ is symmetric.

Exercise 1.38

Show that if $\mathbb{V}$ is symmetric and invertible, then $\mathbb{V}^{-1}$ is also symmetric.

❧ ❋ ❧

While the spectral decomposition is for symmetric matrices, a closely related decomposition holds in more generality. If $\mathbb{M}$ is an $\mathbb{R}^{n\times m}$ matrix, then

$$\mathbb{M} = \mathbb{U}\mathbb{D}\mathbb{V}^T$$

where $\mathbb{U} \in \mathbb{R}^{n\times m}$ and $\mathbb{V} \in \mathbb{R}^{n\times d}$ both have orthonormal columns and $\mathbb{D} \in \mathbb{R}^{m\times d}$ is diagonal and has only non-negative values called singular values.[3] Any such representation is called a

[3] This matrix product is also the sum $\sigma_1\mathbf{u}_1\mathbf{v}_1^T + \ldots + \sigma_d\mathbf{u}_d\mathbf{v}_d^T$, analogously to the spectral decomposition.

singular value decomposition (SVD) of the matrix.[4] (Proofs of the existence of SVDs can be found elsewhere for the interested reader.)

Let's consider the relationship between SVD and spectral decomposition. The singular value decomposition is essentially the statement that there is an orthonormal basis in the domain that maps to a rescaled version of an orthonormal basis in the range. The spectral decomposition says that for symmetric matrices, these orthonormal bases can be chosen to be the exact same basis. A subtle difference is that, by convention, the basis vectors in an SVD are chosen to make the scaling values (diagonals of $\mathbb{D}$) non-negative, whereas a spectral decomposition can have negative scaling values (diagonals of $\mathbb{\Lambda}$).

Exercise 1.39

Use a singular value decomposition for $\mathbb{M}$ to find a spectral decomposition of $\mathbb{M}^T\mathbb{M}$.

❧ ❋ ❧

Homework 1: Slantlet basis

There are many common alternative choices of basis for $\mathbb{R}^n$; different choices have proven convenient for different types of data. As a short project, work through the file *slantlet-basis-homework.R* to see an example of how alternative bases can be used for lossy data compression.

1.7. Quadratic forms

Given a symmetric matrix $\mathbb{V}$ and a vector $\mathbf{w}$, the quantity $\mathbf{w}^T\mathbb{V}\mathbf{w}$ is called a **quadratic form**.

Exercise 1.40

If $\mathbf{q}_1$ is a unit eigenvector of the symmetric matrix $\mathbb{V}$, find the value of the quadratic form $\mathbf{q}_1^T\mathbb{V}\mathbf{q}_1$.

◇

Solution: By a spectral decomposition,

$$\begin{aligned}\mathbf{q}_1^T\mathbb{V}\mathbf{q}_1 &= \mathbf{q}_1^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{q}_1 \\ &= \lambda_1 \underbrace{\mathbf{q}_1^T\mathbf{q}_1}_{1}\underbrace{\mathbf{q}_1^T\mathbf{q}_1}_{1} + \ldots + \lambda_n \underbrace{\mathbf{q}_1^T\mathbf{q}_n}_{0}\underbrace{\mathbf{q}_n^T\mathbf{q}_1}_{0} \\ &= \lambda_1.\end{aligned}$$

Exercise 1.41

If $\mathbb{V}$ is a symmetric matrix, what unit vector $\mathbf{u}$ maximizes the quadratic form $\mathbf{u}^T\mathbb{V}\mathbf{u}$?

◇

[4] Such a representation exists for every d between the rank of $\mathbb{M}$ and $\min\{n, m\}$. If d is larger than the rank of $\mathbb{M}$, then the additional diagonal values are zero.

Solution: By a spectral decomposition,

$$\begin{aligned}\mathbf{u}^T\mathbb{V}\mathbf{u} &= \mathbf{u}^T(\lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T)\mathbf{u} \\ &= \lambda_1\mathbf{u}^T\mathbf{q}_1\mathbf{q}_1^T\mathbf{u} + \ldots + \lambda_n\mathbf{u}^T\mathbf{q}_n\mathbf{q}_n^T\mathbf{u} \\ &= \lambda_1\langle\mathbf{u},\mathbf{q}_1\rangle^2 + \ldots + \lambda_n\langle\mathbf{u},\mathbf{q}_n\rangle^2.\end{aligned}$$

We know that $\langle\mathbf{u},\mathbf{q}_1\rangle,\ldots,\langle\mathbf{u},\mathbf{q}_n\rangle$ provide the coordinates of $\mathbf{u}$ with respect to the basis $\mathbf{q}_1,\ldots,\mathbf{q}_n$. Because $\mathbf{u}$ is a unit vector, the sum of these squared coordinates has to be 1. Therefore, the expression we've derived shows that $\mathbf{u}^T\mathbb{V}\mathbf{u}$ can be understood as a weighted average of the eigenvalues. This weighted average is maximized by placing all of the weight on the largest eigenvalue, that is, by letting $\mathbf{u}$ be equal to a unit eigenvector corresponding to the largest eigenvalue. Such a choice of $\mathbf{u}$ makes $\mathbf{u}^T\mathbb{V}\mathbf{u}$ equal to the largest eigenvalue of $\mathbb{V}$.

❧ ❋ ❧

A symmetric matrix $\mathbb{V}$ is called **positive definite** if for every vector $\mathbf{w} \neq \mathbf{0}$, the quadratic form $\mathbf{w}^T\mathbb{V}\mathbf{w}$ is greater than zero. A weaker condition, **positive semi-definiteness**, only requires every quadratic form to be greater than or equal to zero.

Exercise 1.42

Given a matrix $\mathbb{V}$, show that $\mathbb{V}^T\mathbb{V}$ is positive semi-definite.

⋄

Solution: The quadratic form

$$\begin{aligned}\mathbf{w}^T(\mathbb{V}^T\mathbb{V})\mathbf{w} &= (\mathbf{w}^T\mathbb{V}^T)(\mathbb{V}\mathbf{w}) \\ &= (\mathbb{V}\mathbf{w})^T(\mathbb{V}\mathbf{w})\end{aligned}$$

equals the squared norm of the vector $\mathbb{V}\mathbf{w}$ which is non-negative.

Exercise 1.43

Show that a symmetric matrix $\mathbb{V}$ is positive semi-definite if and only if its eigenvalues are all non-negative.

❧ ❋ ❧

1.8. Principal components

In Homework 1, we used the "slantlets" to understand how a preselected basis can be used to approximate a data vector by only keeping a subset of the vector's most important coefficients with respect to that basis. Another approach is to *design the basis using the data*.

Exercise 1.44

Let $\mathbf{v}_1,\ldots,\mathbf{v}_n$ be the rows of the matrix $\mathbb{M}$. Show that $\mathbf{u}^T(\frac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{u}$ is equal to the average of the squares of the coefficients of $\mathbf{v}_1,\ldots,\mathbf{v}_n$ with respect to $\mathbf{u}$.

⋄

Solution: We'll first express the quadratic form as the squared norm of a vector.

$$\begin{aligned}\mathbf{u}^T(\tfrac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{u} &= \tfrac{1}{n}(\mathbb{M}\mathbf{u})^T(\mathbb{M}\mathbf{u}) \\ &= \tfrac{1}{n}\|\mathbb{M}\mathbf{u}\|^2\end{aligned}$$

The entries of the vector $\mathbb{M}\mathbf{u}$ are the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with respect to $\mathbf{u}$.

$$\begin{aligned}\mathbb{M}\mathbf{u} &= \begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_n & - \end{bmatrix}\mathbf{u} \\ &= \begin{bmatrix} \mathbf{v}_1^T\mathbf{u} \\ \vdots \\ \mathbf{v}_n^T\mathbf{u} \end{bmatrix}\end{aligned}$$

Its squared norm is the sum of its squared entries, so $\frac{1}{n}\|\mathbb{M}\mathbf{u}\|^2$ is the average of those squared entries.

Exercise 1.45

If the zero vector is used to approximate every data vector $\mathbf{v}_1, \ldots, \mathbf{v}_n$, then the average squared approximation error is simply the average squared length $\frac{1}{n}\sum_i \|\mathbf{v}_i\|^2$. Show that if instead $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are approximated by $\langle \mathbf{u}, \mathbf{v}_1\rangle\mathbf{u}, \ldots \langle \mathbf{u}, \mathbf{v}_n\rangle\mathbf{u}$ respectively, then the average squared approximation error is reduced by the average of the squared coefficients $\langle \mathbf{u}, \mathbf{v}_1\rangle^2, \ldots \langle \mathbf{u}, \mathbf{v}_n\rangle^2$.

◇

Solution: For each i, we know that $\mathbf{v}_i - \langle \mathbf{v}_i, \mathbf{u}\rangle\mathbf{u}$ is orthogonal to $\mathbf{u}$. As in Exercise 1.24, the Pythagorean identity implies that

$$\begin{aligned}\|\mathbf{v}_i - \langle \mathbf{v}_i, \mathbf{u}\rangle\mathbf{u}\|^2 &= \|\mathbf{v}_i\|^2 - \|\langle \mathbf{v}_i, \mathbf{u}\rangle\mathbf{u}\|^2 \\ &= \|\mathbf{v}_i\|^2 - \langle \mathbf{v}_i, \mathbf{u}\rangle^2\end{aligned}$$

The average squared approximation error is therefore

$$\begin{aligned}\frac{1}{n}\sum_i \|\mathbf{v}_i - \langle \mathbf{u}, \mathbf{v}_i\rangle\mathbf{u}\|^2 &= \frac{1}{n}\sum_i (\|\mathbf{v}_i\|^2 - \langle \mathbf{v}_i, \mathbf{u}\rangle^2) \\ &= \frac{1}{n}\sum_i \|\mathbf{v}_i\|^2 - \frac{1}{n}\sum_i \langle \mathbf{v}_i, \mathbf{u}\rangle^2.\end{aligned}$$

Exercise 1.46

Identify the unit vector $\mathbf{u}$ that creates the smallest average squared approximation error when $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are approximated by $\langle \mathbf{v}_1, \mathbf{u}\rangle\mathbf{u}, \ldots \langle \mathbf{v}_n, \mathbf{u}\rangle\mathbf{u}$ respectively.

◇

Solution: Exercise 1.45 verified that

$$\frac{1}{n}\sum_i \|\mathbf{v}_i - \langle \mathbf{v}_i, \mathbf{u}\rangle\mathbf{u}\|^2 = \frac{1}{n}\sum_i \|\mathbf{v}_i\|^2 - \frac{1}{n}\sum_i \langle \mathbf{v}_i, \mathbf{u}\rangle^2.$$

From this representation, it is clear that the average squared error is minimized when the average of the squared coefficients of $\mathbf{u}$ is maximized. Exercise 1.44 showed that the average

of the squared coefficients is equal to the quadratic form $\mathbf{u}^T(\frac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{u}$ where $\mathbb{M}$ denotes the matrix with $\mathbf{v}_1, \ldots, \mathbf{v}_n$ as its rows. By Exercise 1.41, this quadratic form is maximized when $\mathbf{u}$ is an eigenvector corresponding to the largest eigenvalue of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$.

Exercise 1.47

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the rows of the matrix $\mathbb{M}$. Show that the average squared length $\frac{1}{n}\sum_i \|\mathbf{v}_i\|^2$ equals the sum of the eigenvalues $\lambda_1 + \ldots + \lambda_n$ of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$.

Exercise 1.48

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the rows of the matrix $\mathbb{M}$. If the zero vector is used to approximate every data vector $\mathbf{v}_1, \ldots, \mathbf{v}_n$, then the average squared approximation error is simply the average squared length $\frac{1}{n}\sum_i \|\mathbf{v}_i\|^2$. Let $\mathbf{q}_j$ be a unit eigenvector of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$. Show that if we instead approximate the data vectors by $\langle \mathbf{q}_j, \mathbf{v}_1\rangle \mathbf{q}_j, \ldots \langle \mathbf{q}_j, \mathbf{v}_n\rangle \mathbf{q}_j$ respectively, the average squared approximation error is reduced by exactly λ_j, the eigenvalue corresponding to eigenvector $\mathbf{q}_j$.

❧ ❋ ❧

Let $\mathbb{M}$ be an $n \times m$ matrix whose rows $\mathbf{v}_1, \ldots, \mathbf{v}_n$ have the zero vector as their average. We will refer to the unit eigenvectors of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$ as the **principal components** of $\mathbb{M}$. In particular, if $\lambda_1 \geq \ldots \geq \lambda_m$ and

$$\frac{1}{n}\mathbb{M}^T\mathbb{M} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_m \mathbf{q}_m \mathbf{q}_m^T,$$

then $\mathbf{q}_1$ is called the *first principal component*, $\mathbf{q}_2$ is called the *second principal component*, and so on.

The eigenvectors of $\mathbb{M}^T\mathbb{M}$ are the same as the eigenvectors of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$ (see Exercise 1.29), but the matrix with the $1/n$ factor has a more convenient interpretation for some purposes.

Exercise 1.49

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the rows of the matrix $\mathbb{M}$. Show that $\frac{1}{n}\mathbb{M}^T\mathbb{M}$ is the matrix whose (j,k)-entry is the average of the product of the jth and kth coordinates of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Exercise 1.50

Explain why the average of the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with respect to a unit vector $\mathbf{u}$ is exactly the same as the coefficient of their average $\frac{1}{n}\sum_i \mathbf{v}_i$ with respect to $\mathbf{u}$?

◇

Solution: Using the formula for the coefficient with respect to $\mathbf{u}$, along with linearity of inner product,

$$\frac{1}{n}\sum_i \langle \mathbf{v}_i, \mathbf{u}\rangle = \left\langle \left(\frac{1}{n}\sum_i \mathbf{v}_i\right), \mathbf{u}\right\rangle.$$

Exercise 1.51

If the mean of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is the zero vector, what is the average of their coefficients with respect to $\mathbf{u}$?

Exercise 1.52

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors whose average $\frac{1}{n}\sum \mathbf{v}_i$ is the zero vector. Write a quadratic form to express the empirical variance of the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with respect to the unit vector $\mathbf{u}$.

◇

Solution: The empirical variance of the coefficients is their average squared deviation from their mean. We know from Exercise 1.51 that the mean of the coefficients is zero. Therefore, the empirical variance of the coefficients is simply the average of their squared values, which we found to be $\mathbf{u}^T(\frac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{u}$ in Exercise 1.44 with $\mathbb{M}$ denoting the matrix whose rows are $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

Exercise 1.53

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be the rows of the matrix $\mathbb{M}$, and suppose the mean of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ is the zero vector. If $\mathbf{q}_j$ is a unit eigenvector of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$, what is the variance of the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with respect to $\mathbf{q}_j$?

◇

Solution: From Exercise 1.52, the empirical variance is $\mathbf{q}_j^T(\frac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{q}_j$. Exercise 1.40 shows that this is λ_j, the eigenvalue of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$ corresponding to eigenvector $\mathbf{q}_j$.

Exercise 1.54

Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be vectors whose average $\frac{1}{n}\sum \mathbf{v}_i$ is the zero vector. Identify a unit vector $\mathbf{u}$ for which the empirical variance of the coefficients of $\mathbf{v}_1, \ldots, \mathbf{v}_n$ with respect to $\mathbf{u}$ is maximized.

◇

Solution: Let $\mathbb{M}$ denote the matrix whose rows are $\mathbf{v}_1, \ldots, \mathbf{v}_n$. Because the vectors have an average of zero, we observed in Exercise 1.52 that the empirical variance of their coefficients is equal to the average of their squared coefficients which is $\mathbf{u}^T(\frac{1}{n}\mathbb{M}^T\mathbb{M})\mathbf{u}$. Recall from Exercise 1.41 that this is maximized by any unit eigenvector of $\frac{1}{n}\mathbb{M}^T\mathbb{M}$ corresponding to the largest eigenvalue, i.e. by the first principal component of $\mathbb{M}$.

Exercise 1.55

Define the vector $\mathbf{m}$ to be the mean $\frac{1}{n}\sum_i \mathbf{v}_i$, and define $\mathbf{w}_i := \mathbf{v}_i - \mathbf{m}$ for $i \in \{1, \ldots, n\}$. What's the mean of $\mathbf{w}_1, \ldots, \mathbf{w}_n$?

❧ ❋ ❧

Earlier we defined the principal components for a set of vectors with mean zero; now we will provide the general definition. If, as in Exercise **??**, we define $\mathbf{m}$ to be the mean $\frac{1}{n}\sum_i \mathbf{v}_i$, the

principal components are the unit eigenvectors of

$$\mathbf{\Sigma} := \frac{1}{n}\begin{bmatrix} - & \mathbf{v}_1 - \mathbf{m} & - \\ & \vdots & \\ - & \mathbf{v}_n - \mathbf{m} & - \end{bmatrix}^T \begin{bmatrix} - & \mathbf{v}_1 - \mathbf{m} & - \\ & \vdots & \\ - & \mathbf{v}_n - \mathbf{m} & - \end{bmatrix}.$$

Notice that the vectors need only be replaced by their *centered* versions. Subtracting the mean from each vectors is called *centering* them. Picturing the vectors as points on a scatterplot, centering them involves translating every point by the same amount. This translation of the entire set doesn't affect the empirical variance of the coefficients in any direction: the variance of the centered vectors is exactly the same as the variance of the original vectors. Therefore, the results of Exercises 1.53 and 1.54 hold more generally: the variances in the principal component directions are the eigenvalues of $\mathbf{\Sigma}$, and the direction of largest variance is the first principal component, that is, the eigenvector of $\mathbf{\Sigma}$ corresponding to the largest eigenvalue.

Homework 2: Principal components compression

Work through the file *nist-faces-homework.R* to understand how principal component basis vectors provide an alternative approach to achieving lossy compression for a batch of data.

1.9. Orthogonal projection

Given a subspace $\mathcal{S} \subseteq \mathbb{R}^n$, the set of all vectors that are orthogonal to $\mathcal{S}$ is called the **orthogonal complement** of $\mathcal{S}$ and is denoted $\mathcal{S}^\perp$.

Exercise 1.56

Given a subspace $\mathcal{S}$, show that $\mathcal{S}^\perp$ is also a subspace.

◇

Solution: If $\mathbf{w}_1$ and $\mathbf{w}_2$ are in $\mathcal{S}^\perp$, then we need to show that an arbitrary linear combination $b_1\mathbf{w}_1 + b_2\mathbf{w}_2$ is also in $\mathcal{S}^\perp$. Letting $\mathbf{v}$ be an arbitrary vector in $\mathcal{S}$,

$$\begin{aligned}\langle b_1\mathbf{w}_1 + b_2\mathbf{w}_2, \mathbf{v}\rangle &= b_1 \underbrace{\langle \mathbf{w}_1, \mathbf{v}\rangle}_{0} + b_2 \underbrace{\langle \mathbf{w}_2, \mathbf{v}\rangle}_{0} \\ &= 0.\end{aligned}$$

❧ ❋ ❧

If $\mathcal{S}$ is a subspace of $\mathbb{R}^n$, any given vector $\mathbf{y} \in \mathbb{R}^n$ has a *unique* representation as the sum of a vector in $\mathcal{S}$ and a vector in $\mathcal{S}^\perp$. The vector in $\mathcal{S}$ in this representation is called the **orthogonal projection** of $\mathbf{y}$ onto $\mathcal{S}$. This observation (Exercise 1.60) generalizes our earlier result in which the subspace was the span of a single vector (Exercise 1.19).

Exercise 1.57

Let $\hat{\mathbf{y}}$ denote the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. How do we know that $\mathbf{y} - \hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}^\perp$?

Exercise 1.58

Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. Show that the vector in $\mathcal{S}$ that is closest to $\mathbf{y}$ is exactly the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$.

◇

Solution: Let $\mathbf{v}$ be an arbitrary vector in $\mathcal{S}$ and let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. Realizing that $\mathbf{v} - \hat{\mathbf{y}}$ is in $\mathcal{S}$ and that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\mathcal{S}$, we observe a right triangle with sides $\mathbf{y} - \mathbf{v}$, $\hat{\mathbf{y}} - \mathbf{v}$, and $\mathbf{y} - \hat{\mathbf{y}}$. By the Pythagorean identity,

$$\|\mathbf{y} - \mathbf{v}\|^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \|\mathbf{v} - \hat{\mathbf{y}}\|^2.$$

The first term on the right doesn't depend on the choice of $\mathbf{v}$, so the quantity is minimized by choosing $\mathbf{v}$ equal to $\hat{\mathbf{y}}$ to make the second term zero.

Exercise 1.59

If a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ has dimension m, what is the dimension of $\mathcal{S}^\perp$?

Exercise 1.60

Let $\mathbf{u}_1, \ldots, \mathbf{u}_m$ be an orthonormal basis for a subspace $\mathcal{S} \subseteq \mathbb{R}^n$. Show that

$$\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m$$

is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$.

◇

Solution: Let $\mathbf{u}_{m+1}, \ldots, \mathbf{u}_n$ be additional vectors that make $\mathbf{u}_1, \ldots, \mathbf{u}_n$ into an orthonormal basis. Then as in Exercise 1.22, $\mathbf{y}$ has the representation

$$\begin{aligned}\mathbf{y} &= \langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_n \rangle \mathbf{u}_n \\ &= \underbrace{\langle \mathbf{y}, \mathbf{u}_1 \rangle \mathbf{u}_1 + \ldots + \langle \mathbf{y}, \mathbf{u}_m \rangle \mathbf{u}_m}_{\in \mathcal{S}} + \underbrace{\langle \mathbf{y}, \mathbf{u}_{m+1} \rangle \mathbf{u}_{m+1} + \ldots + \langle \mathbf{y}, \mathbf{u}_n \rangle \mathbf{u}_n}_{\perp \mathcal{S}}\end{aligned}$$

with all coefficients uniquely determined by these inner products. Thus the specified vector satisfies the definition of orthogonal projection.

❧ ❋ ❧

The **column space** of a matrix is the span of its column vectors; we will denote it by $C(\mathbb{V})$. The **rank** of a matrix is the dimension of its column space.

In the next exercises, we'll work toward the task of finding the coefficient vector $\hat{\mathbf{b}}$ for which $\mathbb{V}\hat{\mathbf{b}}$ equals the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbb{V}$.

Exercise 1.61

Let $\mathbb{V} \in \mathbb{R}^{n \times m}$ be a matrix. Given $\mathbf{y} \in \mathbb{R}^n$, explain why the *Normal equation*

$$\mathbb{V}^T \mathbb{V} \hat{\mathbf{b}} = \mathbb{V}^T \mathbf{y}$$

is satisfied by the coefficient vector $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_m) \in \mathbb{R}^m$ if and only if $\mathbb{V}\hat{\mathbf{b}}$ is the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{V})$.

◇

Solution: The orthogonal projection $\mathbb{V}\hat{\mathbf{b}}$ is the unique vector in $C(\mathbb{V})$ with the property that $\mathbf{y} - \mathbb{V}\hat{\mathbf{b}} \perp C(\mathbb{V})$. It is equivalent to check that $\mathbf{y} - \mathbb{V}\hat{\mathbf{b}}$ is orthogonal to every column of $\mathbb{V}$. Equivalently the following quantity should be equal to the zero vector:

$$\mathbb{V}^T(\mathbf{y} - \mathbb{V}\hat{\mathbf{b}}) = \begin{bmatrix} - & \mathbf{v}_1 & - \\ & \vdots & \\ - & \mathbf{v}_m & - \end{bmatrix} (\mathbf{y} - \mathbb{V}\hat{\mathbf{b}}) = \begin{bmatrix} \mathbf{v}_1^T(\mathbf{y} - \mathbb{V}\hat{\mathbf{b}}) \\ \vdots \\ \mathbf{v}_m^T(\mathbf{y} - \mathbb{V}\hat{\mathbf{b}}) \end{bmatrix}.$$

Setting this vector $\mathbb{V}^T(\mathbf{y} - \mathbb{V}\hat{\mathbf{b}})$ equal to the zero vector results in the Normal equation.

Exercise 1.62

Explain how you know that $\mathbb{V}^T\mathbb{V}$ is invertible if and only if the columns of $\mathbb{V}$ are linearly independent.

Exercise 1.63

Let $\mathbb{V} \in \mathbb{R}^{n\times m}$ be a matrix with linearly independent columns. Given $\mathbf{y} \in \mathbb{R}^n$, provide a formula for the coefficient vector $\mathbf{b}$ for which $\mathbb{V}\hat{\mathbf{b}}$ equals the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{V})$.

Solution: Because the columns are linearly independent, we know that $\mathbb{V}^T\mathbb{V}$ is invertible and thus the Normal equation

$$\mathbb{V}^T\mathbb{V}\hat{\mathbf{b}} = \mathbb{V}^T\mathbf{y}$$

is uniquely solved by $\hat{\mathbf{b}} = (\mathbb{V}^T\mathbb{V})^{-1}\mathbb{V}^T\mathbf{y}$.

Exercise 1.64

Let $\mathcal{S} \subseteq \mathbb{R}^n$ be a subspace. If $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$ and $\hat{\mathbf{z}}$ is the orthogonal projection of $\mathbf{z}$ onto $\mathcal{S}$, find the orthogonal projection of $\mathbf{y} + \mathbf{z}$ onto $\mathcal{S}$.

1.10. Orthogonal projection matrices

Given a subspace $\mathcal{S}$, there is a unique matrix that maps any vector to its orthogonal projection onto $\mathcal{S}$. This matrix is called the **orthogonal projection matrix** onto $\mathcal{S}$.

Exercise 1.65

Let $\mathbb{H}$ be the orthogonal projection matrix onto a subspace $\mathcal{S} \subseteq \mathbb{R}^n$. Show that every vector in $\mathcal{S}$ is an eigenvector of $\mathbb{H}$.

Solution: If $\mathbf{w}$ is in $\mathcal{S}$, then clearly $\mathbf{w} = \mathbf{w} + \mathbf{0}$ is the unique representation of $\mathbf{w}$ as the sum of a vector in $\mathcal{S}$ and a vector orthogonal to $\mathcal{S}$. Therefore $\mathbb{H}\mathbf{w} = \mathbf{w}$, which means that $\mathbf{w}$ is an eigenvector with eigenvalue 1.

Exercise 1.66

Let $\mathbb{H}$ be the orthogonal projection matrix onto a subspace $\mathcal{S} \subseteq \mathbb{R}^n$. Show that every vector in $\mathcal{S}^\perp$ is an eigenvector of $\mathbb{H}$.

$\diamond$

Solution: If $\mathbf{w} \perp \mathcal{S}$, then clearly $\mathbf{w} = \mathbf{0} + \mathbf{w}$ is the unique representation of $\mathbf{w}$ as the sum of a vector in $\mathcal{S}$ and a vector orthogonal to $\mathcal{S}$. Therefore $\mathbb{H}\mathbf{w} = \mathbf{0}$, which means that $\mathbf{w}$ is an eigenvector with eigenvalue 0.

Exercise 1.67

If $\mathbb{H}$ is an orthogonal projection matrix, show that it is *idempotent* (meaning $\mathbb{H}\mathbb{H} = \mathbb{H}$).

Exercise 1.68

Use the result of Exercise 1.31 to prove that every orthogonal projection matrix is symmetric.

Exercise 1.69

If $\mathbb{H}$ is an orthogonal projection matrix, describe a spectral decomposition for $\mathbb{H}$.

Exercise 1.70

Let $\mathbb{H}$ be a symmetric matrix whose only eigenvalues are 0 and 1. Show that $\mathbb{H}$ is the orthogonal projection matrix onto the eigenspace corresponding to eigenvalue 1.

Exercise 1.71

If $\mathbb{H}$ is an orthogonal projection matrix, show that its trace equals the dimension of the subspace that it projects onto.

Exercise 1.72

Let $\mathbb{V}$ be a matrix. Explain why the rank of the orthogonal projection matrix onto $C(\mathbb{V})$ must be exactly the same as the rank of $\mathbb{V}$.

If $\mathcal{S}_0$ is a subspace of $\mathcal{S}_1$ which is a subspace of $\mathbb{R}^n$, we can define the *orthogonal complement* of $\mathcal{S}_0$ within $\mathcal{S}_1$ to be the set of vectors in $\mathcal{S}_1$ that are orthogonal to $\mathcal{S}_0$. We've seen from Exercise 1.56 that orthogonal complements are subspaces, and the same reasoning applies here.

Exercise 1.73

Let $\mathbb{H}_1$ be the orthogonal projection matrix onto $\mathcal{S}_1$, and let $\mathbb{H}_0$ be the orthogonal projection matrix onto $\mathcal{S}_0 \subseteq \mathcal{S}_1$. Show that $\mathbb{H}_1 - \mathbb{H}_0$ is the orthogonal projection matrix onto the orthogonal complement of $\mathcal{S}_0$ within $\mathcal{S}_1$.

Exercise 1.74

Let $\mathcal{S}_0 \subseteq \mathcal{S}_1$ be subspaces, and let $\mathbb{H}_0$ and $\mathbb{H}_1$ be orthogonal projection matrices onto $\mathcal{S}_0$ and $\mathcal{S}_1$ respectively. Verify that $\mathbb{H}_0 \circ \mathbb{H}_1 = \mathbb{H}_1 \circ \mathbb{H}_0 = \mathbb{H}_0$.

Finally, let's derive a specific formula for calculating orthogonal projection matrices.

Exercise 1.75

Let $\mathbb{V} \in \mathbb{R}^{n \times m}$ be a matrix with linearly independent columns. Provide a formula for the orthogonal projection matrix onto $C(\mathbb{V})$.

◇

Solution: We've already derived in Exercise 1.63 a formula for the desired coefficient vector $\hat{\mathbf{b}} = (\mathbb{V}^T\mathbb{V})^{-1}\mathbb{V}^T\mathbf{y}$, so we simply plug this into $\mathbb{V}\hat{\mathbf{b}}$ to find the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{V})$.

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbb{V}\hat{\mathbf{b}} \\ &= \mathbb{V}(\mathbb{V}^T\mathbb{V})^{-1}\mathbb{V}^T\mathbf{y}\end{aligned}$$

Therefore, we see that $\mathbf{y}$ is mapped to its orthogonal projection onto $C(\mathbb{V})$ by the matrix $\mathbb{V}(\mathbb{V}^T\mathbb{V})^{-1}\mathbb{V}^T$.

If the columns of $\mathbb{V}$ aren't linearly independent, the orthogonal projection matrix onto $C(\mathbb{V})$ can still be readily found. Construct a matrix $\mathbb{W}$ whose columns are a basis for $C(\mathbb{V})$. Then $\mathbb{W}(\mathbb{W}^T\mathbb{W})^{-1}\mathbb{W}^T$ is the orthogonal projection matrix onto $C(\mathbb{W})$ which is exactly the same subspace as $C(\mathbb{V})$.

Exercise 1.76

Provide a formula for the matrix that maps any $\mathbf{y} \in \mathbb{R}^n$ to its orthogonal projection onto the span of a unit vector $\mathbf{u}$.

◇

Solution: We can apply our formula from Exercise 1.75 to the matrix that has $\mathbf{u}$ as its only column to get $\mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}$ which simplifies to $\mathbf{u}\mathbf{u}^T$ because $\mathbf{u}^T\mathbf{u} = \|\mathbf{u}\|^2 = 1$. Notice that $\mathbf{u}\mathbf{u}^T$ maps $\mathbf{y}$ to $(\mathbf{u}^T\mathbf{y})\mathbf{u}$ which is indeed what we've previously derived as the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{u}$ according to Exercise 1.20. Notice that we can now interpret spectral decompositions as representing each symmetric matrix as a linear combination of orthogonal projection matrices onto the spans of the eigenvectors.

CHAPTER

2

LEAST-SQUARES REGRESSION

Regression means deciding on a function of the explanatory variable(s) that fits or helps predict a quantitative response variable. Each possible function creates a **residual** for every observation, which is defined as the value of the response variable minus the fitted value (the value predicted by that function). A commonly used criterion for selecting a function is *minimization of the sum of squared residuals*; the optimization may also include a penalty term that increases with the number of parameters or their magnitudes.

Minimizing the sum of squared residuals takes on a special meaning in the case of *probabilistic modeling* with iid Normal additive errors, as you will show in Exercise 6.1. However, this procedure makes intuitive sense even if we don't make any assumptions about the "true relationships" among the variables based on the physical mechanism that generates them. We can still say that the regression function is designed to *summarize* the relationship among the variables *in the data*; furthermore, such procedures can be used to *compress data*.

2.1. Visualizing the observations

The most natural way to understand the idea of least-squares regression is to visualize the data *observations* as points in a space.

2.1.1. Least-squares point

As we learn about linear regression and modeling, a theme will be to consider what happens when *one more* explanatory variable is included. The following exercises are designed to prime your thinking with the most basic case: zero explanatory variables; aspects of this case will remain important throughout our course of study.

Without explanatory variables to distinguish the observations, it's only natural to use the same single number to predict every response value. The number that minimizes the sum of squared residuals turns out to be the sample average, denoted $\bar{y}$; we might call it the *least-squares point*.

Exercise 2.1

Assume that you have data values $y_1, \ldots, y_n$ without any explanatory variables. Use calculus or completing the square to show that the fitted value for the response variable that minimizes the sum of squared residuals is the sample average.

Exercise 2.2

Let $x_1, \ldots, x_n$ be real numbers. Let N be uniformly distributed on $\{1, \ldots, n\}$. (The distribution of x_N is called the *empirical distribution* of $x_1, \ldots, x_n$.) What is the expected value of x_N?

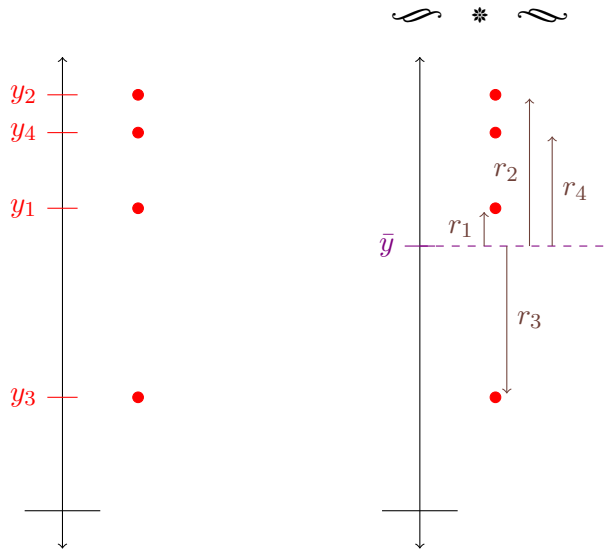


Figure 2.1: Left: Generic values of a single quantitative (response) variable. Right: Arrows represent the residuals produced when the constant $\bar{y}$ is used to fit the variable.

2.1.2. Least-squares line

Let $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be paired measurements of a quantitative explanatory and a quantitative response variable. The data can be neatly visualized on a *scatterplot*, as in Figure 2.2.

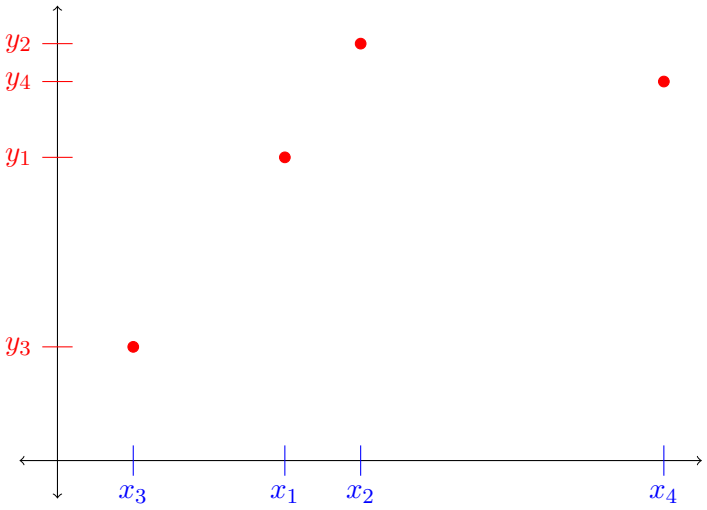


Figure 2.2: The response variable values are the same as in Figure 2.1, but now we also have an explanatory variable value for each observation which allows us to draw a scatterplot. In a scatterplot, every observation in the data is represented by a point.

The scatterplot could indicate any number of types of relationship between the data variables. Here, we're interested in summarizing the relationship with a *line*. If we think of the potential lines as *predicting* the response values based on the explanatory values, then it's natural to quantify the quality of each line according to how much those predictions differ from the actual

response values. Recall that these differences (actual minus predicted) are called the residuals produced by that line.

Selecting a line to fit the data is called *simple linear regression*. The **least-squares line** is the line that produces the smallest sum of squared residuals; see Figure 2.3.

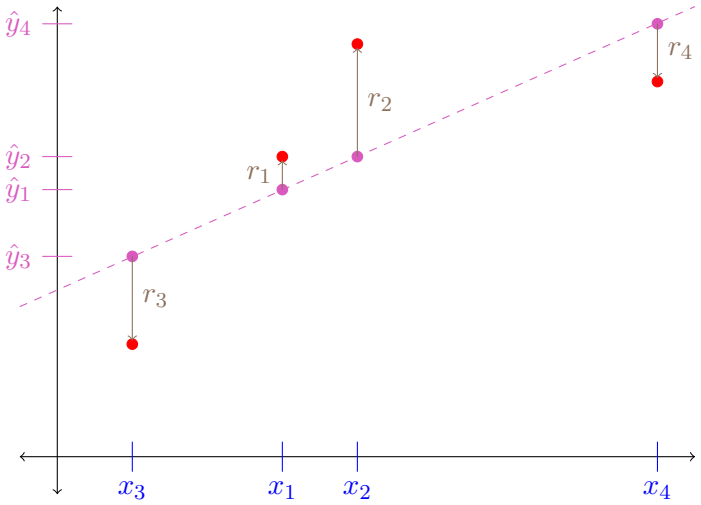


Figure 2.3: Arrows represent the residuals produced when the dotted line is used to fit the variable. In fact, the particular line shown is the least-squares line for this data, that is, the line that produces the smallest sum of squared residuals.

Exercise 2.3

Is it possible for the least-squares line's sum of squared residuals to be greater than the least-squares point's sum of squared residuals?

Exercise 2.4

Assume that you have response values $y_1, \ldots, y_n$ and explanatory values $x_1, \ldots, x_n$. Use calculus to find a formula for the least-squares line.

2.1.3. Least-squares hyperplane

To make sure you understand how the picture continues to generalize, we'll now consider the case in which the data comprises a quantitative response variables and *two* quantitative explanatory variables. The ordinary scatterplot doesn't have enough dimensions to let us visualize these points. We require a two-dimensional real plane just to index all the possible pairs of explanatory variable values. We need a third axis rising perpendicularly from this plane to index the possible response variable values. This picture is called a *3D scatterplot*; see Figure 2.4.

Let $x_1^{(1)}, \ldots, x_n^{(1)}$ and $x_1^{(2)}, \ldots, x_n^{(2)}$ be the values of two quantitative explanatory variables. Now we consider regression functions of the form $f_{a,b,c}(x^{(1)}, x^{(2)}) = a + bx^{(1)} + cx^{(2)}$ with (a, b, c) ranging over $\mathbb{R}^3$. Each possible (a, b, c) defines a different *plane*. Each plane creates a residual at each observation defined, as always, to be the value of the response variable minus the fitted value; in this case, the fitted value is the height of the plane at the location of the two explanatory variable values. The *least-squares plane* is the plane that has the smallest sum of squared residuals; see Figure 2.5.

We won't derive the formulas for the least-squares plane's coefficients here; that derivation becomes much easier once we learn how to visualize the *variables*. In fact, we'll work out a least-squares coefficient formula that's valid for any number of dimensions. A *hyperplane* generalizes

the concept of line and plane to arbitrarily many dimensions. With m explanatory variables, one can imagine the observations as points in $\mathbb{R}^{m+1}$ and seek the m-dimensional hyperplane that minimizes the sum of squared residuals; the solution will be derived in Section 2.2.3.

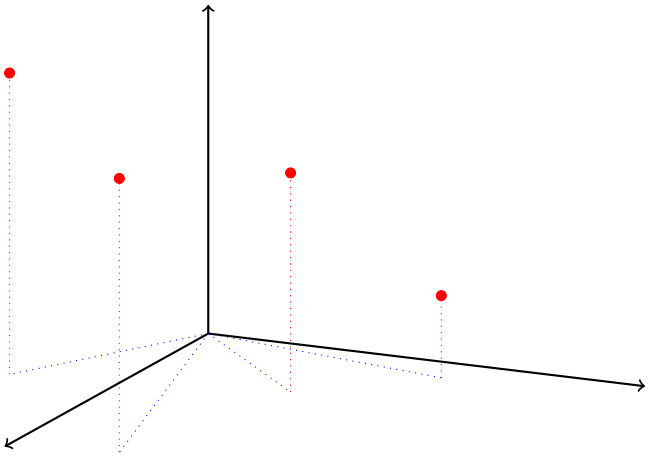


Figure 2.4: The response variable and the first explanatory variable are the same as in the previous plots, but now we also have a second explanatory variable value for each observation which allows us to draw a 3D scatterplot. The explanatory variables correspond to the horizontal plane while the response variable corresponds to the height that the point is placed above that plane.

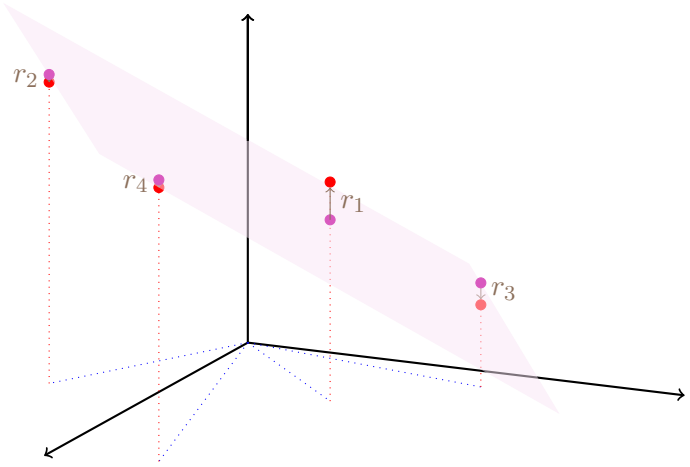


Figure 2.5: The fitted values are shown on the least-squares plane. Arrows represent the residuals.

2.2. Visualizing the variables

A more challenging, but ultimately more powerful, way to understand least-squares regression comes from visualizing the data *variables* as vectors in a space. We think of the *vector* of response values $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ along with a vector of predicted values $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_n) \in \mathbb{R}^n$. Together, they define the vector of residuals $\mathbf{r} := (y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n) = \mathbf{y} - \hat{\mathbf{y}} \in \mathbb{R}^n$. We see that the sum of squared residuals is exactly the squared distance from $\mathbf{y}$ to $\hat{\mathbf{y}}$:

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \sum_i (y_i - \hat{y}_i)^2.$$

Recall from Exercise 1.58 that, given any vector $\mathbf{y} \in \mathbb{R}^n$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^n$, the vector in $\mathcal{S}$ that is closest (in Euclidean distance) to $\mathbf{y}$ is the *orthogonal projection* of $\mathbf{y}$ onto $\mathcal{S}$. Therefore, we draw a key conclusion: *whenever the set of possible vectors of predicted values comprise a subspace, the least-squares predictions are exactly the orthogonal projection of* $\mathbf{y}$ *onto that subspace.*

2.2.1. Least-squares point

Again, we start with the simple case in which there are no explanatory variables. But instead of thinking of $y_1, \ldots, y_n$ as n separate numbers, think of them together as a single vector $\mathbf{y} \in \mathbb{R}^n$. If we wish to select a single number a to predict $y_1, \ldots, y_n$, it's convenient to identify that number with a vector in $\mathbb{R}^n$ as well: the *constant vector* $(a, \ldots, a)$ that has a as all n of its components.

An even more convenient representation of this constant vector is $a\mathbf{1}$ where $\mathbf{1} := (1, \ldots, 1) \in \mathbb{R}^n$. Thus, selecting a real constant to predict $y_1, \ldots, y_n$ is equivalent to selecting an a for which $a\mathbf{1}$ predicts $\mathbf{y}$. The span of $\mathbf{1}$, i.e. the set $\{a\mathbf{1} : a \in \mathbb{R}\}$, can be called *the constant subspace* because each vector in span$\{\mathbf{1}\}$ has the same value for all of its entries.

Given any constant prediction a, the residual vector is $(y_1 - a, \ldots, y_n - a) = \mathbf{y} - a\mathbf{1}$. The sum of squared residuals is minimized when $a\mathbf{1}$ is the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$.

Exercise 2.5

Use the formula from Exercise 1.19 to find the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$.

◇

Solution: The squared length of $\mathbf{1} \in \mathbb{R}^n$ is $1^2 + \ldots + 1^2 = n$, while its inner product with $\mathbf{y}$ is simply the sum of $y_1, \ldots, y_n$. According to our formula, the orthogonal projection is

$$\frac{\langle \mathbf{y}, \mathbf{1} \rangle}{\|\mathbf{1}\|^2}\mathbf{1} = (\underbrace{\tfrac{1}{n}\sum_i y_i}_{\text{“}\bar{y}\text{”}})\mathbf{1}.$$

❧ ❋ ❧

The vector $\bar{y}\mathbf{1}$ we will also denote $\bar{\mathbf{y}}$.

Exercise 2.6

Use the Pythagorean identity to decompose the average of the squared differences between the response values and $a \in \mathbb{R}$

$$\tfrac{1}{n}\sum_i (y_i - a)^2$$

into two terms, one of which is the empirical variance of $y_1, \ldots, y_n$.

◇

Solution: We can write $\sum_i (y_1 - a)^2$ as the squared norm $\|\mathbf{y} - a\mathbf{1}\|^2$. The vector $\mathbf{y} - a\mathbf{1}$ is the hypotenuse of the right triangle whose other two sides are $\mathbf{y} - \bar{\mathbf{y}}$ and $\bar{\mathbf{y}} - a\mathbf{1}$. By the Pythagorean identity

$$\begin{aligned}
\tfrac{1}{n}\sum_i (y_i - a)^2 &= \tfrac{1}{n}\|\mathbf{y} - a\mathbf{1}\|^2 \\
&= \tfrac{1}{n}[\|\mathbf{y} - \bar{\mathbf{y}}\|^2 + \|\bar{\mathbf{y}} - a\mathbf{1}\|^2] \\
&= \tfrac{1}{n}\left[\sum_i (y_i - \bar{y})^2 + n(\bar{y} - a)^2\right] \\
&= \tfrac{1}{n}\sum_i (y_i - \bar{y})^2 + (\bar{y} - a)^2
\end{aligned}$$

Exercise 2.7

Let $\mathbf{y} = (-2, 3, 11)$. Calculate the least-squares prediction vector $\bar{\mathbf{y}}$ along with the least-squares residual vector $\mathbf{y} - \bar{\mathbf{y}}$.

With only 3 observations, as in Exercise 2.7, it's easy to envision or to draw the relevant vectors in $\mathbb{R}^3$. In general, the picture takes place in $\mathbb{R}^n$, where n is the number of observations. What if n is larger than 3 or is left unspecified? We can still draw essentially the same picture, realizing that we're only depicting a three-dimensional subspace.[1] As long as no more than three vectors are being depicted, we know that they will all lie in a three-dimensional subspace of $\mathbb{R}^n$ and can therefore be accurately depicted in a sketch; see Figures 2.6 and 2.7.

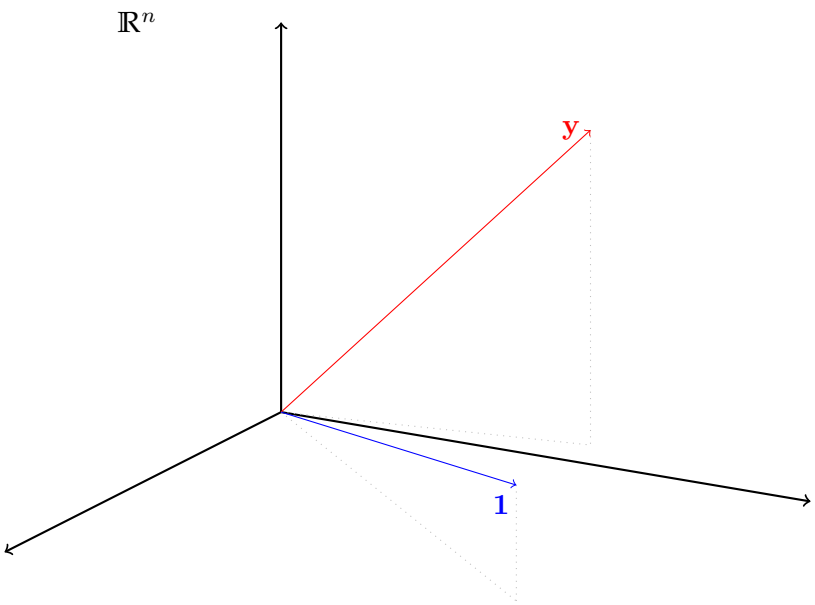


Figure 2.6: A generic picture of the constant vector $\mathbf{1}$ *and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. We know that there exist infinitely many three-dimensional subspaces that includes the origin along with these two vectors' endpoints, so we can assume that the view shown here is one such three-dimensional subspace of* $\mathbb{R}^n$*.*

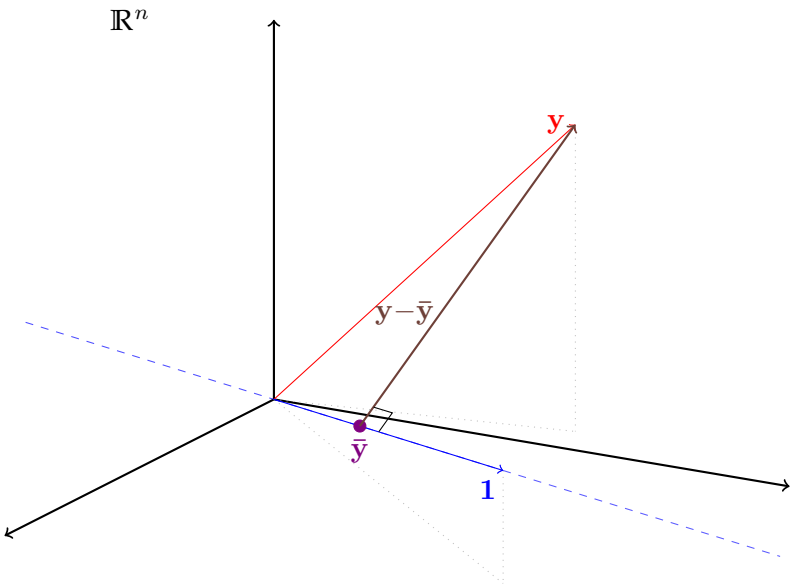


Figure 2.7: A generic picture of the constant vector $\mathbf{1}$ *and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dotted line follows a portion of the span of* $\mathbf{1}$*.* $\bar{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto that subspace, and* $\mathbf{y} - \bar{\mathbf{y}}$ *is orthogonal to it.*

[1] If n is left unspecified, then a case with $n < 3$ can generally be thought of as occupying a subspace of our three-dimensional picture; the three dimensional drawing remains valid.

2.2.2. Least-squares line

Let's think about how the picture looks when there's also an explanatory variable vector $\mathbf{x} = (x_1, \ldots, x_n)$. The vector of fitted values for $\mathbf{y}$ now has the form $b_0\mathbf{1} + b_1\mathbf{x}$ for $b_0, b_1 \in \mathbb{R}$. That means that the set of possible fitted values is exactly the span of $\{\mathbf{1}, \mathbf{x}\}$, which forms a two-dimensional subspace[2] of $\mathbb{R}^n$.

Again, regardless of how many data points there are, the relevant vectors can be accurately depicted in a three-dimensional subspace of $\mathbb{R}^n$ that includes $\mathbf{1}$, $\mathbf{x}$ and $\mathbf{y}$; see Figures 2.8 and 2.9. All of the familiar rules of Euclidean geometry apply, of course, in this three-dimensional subspace.

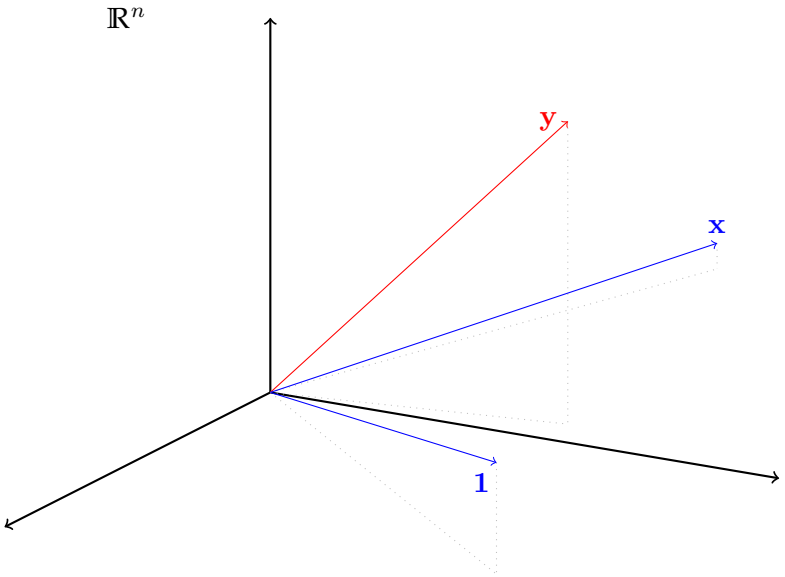


Figure 2.8: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. We know that there exists a three-dimensional subspace that includes these three vectors, so we can assume that the view shown here is such a three-dimensional slice of* $\mathbb{R}^n$*.*

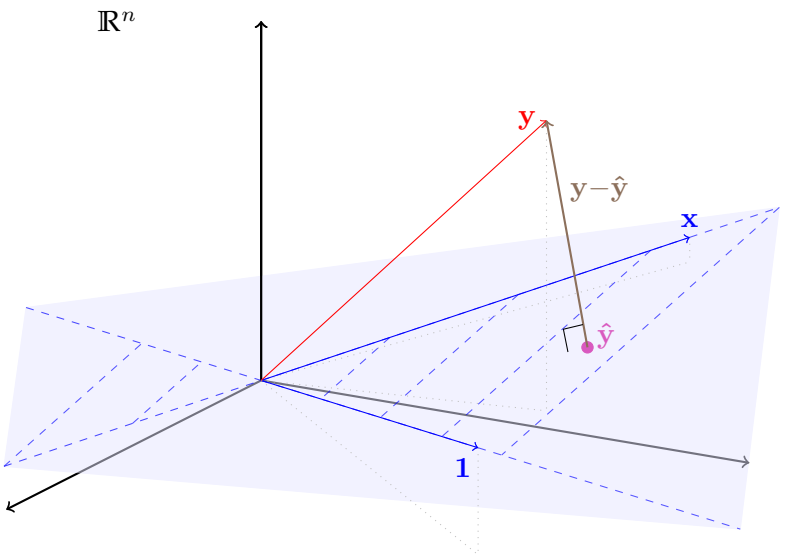


Figure 2.9: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dashed lines outline a portion of the span of* $\mathbf{1}$ *and* $\mathbf{x}$*.* $\hat{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto that subspace, and* $\mathbf{y} - \hat{\mathbf{y}}$ *is orthogonal to it.*

Exercise 2.8

The variables picture provides us with a more specific answer to Exercise 2.3. Use the Pythagorean identity to quantify the difference between the least-squares point's sum of squared residuals and the least-squares line's sum of squared residuals.

◇

Solution: Because $\bar{\mathbf{y}}$ is in the span of $\mathbf{1}$ and $\mathbf{x}$, we see that the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ must

[2]This is true unless $\mathbf{x}$ is constant, in which case span$\{\mathbf{1}, \mathbf{x}\}$ is one-dimensional. In that case, the explanatory variable vector doesn't add anything to the set of possible fits.

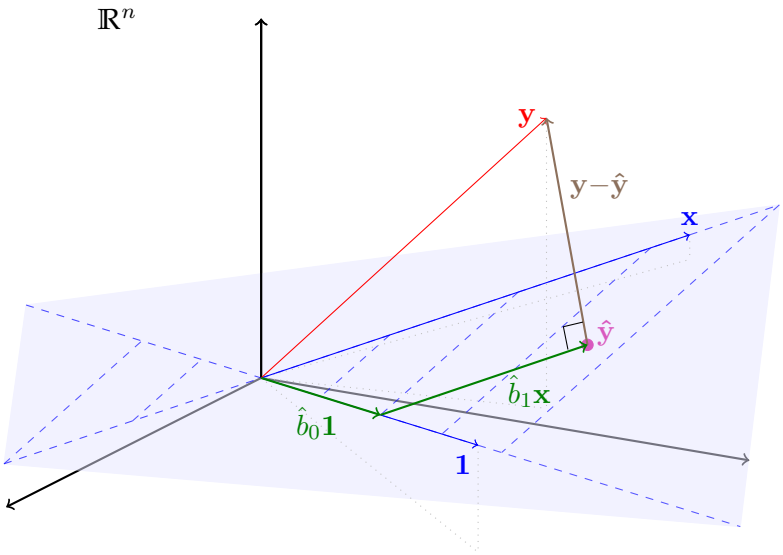


Figure 2.10: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, a response variable vector* $\mathbf{y}$*, and its projection onto the span of* $\mathbf{1}$ *and* $\mathbf{x}$*. The least-squares coefficients* $\hat{b}_0$ *and* $\hat{b}_1$ *are the unique coefficients of* $\mathbf{1}$ *and* $\mathbf{x}$ *for which* $\hat{b}_0\mathbf{1}+\hat{b}_1\mathbf{x}=\hat{\mathbf{y}}$*.*

be orthogonal to $\hat{\mathbf{y}}-\bar{\mathbf{y}}$. Thus by the Pythagorean identity,

$$\begin{aligned}\|\mathbf{y}-\bar{\mathbf{y}}\|^2 &= \|\mathbf{y}-\hat{\mathbf{y}}\|^2+\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2\\ \Rightarrow\|\mathbf{y}-\bar{\mathbf{y}}\|^2-\|\mathbf{y}-\hat{\mathbf{y}}\|^2 &= \|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2.\end{aligned}$$

Exercise 2.9

Let $\mathbf{y}=(y_1,\ldots,y_n)$ be the response variable, and assume $\mathbf{x}=c\mathbf{1}$ for some $c\in\mathbb{R}$. Find the set of (b_0,b_1) for which $b_0\mathbf{1}+b_1\mathbf{x}$ is equal to the least-squares fit.

Exercise 2.10

Use the Gram-Schmidt algorithm to find an orthonormal basis for the span of $\mathbf{1}$ and $\mathbf{x}$.

◇

Solution: We first normalize $\mathbf{1}$ by dividing by its length $\sqrt{n}$. Then subtract from $\mathbf{x}$ its orthogonal projection onto $\frac{\mathbf{1}}{\sqrt{n}}$ which is $\bar{\mathbf{x}}:=\bar{x}\mathbf{1}$, and divide the resulting vector by its length. The resulting orthonormal basis vectors for $\text{span}\{\mathbf{1},\mathbf{x}\}$ are

$$\frac{\mathbf{1}}{\sqrt{n}} \quad \text{and} \quad \frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}.$$

Exercise 2.11

Use the observation from Exercise 1.60 along with the orthonormal basis derived in Exercise 2.10 to find the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$ and $\mathbf{x}$.

◇

Solution: The orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$ and $\mathbf{x}$ is simply the sum of its orthogonal projections onto the spans of the orthonormal basis vectors $\frac{\mathbf{1}}{\sqrt{n}}$ and $\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}$. The orthogonal projection of $\mathbf{y}$ onto the span of $\frac{\mathbf{1}}{\sqrt{n}}$ is the same as its orthogonal projection onto $\mathbf{1}$ which we've already found to be $\bar{\mathbf{y}}$. Using the formula from Exercise 1.20, the orthogonal

projection onto $\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}$ must be

$$\left\langle \mathbf{y}, \tfrac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|} \right\rangle \tfrac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|} = \frac{\langle \mathbf{y}, \mathbf{x}-\bar{\mathbf{x}} \rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}(\mathbf{x}-\bar{\mathbf{x}}).$$

Putting these terms together, we find the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$ and $\mathbf{x}$ to be

$$\hat{\mathbf{y}} = \bar{\mathbf{y}} + \frac{\langle \mathbf{y}, \mathbf{x}-\bar{\mathbf{x}} \rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}(\mathbf{x}-\bar{\mathbf{x}}).$$

Exercise 2.12

Show that $\langle \mathbf{x}-\bar{\mathbf{x}}, \mathbf{y} \rangle = \langle \mathbf{x}-\bar{\mathbf{x}}, \mathbf{y}-\bar{\mathbf{y}} \rangle$.

❧ ❋ ❧

We will use the notation $\sigma_{\mathbf{x}}^2$ to denote the empirical variance of $\mathbf{x} = (x_1, \ldots, x_n)$:

$$\begin{aligned} \sigma_{\mathbf{x}}^2 &:= \tfrac{1}{n} \sum_i (x_i - \bar{x})^2 \\ &= \tfrac{1}{n} \|\mathbf{x}-\bar{\mathbf{x}}\|^2 \end{aligned}$$

and likewise for $\sigma_{\mathbf{y}}$. The empirical *covariance* we will denote by

$$\begin{aligned} \sigma_{\mathbf{x},\mathbf{y}} &:= \tfrac{1}{n} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ &= \tfrac{1}{n} \langle \mathbf{x}-\bar{\mathbf{x}}, \mathbf{y}-\bar{\mathbf{y}} \rangle. \end{aligned}$$

Let $\rho_{\mathbf{x},\mathbf{y}}$ denote the *correlation* of $(x_1, y_1), \ldots, (x_n, y_n)$:

$$\begin{aligned} \rho_{\mathbf{x},\mathbf{y}} &:= \frac{\sigma_{\mathbf{x},\mathbf{y}}}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}} \\ &= \frac{(1/n)\langle \mathbf{x}-\bar{\mathbf{x}}, \mathbf{y}-\bar{\mathbf{y}} \rangle}{(\sqrt{1/n}\|\mathbf{x}-\bar{\mathbf{x}}\|)(\sqrt{1/n}\|\mathbf{y}-\bar{\mathbf{y}}\|)} \\ &= \frac{\langle \mathbf{x}-\bar{\mathbf{x}}, \mathbf{y}-\bar{\mathbf{y}} \rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\| \|\mathbf{y}-\bar{\mathbf{y}}\|}. \end{aligned}$$

Take note of the Euclidean interpretations of the quantities we've discussed here.

Exercise 2.13

Re-express $\frac{\langle \mathbf{y}, \mathbf{x}-\bar{\mathbf{x}} \rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}$ in terms of $\rho_{\mathbf{x},\mathbf{y}}$, $\sigma_{\mathbf{x}}$, and $\sigma_{\mathbf{y}}$.

Exercise 2.14

Inspect the orthogonal projection found in Exercise 2.11 to determine the slope and intercept of the least-squares line, and use Exercise 2.13 to express your answer.

◇

Solution: We substitute $\rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}$ for $\frac{\langle \mathbf{y}, \mathbf{x}-\bar{\mathbf{x}}\rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}$ and rearrange to produce

$$\begin{aligned}\hat{\mathbf{y}} &= \bar{\mathbf{y}} + \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}(\mathbf{x}-\bar{\mathbf{x}}) \\ &= \left(\bar{\mathbf{y}} - \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}\bar{\mathbf{x}}\right) + \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}\mathbf{x} \\ &= \left(\bar{y} - \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}\bar{x}\right)\mathbf{1} + \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}\mathbf{x}.\end{aligned}$$

Each entry of the first vector is equal to the least-squares line's intercept: $\bar{y} - \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}\bar{x}$; the coefficient of the second vector equals the least-squares line's slope: $\rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}$.

Rather than the empirical standard deviation and covariance, people sometimes prefer the *unbiased estimators* of variances and covariance which use $\frac{1}{n-1}$ rather[3] than $\frac{1}{n}$. These estimates can be also used in the formulas of the preceding exercises because the common factor cancels out anyway. Similarly, the correlation can be defined using either the empirical or the estimated quantities without making any difference.

If both $\mathbf{x}$ and $\mathbf{y}$ have been *standardized* by subtracting their averages and dividing by either their standard deviations, the least-squares line goes through the origin and has slope equal to the correlation.[4]

$$\begin{aligned}\hat{\mathbf{y}} &= \bar{\mathbf{y}} + \rho_{\mathbf{x},\mathbf{y}}\frac{\sigma_\mathbf{y}}{\sigma_\mathbf{x}}(\mathbf{x}-\bar{\mathbf{x}}) \\ &\Updownarrow \\ \frac{\hat{\mathbf{y}}-\bar{\mathbf{y}}}{\sigma_\mathbf{y}} &= \rho_{\mathbf{x},\mathbf{y}}\frac{\mathbf{x}-\bar{\mathbf{x}}}{\sigma_\mathbf{x}}\end{aligned} \tag{2.1}$$

In other words, if the explanatory variable is z standard deviations above its mean, then the least-squares line predicts the response variable to be $\rho_{\mathbf{x},\mathbf{y}}z$ standard deviations above its mean. Thus the response variable is predicted to be *less extreme* than the explanatory variable. Least-squares fitting was pioneered in the late nineteenth century by the eminent British intellectual Sir Francis Galton who first demonstrated it to predict men's heights using their fathers' heights. The men's heights were indeed found to be less extreme on average than their fathers' heights, a phenomenon called by Galton "regression toward mediocrity"[5]; the term *regression* caught on and is now used broadly for fitting quantitative response data.

2.2.3. Least-squares hyperplane

The visualizations and the reasoning we've seen in this section can be extended to the case of arbitrarily many explanatory variables. With m explanatory variables, we'll consider fitting the response vector with prediction vectors of the form

$$b_0\mathbf{1} + b_1\mathbf{x}^{(1)} + \ldots + b_m\mathbf{x}^{(m)}$$

[3] More precisely, these statistics are unbiased estimates of the variances and covariance between the random variables X and Y, assuming $(x_1, y_1), \ldots, (x_n, y_n)$ were iid draws from the joint distribution of (X, Y).

[4] This holds whether the empirical variances are used or the unbiased estimators are used; as before, the constant factor cancels itself out.

[5] Of course, individuals also have a chance to be more exceptional that their parents in any given characteristic, and as a result aggregate population characteristics remain relatively stable.

with $b_0, \ldots, b_m \in \mathbb{R}$. A more compact expression for these prediction vectors is $\mathbb{X}\mathbf{b}$ where[6]

$$\mathbb{X} := \begin{bmatrix} | & | & & | \\ \mathbf{1} & \mathbf{x}^{(1)} & \ldots & \mathbf{x}^{(m)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n\times(m+1)} \quad \text{and} \quad \mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{bmatrix}.$$

With the vector of coefficients $\mathbf{b}$ ranging over $\mathbb{R}^{m+1}$, the set of possible prediction vectors is precisely the span of the columns $\mathbf{1}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, which is a subspace of $\mathbb{R}^n$. The prediction vector minimizing the sum of squared residuals is exactly the one that is closest to $\mathbf{y}$ in Euclidean distance. Recalling our thinking from Section 2.1, if we envision the observations as data points in $\mathbb{R}^{m+1}$, each possible fit function is an m-dimensional hyperplane defined by $f_{\mathbf{b}}([x^{(1)}, \ldots, x^{(m)}]) = b_0 + b_1 x^{(1)} + \ldots + b_m x^{(m)}$ for any explanatory observation $(x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^m$. To be consistent with our earlier terminology, we can call the optimal such function a *least-squares hyperplane*, but it is much more common to say simply *least-squares fit*.

The vector of least-squares prediction is the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{X})$; as long as the columns are linearly independent we can use formulas from Chapter 1. By Exercise 1.63, the coefficients leading to the least-squares predictions are

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

while the vector of least-squares predictions is $\mathbb{X}\hat{\mathbf{b}}$ which can also be expressed $\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y}$.

Exercise 2.15

Assuming the columns of the design matrix are linearly independent, show that $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\hat{\mathbf{y}}$ is an alternative expression for the least-squares coefficient vector.

◇

Solution: There's an intuitive explanation for this. You can think of $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ as the matrix that maps any vector in $\mathbb{R}^n$ to the coefficients of the columns of $\mathbb{X}$ that lead to the orthogonal projection of that vector onto $C(\mathbb{X})$. Because the orthogonal projection of $\hat{\mathbf{y}}$ onto $C(\mathbb{X})$ is exactly the same as the orthogonal projection of $\mathbf{y}$ onto $C(\mathbb{X})$ (namely, both are $\hat{\mathbf{y}}$), the coefficients leading to this orthogonal projection must be the same.

By multiplying and dividing by n, we can also express the least-squares coefficients as

$$\begin{aligned} \hat{\mathbf{b}} &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y} \\ &= (\tfrac{1}{n}\mathbb{X}^T\mathbb{X})^{-1}(\tfrac{1}{n}\mathbb{X}^T\mathbf{y}). \end{aligned}$$

From Exercise 1.49, the (j,k) entry of $\frac{1}{n}\mathbb{X}^T\mathbb{X}$ is the average of the products of the jth and kth coordinates of the rows of $\mathbb{X}$. (This matrix is known as the *empirical second moments matrix* for the columns of $\mathbb{X}$.) The other vector in our expression, $\frac{1}{n}\mathbb{X}^T\mathbf{y}$, has as its jth entry the average of the products of the response values with the jth coordinate of the corresponding the row of $\mathbb{X}$.

When the first column of $\mathbb{X}$ is $\mathbf{1}$ (a case called *multiple linear regression*), we can observe the role of more familiar quantities. Instead of using the variables $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}$, consider using the centered variables $\mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}, \ldots, \mathbf{x}^{(d)} - \bar{\mathbf{x}}^{(d)}$, where $\bar{\mathbf{x}}^{(1)}, \ldots, \bar{\mathbf{x}}^{(d)}$ represent constant vectors for the variables' averages $\bar{x}^{(1)}, \ldots, \bar{x}^{(d)}$. The span of $\{\mathbf{1}, \mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}, \ldots, \mathbf{x}^{(d)} - \bar{\mathbf{x}}^{(d)}\}$ is exactly

[6] $\mathbb{X}$ is called the **design matrix** or the *model matrix*.

the same as the span of $\{\mathbf{1}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\}$, so using the centered vectors must result in the exact same orthogonal projection $\hat{\mathbf{y}}$. In fact, we can readily rewrite $\hat{\mathbf{y}}$ with the centered versions of the variables as the terms:

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{b}_0\mathbf{1} + \hat{b}_1\mathbf{x}^{(1)} + \ldots + \hat{b}_d\mathbf{x}^{(d)} \\ &= (\underbrace{\hat{b}_0 + \hat{b}_1\bar{\mathbf{x}}^{(1)} + \ldots + \hat{b}_d\bar{\mathbf{x}}^{(d)}}_{\text{“}\hat{a}_0\text{”}} + \hat{b}_1(\mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}) + \ldots + \hat{b}_d(\mathbf{x}^{(d)} - \bar{\mathbf{x}}^{(d)}). \qquad (2.2)\end{aligned}$$

We observe that the least-squares coefficients of the centered variables are exactly the same as the least-squares coefficients of the original variables; only the least-squares constant changes when the centered variables are used. Based on this, we realize that we can find the least-squares coefficients by using the centered variables in our formula. When we do this, the second moments become covariances. With $\boldsymbol{\Sigma}$ denoting the empirical covariance matrix (see Section 3.2) for the explanatory variables,

$$\begin{aligned}\begin{bmatrix}\hat{a}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_d\end{bmatrix} &= \begin{bmatrix}1 & 0 \\ 0 & \boldsymbol{\Sigma}\end{bmatrix}^{-1} \begin{bmatrix}\frac{1}{n}\langle \mathbf{y}, \mathbf{1}\rangle \\ \frac{1}{n}\langle \mathbf{y}, \mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}\rangle \\ \vdots \\ \frac{1}{n}\langle \mathbf{y}, \mathbf{x}^{(d)} - \bar{\mathbf{x}}^{(d)}\rangle\end{bmatrix} \\ &= \begin{bmatrix}1 & 0 \\ 0 & \boldsymbol{\Sigma}^{-1}\end{bmatrix} \begin{bmatrix}\bar{y} \\ \sigma_{\mathbf{y},\mathbf{x}^{(1)}} \\ \vdots \\ \sigma_{\mathbf{y},\mathbf{x}^{(d)}}\end{bmatrix}\end{aligned}$$

using the covariance trick from Exercise 2.12. (You can easily verify the matrix inverse step by realizing that block diagonal matrices multiply block-by-block.) We can also write this solution as $\hat{a}_0 = \bar{y}$ and

$$\begin{aligned}\begin{bmatrix}\hat{b}_1 \\ \vdots \\ \hat{b}_d\end{bmatrix} &= \boldsymbol{\Sigma}^{-1} \begin{bmatrix}\sigma_{\mathbf{y},\mathbf{x}^{(1)}} \\ \vdots \\ \sigma_{\mathbf{y},\mathbf{x}^{(d)}}\end{bmatrix} \\ &= \boldsymbol{\Sigma}^{-1}(\tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{y})\end{aligned}$$

where $\widetilde{\mathbb{X}}$ is the matrix whose columns are the centered versions of the explanatory vectors. Finally, making reference to Equation 2.2, the least-squares constant coefficient for the original uncentered variables can then be found:

$$\hat{b}_0 = \bar{y} - (\hat{b}_1\bar{x}^{(1)} + \ldots + \hat{b}_d\bar{x}^{(d)}).$$

We will continue to draw pictures to represent the variable vectors in $\mathbb{R}^n$, but at this point we need to think more carefully about how to depict the vectors and subspaces accurately. Figures 2.11 and 2.12 are the prime examples of pictures with potentially more than one explanatory variable, and we will consider each figure in turn.

Let's first look at Figure 2.11 and ask ourselves how accurate it is. We know that there exists *some* three-dimensional subspace that includes $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\bar{\mathbf{y}}$ (in addition to the origin). The question is, does $\text{span}\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ (which we'll call $\mathcal{S}$) intersect this subspace in a *line*? Assume that $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\bar{\mathbf{y}}$ are distinct vectors and that $\mathbf{1}$ is not in the span of the explanatory variables; typically, this is indeed the case. Because $\hat{\mathbf{y}}$ is in the span of $\{\mathbf{1}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$, we know that it can be represented as a linear combination

$$\hat{\mathbf{y}} = \hat{b}_0\mathbf{1} + \hat{b}_1\mathbf{x}^{(1)} + \ldots + \hat{b}_m\mathbf{x}^{(m)}.$$

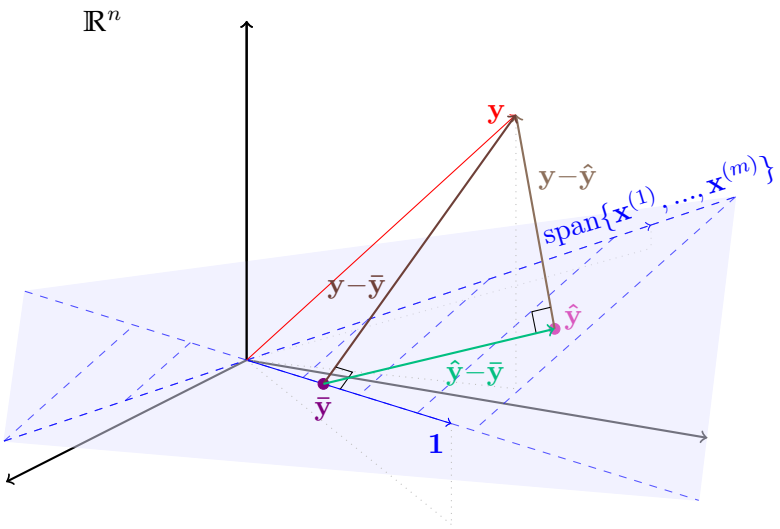


Figure 2.11: A generic picture showing the constant subspace (the span of $\mathbf{1}$*) and the response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dashed lines represent a portion of the span of* $\mathbf{1}$ *and the explanatory variables together.* $\hat{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto that subspace, and* $\mathbf{y}-\hat{\mathbf{y}}$ *is orthogonal to it.* $\bar{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto the span of* $\mathbf{1}$*, and* $\mathbf{y}-\bar{\mathbf{y}}$ *is orthogonal to it. In general, the span of the explanatory variables has a one-dimensional intersection with this three-dimensional perspective.*

Let $\mathbf{v} := \hat{b}_1\mathbf{x}^{(1)} + \ldots + \hat{b}_m\mathbf{x}^{(m)}$, and note that it's in $\mathcal{S}$. Because the $\hat{b}_0\mathbf{1}$ and $\hat{\mathbf{y}}$ are both in our three-dimensional picture, so is $\mathbf{v} = \hat{\mathbf{y}} - \hat{b}_0\mathbf{1}$. And because $\mathbf{v}$ is in our three-dimensional subspace, so is $a\mathbf{v}$ for every real a. Therefore, there is *at least* a line of intersection with $\mathcal{S}$ in our picture. Could the intersection be *more* than a line? If it were a plane that didn't include $\mathbf{1}$, then $\text{span}\{\mathbf{1}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ would occupy all three dimensions in our picture and would therefore include $\mathbf{y}$, but that would contradict our assumption that $\hat{\mathbf{y}}$ and $\mathbf{y}$ are distinct.

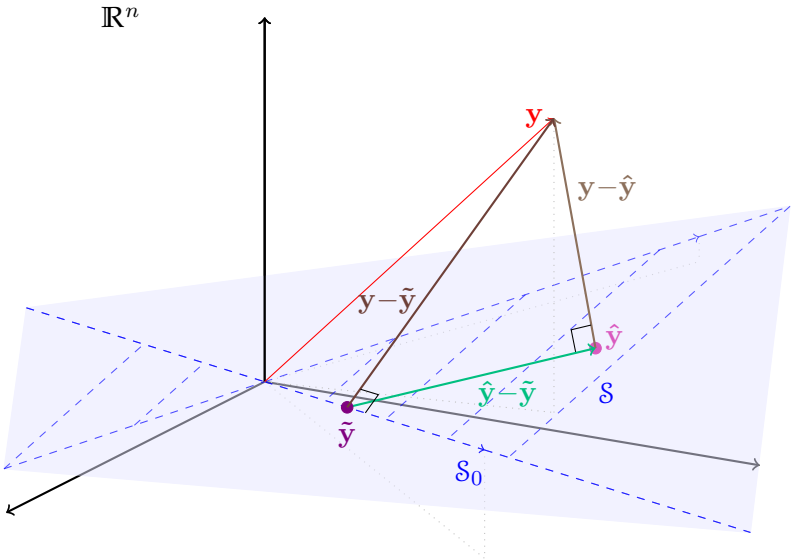


Figure 2.12: A generic picture showing a response variable vector $\mathbf{y}$ *in* $\mathbb{R}^n$*, its orthogonal projection* $\tilde{\mathbf{y}}$ *onto a subspace* $\mathcal{S}_0$*, and its orthogonal projection* $\hat{\mathbf{y}}$ *onto a subspace* $\mathcal{S} \supseteq \mathcal{S}_0$*. In general,* $\mathcal{S}_0$ *intersects this three-dimensional perspective in a line, and* $\mathcal{S}$ *intersects it in a plane. The dashed lines represent a portion of* $\mathcal{S}$*.*

Next, how accurate is Figure 2.12? We know that there exists *some* three-dimensional subspace that includes $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\tilde{\mathbf{y}}$ (in addition to the origin). The question is, does $\mathcal{S}_0$ intersect this subspace in a *line*, and does $\mathcal{S}$ intersect it in a plane? Assume that $\mathbf{y}$, $\tilde{\mathbf{y}}$ and $\hat{\mathbf{y}}$ are distinct vectors. $\{a\tilde{\mathbf{y}} : a \in \mathbb{R}\} \subseteq \mathcal{S}_0$ is in our three-dimensional pictures, so the intersection with $\mathcal{S}_0$ is at least a line. And $\text{span}\{\tilde{\mathbf{y}}, \hat{\mathbf{y}}\} \subseteq \mathcal{S}$ is a plane that includes the line $\{a\tilde{\mathbf{y}} : a \in \mathbb{R}\}$. If the intersection with $\mathcal{S}$ was three-dimensional, then the entire picture would be in $\mathcal{S}$ which contradicts the assumption that $\hat{\mathbf{y}}$ and $\mathbf{y}$ are distinct. Because we've assumed $\tilde{\mathbf{y}} \neq \hat{\mathbf{y}}$, we can conclude that this plane is not in $\mathcal{S}_0$. Finally, if the intersection of this three-dimensional perspective with $\mathcal{S}_0$ had another dimension (outside of $\text{span}\{\tilde{\mathbf{y}}, \hat{\mathbf{y}}\}$), this would also imply that the intersection with $\mathcal{S}$ is three-dimensional.

What if the vectors and subspaces of interest aren't linearly independent in the ways we

assumed above? In that case, the true picture *may* differ from our depiction. For example, if $\mathbf{y} \in \mathcal{S}$, then $\hat{\mathbf{y}} = \mathbf{y}$. It's good to be aware of these types of possibilities, but our drawings represent the typical case of linearly independence. And even when linear dependence causes the picture to be imperfect, the conclusions drawn from the picture are often still valid. To be completely thorough, one must think through each possible way in which the picture can differ from reality and make sure that the result holds in those cases.

Based on the intuition you're developing by working through this chapter, hopefully it's becoming clear to you that *whenever the set of fits under consideration comprises a subspace of* $\mathbb{R}^n$*, the fit that minimizes the sum of squared residuals is precisely the orthogonal projection of the data onto that subspace.*

Exercise 2.16

Let $\mathbf{y} \in \mathbb{R}^n$ be a response variable vector and $\mathbf{x} \in \mathbb{R}^n$ be an explanatory variable vector. Consider fitting the response variable by using quadratic functions of the explanatory variable:

$$\{f_{a,b,c}(x) = a + bx + cx^2 : a, b, c \in \mathbb{R}\}.$$

Show that the set of possible vectors of fitted values is a subspace of $\mathbb{R}^n$.

Exercise 2.17

Let $\mathbf{y} \in \mathbb{R}^n$ be a response variable vector and $\mathbf{x} \in \mathbb{R}^n$ be an explanatory variable vector. Consider fitting the response variable by using quadratic functions of the explanatory variable:

$$\{f_{a,b,c}(x) = a + bx + cx^2 : a, b, c \in \mathbb{R}\}.$$

Explain how to find the coefficients $(\hat{a}, \hat{b}, \hat{c})$ of the quadratic function that minimizes the sum of squared residuals.

❧ ❋ ❧

As Exercises 2.16 and 2.17 demonstrate, you aren't limited to the explanatory variables you started with: you can use any functions of those variables to construct terms for your fit. As long as the free parameters are exactly the coefficients of the terms, the set of possible fits is a subspace of $\mathbb{R}^n$. The set of possible functions should take the general form

$$\{f_{b_0,\ldots,b_d}(x^{(1)}, \ldots, x^{(m)}) := b_0 g_0(x^{(1)}, \ldots, x^{(m)}) + \ldots + b_d g_d(x^{(1)}, \ldots, x^{(m)}) : (b_0, \ldots, b_d) \in \mathbb{R}^{d+1}\}.$$

After constructing the design matrix $\mathbb{X}$ with the desired terms as its columns

$$\mathbb{X} := \begin{bmatrix} g_0(x_1^{(1)}, \ldots, x_1^{(m)}) & \ldots & g_d(x_1^{(1)}, \ldots, x_1^{(m)}) \\ \vdots & \ddots & \vdots \\ g_0(x_n^{(1)}, \ldots, x_n^{(m)}) & \ldots & g_d(x_n^{(1)}, \ldots, x_n^{(m)}) \end{bmatrix},$$

the least-squares coefficients are $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{y}$ if the columns are linearly independent.

This idea can also be used for least-squares fitting involving categorical explanatory variables. With a single categorical explanatory variable taking the values $\{1, \ldots, k\}$, a natural set of functions for fitting the data is

$$\{f_{b_1,\ldots,b_k}(x) = b_1 \mathbb{I}(x = 1) + \ldots + b_k \mathbb{I}(x = k) : (b_1, \ldots, b_k) \in \mathbb{R}^k\}.$$

In other words, all the observations of group j will be predicted by some number b_j. We know from Section 2.1.1 that the least-squares coefficients must be simply the groups' averages. What does the corresponding design matrix look like? With $x_i = j$, the ith row of the design matrix has a 1 in column j and zeros elsewhere. Notice that the columns are all orthogonal to each other, so linear independence of the columns always holds in this context.

When both quantitative and categorical variables are available, more complex sets of possible fits can be defined involving both types as we will see later in Homework 4.

Homework 3: Least-squares fitting

Let's use what we've learned to do some least-squares fitting with real data: work through *least-squares-fitting.Rmd*.

2.3. Decomposing sums of squares

With the variables picture in mind, we derived least-squares coefficients and predictions by understanding them in terms of orthogonal projection theory which was covered at some length in Chapter 1. Closely related to this, another benefit of the variables picture is that it enables us to easily derive a variety of useful decompositions of certain sums of squares by observing right triangles in the picture and simply invoking the Pythagorean identity.

Exercise 2.18

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$, and let $\tilde{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0 \subseteq \mathcal{S}$. Use the result of Exercise 1.74 to verify that $\hat{\mathbf{y}} - \tilde{\mathbf{y}}$ is orthogonal to $\mathcal{S}_0$.

⋄

Solution: The vector $\tilde{\mathbf{y}}$ is defined to be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0$. However, it's also the orthogonal projection of $\hat{\mathbf{y}}$ onto $\mathcal{S}_0$ because according to Exercise 1.74, orthogonal projection onto $\mathcal{S}$ followed by orthogonal projection onto $\mathcal{S}_0$ lands you at the exact same vector that a single orthogonal projection onto $\mathcal{S}_0$ does. Finally, $\hat{\mathbf{y}}$ minus its orthogonal projection onto $\mathcal{S}_0$ is orthogonal to $\mathcal{S}_0$.

Exercise 2.19

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$. With $\mathbf{w} \in \mathcal{S}$, use the Pythagorean identity to decompose the squared length of $\mathbf{y} - \mathbf{w}$.

Exercise 2.20

Let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$, and suppose $\mathbf{1} \in \mathcal{S}$. Justify the *ANOVA decomposition*:

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2.$$

⋄

Solution: We know that $\bar{\mathbf{y}}$ is in $\mathcal{S}$. By Exercise 2.19

$$\|\mathbf{y} - \bar{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 + \|\mathbf{y} - \hat{\mathbf{y}}\|^2$$

which is exactly the ANOVA decomposition written in vector notation.

In the case of one categorical explanatory variable, the ANOVA decomposition (Exercise 2.20) is often written by summing over the groups rather than the observations. With k groups, we let $\bar{y}_1, \dots, \bar{y}_k$ represent the groups' means and $n_1, \dots, n_k$ represent the number of observations in each group. The least-squares procedure predicts every observation by its group's mean, so

$$\begin{aligned}\sum_i (y_i - \bar{y})^2 &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \\ &= \sum_i (\bar{y}_{x_i} - \bar{y})^2 + \sum_i (y_i - \bar{y}_{x_i})^2 \\ &= \sum_{j=1}^{k} n_j (\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k} \sum_{i:x_i=j} (y_i - \bar{y}_j)^2.\end{aligned}$$

The terms of the ANOVA decomposition are sometimes called the *total sum of squares*, *regression sum of squares*, and *residual sum of squares*, respectively.[7] The fraction $\frac{\|\hat{\mathbf{y}}-\bar{\mathbf{y}}\|^2}{\|\mathbf{y}-\bar{\mathbf{y}}\|^2}$ is called the R^2 of the regression; it represents the proportion of the total sum of squares that was "explained" by the explanatory variables. Notice that this fraction is the squared cosine of the angle between $\mathbf{y} - \bar{\mathbf{y}}$ and $\hat{\mathbf{y}} - \bar{\mathbf{y}}$. When there is only one explanatory variable, $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ is in the direction of $\mathbf{x} - \bar{\mathbf{x}}$, as we can see in Equation 2.1. So in that case R^2 is the squared cosine of the angle between $\mathbf{y} - \bar{\mathbf{y}}$ and $\mathbf{x} - \bar{\mathbf{x}}$ which, recalling the discussion of correlation in Section 2.2.2, means that the R^2 is the squared correlation between $\mathbf{y}$ and $\mathbf{x}$.[8]

When there are multiple variables, the regression sum of squares can generally be further decomposed. Note that the logic behind Exercise 2.21 can be extended to any number of nested subspaces.

Exercise 2.21

Suppose $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}$, $\tilde{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto $\mathcal{S}_0 \subseteq \mathcal{S}$, and that $\mathbf{1} \in \mathcal{S}_0$. Explain why

$$\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = \|\hat{\mathbf{y}} - \tilde{\mathbf{y}}\|^2 + \|\tilde{\mathbf{y}} - \bar{\mathbf{y}}\|^2.$$

We'll conclude this chapter by working through another important decomposition of $\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2$ that occurs in the context of categorical explanatory variables. The results from this discussion will be used in Chapter 6.

Let $\mathbf{x}$ be a categorical explanatory variable taking values $\{1, \dots, k\}$ with every group having n/k observations.[9] In this context, any vector $\mathbf{c} = (c_1, \dots, c_k) \in \mathbb{R}^k$ that has mean zero is called a **contrast**; its k entries correspond to the k groups. Define $\mathbf{c}_\mathbf{x} := (c_{x_1}, \dots, c_{x_n}) \in \mathbb{R}^n$, the vector

[7] ANOVA stands for analysis of variance. By dividing both sides of the equation by n, you can interpret the quantities as the empirical variance of $\mathbf{y}$, the empirical variance of $\hat{\mathbf{y}}$ (which has mean $\bar{\mathbf{y}}$ according to Exercise 1.74) and the empirical variance of the residual vector $\mathbf{y} - \hat{\mathbf{y}}$ (which has mean zero because they are orthogonal to $\mathbf{1}$).

[8] Galton used the letter R for the "regression coefficient" in (2.1) which we now call the *correlation*. In simple linear regression, R^2 is exactly the squared correlation, which is why the statistic is called R^2. Note that with more than one explanatory variable, this interpretation no longer works.

[9] When all groups have the same number of observations, they are called *balanced*.

whose ith entry is equal to the contrast value for the ith observation's group.

Exercise 2.22

Let $\mathbf{x}$ be a categorical variable of k balanced groups, and let $\mathbf{c} = (c_1, \ldots, c_k) \in \mathbb{R}^k$ has 0 as the average of its entries. Show that $\mathbf{c}_\mathbf{x} := (c_{x_1}, \ldots, c_{x_n}) \in \mathbb{R}^n$ also has mean 0.

Exercise 2.23

Let $\mathbf{x}$ be a categorical variable of k balanced groups, and let $\mathbf{c}^{(1)} := (c_1^{(1)}, \ldots, c_k^{(1)}) \perp \mathbf{c}^{(2)} := (c_1^{(2)}, \ldots, c_k^{(2)})$ in $\mathbb{R}^k$. Show that $\mathbf{c}_\mathbf{x}^{(1)} := (c_{x_1}^{(1)}, \ldots, c_{x_n}^{(1)})$ is orthogonal to $\mathbf{c}_\mathbf{x}^{(2)} := (c_{x_1}^{(2)}, \ldots, c_{x_n}^{(2)})$ in $\mathbb{R}^n$.

❧ ❋ ❧

Let $\mathbf{z}_\mathbf{x}^{(j)} := (\mathbb{I}(x_1 = j), \ldots, \mathbb{I}(x_n = j))$ denote the jth column of the design matrix $\mathbb{X}$ as described at the end of Section 2.2.3. Observe that the sum of the columns of $\mathbb{X}$ equals the $\mathbf{1}$ vector. Furthermore, $\mathbb{X}\mathbf{c}^{(j)} = \mathbf{c}_\mathbf{x}^{(j)}$, so each of $\mathbf{c}_\mathbf{x}^{(1)}, \ldots, \mathbf{c}_\mathbf{x}^{(k-1)}$ is also in $C(\mathbb{X})$. Based on the results of Exercises 2.22 and 2.23, the k vectors $\{\mathbf{1}, \mathbf{c}_\mathbf{x}^{(1)}, \ldots, \mathbf{c}_\mathbf{x}^{(k-1)}\}$ are all orthogonal to each other and therefore linear independent so we can conclude that they must be a basis for the k-dimensional subspace $C(\mathbb{X})$. We can divide each of these vectors by its length to obtain an orthonormal basis. The squared length of $\mathbf{c}_\mathbf{x}^{(j)}$ is

$$\begin{aligned}
\|\mathbf{c}_\mathbf{x}^{(j)}\|^2 &= \sum_i c_{x_i}^2 \\
&= \sum_{j=1}^{k} \frac{n}{k} c_j^2 \\
&= \frac{n}{k} \underbrace{\sum_{j=1}^{k} c_j^2}_{\|\mathbf{c}^{(j)}\|^2}
\end{aligned}$$

which is n times the average of the squared values of $\mathbf{c}^{(j)}$. In particular, $\{\frac{\mathbf{c}_\mathbf{x}^{(1)}}{\sqrt{n/k}\|\mathbf{c}^{(1)}\|}, \ldots, \frac{\mathbf{c}_\mathbf{x}^{(k-1)}}{\sqrt{n/k}\|\mathbf{c}^{(k-1)}\|}\}$ comprise an orthonormal basis for the orthogonal complement of $\mathbf{1}$ in $C(\mathbb{X})$.

Based on the observation of Exercise 1.57, one can determine that $\hat{\mathbf{y}} - \bar{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the orthogonal complement of $\mathbf{1}$ in $C(\mathbb{X})$. From Exercise 1.60, this same orthogonal projection can also be expressed using the orthonormal basis that we've just derived.

$$\hat{\mathbf{y}} - \bar{\mathbf{y}} = \langle \mathbf{y}, \frac{\mathbf{c}_\mathbf{x}^{(1)}}{\sqrt{n/k}\|\mathbf{c}^{(1)}\|} \rangle \frac{\mathbf{c}_\mathbf{x}^{(1)}}{\sqrt{n/k}\|\mathbf{c}^{(1)}\|} + \ldots + \langle \mathbf{y}, \frac{\mathbf{c}_\mathbf{x}^{(k)}}{\sqrt{n/k}\|\mathbf{c}^{(k)}\|} \rangle \frac{\mathbf{c}_\mathbf{x}^{(k-1)}}{\sqrt{n/k}\|\mathbf{c}^{(k-1)}\|} \tag{2.3}$$

Let's work out an alternative expression for the inner product between $\mathbf{c}_\mathbf{x}^{(1)}$ and $\mathbf{y}$ by summing

over the groups:

$$\begin{aligned}\langle \mathbf{y}, \mathbf{c}_{\mathbf{x}}^{(1)} \rangle &= \sum_i c_{x_i}^{(1)} y_i \\ &= \sum_{j=1}^{k} c_j^{(1)} \underbrace{\sum_{i:x_i=j} y_i}_{(n/k)\bar{y}_j} \\ &= (n/k) \sum_{j=1}^{k} c_j^{(1)} \bar{y}_j.\end{aligned}$$

Because $\frac{\mathbf{c}_{\mathbf{x}}^{(1)}}{\sqrt{n/k}\|\mathbf{c}^{(1)}\|}, \ldots, \frac{\mathbf{c}_{\mathbf{x}}^{(k-1)}}{\sqrt{n/k}\|\mathbf{c}^{(k-1)}\|}$ are orthonormal, Equation 2.3 implies

$$\begin{aligned}\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 &= \left(\frac{\langle \mathbf{y}, \mathbf{c}_{\mathbf{x}}^{(1)} \rangle}{\sqrt{n/k}\|\mathbf{c}^{(1)}\|} \right)^2 + \ldots + \left(\frac{\langle \mathbf{y}, \mathbf{c}_{\mathbf{x}}^{(k-1)} \rangle}{\sqrt{n/k}\|\mathbf{c}^{(k-1)}\|} \right)^2 \\ &= \frac{(n/k)(\sum_{j=1}^{k} c_j^{(1)} \bar{y}_j)^2}{\|\mathbf{c}^{(1)}\|^2} + \ldots + \frac{(n/k)(\sum_{j=1}^{k} c_j^{(k-1)} \bar{y}_j)^2}{\|\mathbf{c}^{(k-1)}\|^2}.\end{aligned}$$

Homework 4: Sums of squares

Next, you'll tackle more challenging least-squares regression tasks and verify some of this section's decompositions: try *sums-of-squares.Rmd*.

In Chapter 4, we'll make certain *modeling assumptions* (statements about a probability distribution generating the data), and see how to incorporate them into our visualizations. But first, Chapter 3 will discuss random vectors and some of their key properties.

CHAPTER

3

REVIEW: RANDOM VECTORS

In a probability class, you learn that a *random variable* is a function from a sample space to the real numbers that has a *distribution* on $\mathbb{R}$. You also learn that multiple random variables can have a probabilistic relationship to each other that is defined by their *joint distribution*. Any n random variables with a joint distribution can be considered as the n entries of a vector, called an $\mathbb{R}^n$-valued **random vector**. We will continue to denote non-random numbers and vectors with lower-case letters, but we will use capital letters for random variables and vectors.

In general, results from Chapters 1 and 2 that were true for arbitrary vectors $(y_1, \ldots, y_n)$ also hold[1] for random vectors $(Y_1, \ldots, Y_n)$. For example, the sum of squared differences from a point a is now a random variable $\sum(Y_i - a)^2$, and it's minimized on the entire sample space (i.e. for every possible realization of $Y_1, \ldots, Y_n$) by replacing a with the random variable $\bar{Y} := \frac{1}{n}\sum Y_i$.

So far, we've only worried about approximating or fitting data; we've not asked *why* the data looks like it does. Moving forward, we'll use probabilistic modeling to consider possible mechanisms generating the data. In this chapter, we'll briefly learn how to work with random vectors; they will be the building blocks of our later probabilistic modeling.

3.1. Bias-variance decomposition

The *expectation* of the random vector $\mathbf{Y} := (Y_1, \ldots, Y_n)$ is defined to be $\mathbb{E}\mathbf{Y} = (\mathbb{E}Y_1, \ldots, \mathbb{E}Y_n)$, that is, the coordinates of the expectation vector are simply the expectations of the coordinate random variables.

Exercise 3.1

Let $\mathbf{Y}$ be a random vector and $\mathbf{v}$ be a non-random vector. Show that $\mathbb{E}\langle \mathbf{v}, \mathbf{Y}\rangle = \langle \mathbf{v}, \mathbb{E}\mathbf{Y}\rangle$.

[1] The results are true *point-wise*, meaning that they are true whenever all of the random variables are evaluated at the same point ω in the sample space Ω.

Exercise 3.2

Let $\mathbf{Y}$ be a random vector, $\mathbf{v}$ be a non-random vector, and $\mathbb{M}$ be a matrix. Show that

$$\mathbb{E}(\mathbf{v} + \mathbb{M}\mathbf{Y}) = \mathbf{v} + \mathbb{M}\mathbb{E}\mathbf{Y}.$$

Exercise 3.3

If $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is a random vector, its expected squared length is equal to the sum of its expected squared entries:

$$\begin{aligned}\mathbb{E}\|\mathbf{Y}\|^2 &= \mathbb{E}(Y_1^2 + \ldots + Y_n^2) \\ &= \mathbb{E}Y_1^2 + \ldots + \mathbb{E}Y_n^2.\end{aligned}$$

Verify the more general fact that the expected squared length of a random vector equals the sum of its expected squared coordinates with respect to any orthonormal basis.

Exercise 3.4

Let $\mathbf{Y}$ be a random vector with expectation $\boldsymbol{\mu}$, and let $\mathbf{v}$ be a non-random vector. Prove the **bias-variance decomposition**:

$$\mathbb{E}\|\mathbf{Y} - \mathbf{v}\|^2 = \|\mathbf{v} - \boldsymbol{\mu}\|^2 + \mathbb{E}\|\mathbf{Y} - \boldsymbol{\mu}\|^2.$$

(The terminology comes from thinking of $\mathbf{Y}$ as an estimator for $\mathbf{v}$. The first term is a generalization of the squared bias of $\mathbf{Y}$ for $\mathbf{v}$, while the second term is a generalization of the variance of $\mathbf{Y}$.)

◇

Solution: Adding and subtracting $\boldsymbol{\mu}$, then writing the norm in inner product form,

$$\begin{aligned}\mathbb{E}\|\mathbf{Y} - \mathbf{v}\|^2 &= \mathbb{E}\|(\mathbf{Y} - \boldsymbol{\mu}) - (\mathbf{v} - \boldsymbol{\mu})\|^2 \\ &= \mathbb{E}\langle(\mathbf{Y} - \boldsymbol{\mu}) - (\mathbf{v} - \boldsymbol{\mu}), (\mathbf{Y} - \boldsymbol{\mu}) - (\mathbf{v} - \boldsymbol{\mu})\rangle \\ &= \mathbb{E}[\langle\mathbf{Y} - \boldsymbol{\mu}, \mathbf{Y} - \boldsymbol{\mu}\rangle - 2\langle\mathbf{Y} - \boldsymbol{\mu}, \mathbf{v} - \boldsymbol{\mu}\rangle + \langle\mathbf{v} - \boldsymbol{\mu}, \mathbf{v} - \boldsymbol{\mu}\rangle] \\ &= \mathbb{E}\|\mathbf{Y} - \boldsymbol{\mu}\|^2 - 2\langle\underbrace{\mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})}_{\mathbf{0}}, \mathbf{v} - \boldsymbol{\mu}\rangle + \mathbb{E}\,\underbrace{\|\mathbf{v} - \boldsymbol{\mu}\|^2}_{\text{non-random}} \\ &= \mathbb{E}\|\mathbf{Y} - \boldsymbol{\mu}\|^2 + \|\mathbf{v} - \boldsymbol{\mu}\|^2.\end{aligned}$$

Exercise 3.5

Let $\mathbf{Y}$ be a random vector with expectation $\boldsymbol{\mu}$. Find the non-random vector $\mathbf{v}$ that minimizes $\mathbb{E}\|\mathbf{Y} - \mathbf{v}\|^2$.

◇

Solution: By the bias-variance decomposition, the objective function equals $\|\mathbf{v} - \boldsymbol{\mu}\|^2 + \mathbb{E}\|\mathbf{Y} - \boldsymbol{\mu}\|^2$. The second term doesn't depend on $\mathbf{v}$, so we can minimize the sum by taking $\mathbf{v}$ to be $\boldsymbol{\mu}$ which makes the first term zero.

Exercise 3.6

Let $\mathbf{Y}$ be a random vector that is an *unbiased estimator* for $\boldsymbol{\theta} \in \mathbb{R}^n$, that is $\mathbb{E}\mathbf{Y} = \boldsymbol{\theta}$. If $\lambda \in \mathbb{R}$, find a simple expression for $\|\mathbb{E}(\lambda\mathbf{Y}) - \boldsymbol{\theta}\|^2$, which can be thought of as the *squared bias* of the estimator $\lambda\mathbf{Y}$.

Exercise 3.7

Let $\mathbb{X}$ be a random vector, and let $\lambda \in \mathbb{R}$. Find $\mathbb{E}\|\lambda\mathbf{Y} - \mathbb{E}(\lambda\mathbf{Y})\|^2$ in terms of $\mathbb{E}\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2$.

Exercise 3.8

Let $\mathbf{Y}$ be a random vector that is an *unbiased estimator* for $\boldsymbol{\theta} \in \mathbb{R}^n$. Use the bias-variance decomposition along with your results from Exercises 3.6 and 3.7 to find an expression for $\lambda \in \mathbb{R}$ (in terms of $\|\boldsymbol{\theta}\|^2$ and $\mathbb{E}\|\mathbf{Y} - \boldsymbol{\theta}\|^2$) for which $\mathbb{E}\|\boldsymbol{\theta} - \lambda\mathbf{Y}\|^2$ is as small as possible.

Exercise 3.9

Explain how Exercise 2.6 is an instance of the bias-variance decomposition.

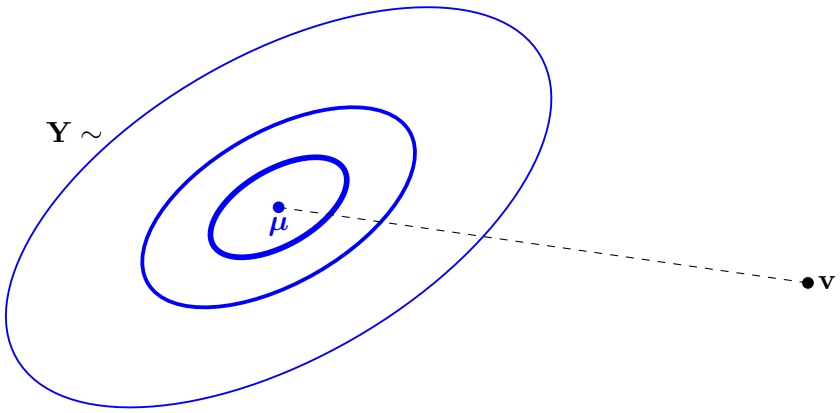


Figure 3.1: The bias-variance decomposition (Exercise 3.4) says that the expected squared distance from $\mathbf{Y}$ to $\mathbf{v}$ is the sum of the expected squared distance from $\mathbf{Y}$ to its expectation $\boldsymbol{\mu}$ plus the squared distance from $\boldsymbol{\mu}$ to $\mathbf{v}$.

It's good to realize that, mathematically speaking, you can't take the existence of expectations in $\mathbb{R}^n$ for granted. Our various results involving expectations of random vectors will depend on assuming that they're finite.

Exercise 3.10

Let X be a discrete random variable whose possible values are the positive integers. In particular, suppose that $\mathbb{P}\{X = k\}$ is proportional to $1/k^2$ for $k \in \{1, 2, \ldots\}$. What is the expected value of X?

3.2. Covariance

The **covariance matrix** of $\mathbf{Y}$ is the $n \times n$ matrix whose (i,j)-entry equals the *covariance* between Y_i and Y_j which is defined to be $\mathbb{E}[(Y_i - \mathbb{E}Y_i)(Y_j - \mathbb{E}Y_j)]$. Notice from this definition that covariance matrices must be symmetric. Notice also that the diagonals are the variances of the coordinate random variables.

The same notation that we used for expectations of random vectors will be applied to matrices with random entries as well: the expectation of the matrix is the matrix of its expected entries.

Exercise 3.11

Show that an alternative expression for the covariance of $\mathbf{Y}$ is $\mathbb{E}[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T]$.

Exercise 3.12

Show that every covariance matrix is positive definite.

◇

Solution: To satisfy the definition of positive definiteness, we need to show that every quadratic form is non-negative. We'll use the covariance expression from Exercise 3.11 and consider its quadratic form for an arbitrary vector $\mathbf{v}$,

$$\begin{aligned}\mathbf{v}^T\mathbb{E}[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T]\mathbf{v} &= \mathbb{E}[\mathbf{v}^T(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T]\mathbf{v} \\ &= \mathbb{E}[\mathbf{v}^T(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T\mathbf{v}] \\ &= \mathbb{E}\langle \mathbf{v}, \mathbf{Y} - \mathbb{E}\mathbf{Y}\rangle^2.\end{aligned}$$

The expectation of a non-negative random variable has to be non-negative.

Exercise 3.13

Let $\mathbf{Y}$ be a random vector with covariance matrix $\mathbb{C}$. Let $\mathbf{v}$ be a non-random vector, and let $\mathbb{M}$ be a real matrix. Use the expression of the covariance matrix from Exercise 3.11 to find a formula for the covariance of $\mathbf{v} + \mathbb{M}\mathbf{Y}$ in terms of $\mathbb{C}$.

◇

Solution: By Exercise 3.2, we know that the expectation of $\mathbf{v} + \mathbb{M}\mathbf{Y}$ equals $\mathbf{v} + \mathbb{M}\mathbb{E}\mathbf{Y}$.

$$\begin{aligned}\operatorname{cov}(\mathbf{v} + \mathbb{M}\mathbf{Y}) &= \mathbb{E}[(\mathbf{v} + \mathbb{M}\mathbf{Y} - \mathbb{E}(\mathbf{v} + \mathbb{M}\mathbf{Y}))(\mathbf{v} + \mathbb{M}\mathbf{Y} - \mathbb{E}(\mathbf{v} + \mathbb{M}\mathbf{Y}))^T] \\ &= \mathbb{E}[(\mathbb{M}\mathbf{Y} - \mathbb{M}\mathbb{E}\mathbf{Y})(\mathbb{M}\mathbf{Y} - \mathbb{M}\mathbb{E}\mathbf{Y})^T] \\ &= \mathbb{E}[(\mathbb{M}(\mathbf{Y} - \mathbb{E}\mathbf{Y}))(\mathbb{M}(\mathbf{Y} - \mathbb{E}\mathbf{Y}))^T] \\ &= \mathbb{M}(\mathbb{E}[(\mathbf{Y} - \mathbb{E}\mathbf{Y})(\mathbf{Y} - \mathbb{E}\mathbf{Y})^T])\mathbb{M}^T \\ &= \mathbb{M}\mathbb{C}\mathbb{M}^T\end{aligned}$$

Exercise 3.14

Let $\mathbf{Y}$ be a random vector with covariance matrix $\sigma^2\mathbb{I}$. Let $\mathbf{v}$ be a non-random vector, and let $\mathbb{H}$ be an orthogonal projection matrix. Use the result of Exercise 3.13 to find the covariance of $\mathbf{v} + \mathbb{H}\mathbf{Y}$.

◇

Solution: According to our result, the covariance is $\mathbb{H}(\sigma^2\mathbb{I})\mathbb{H}^T = \sigma^2\mathbb{H}\mathbb{H}\mathbb{H}^T$. By symmetry and idempotence of orthogonal projection matrices, this simplifies to $\sigma^2\mathbb{H}$.

Exercise 3.15

Let $\mathbf{Y}$ have expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$, and let $\mathbb{C} = \lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T$ be a spectral decomposition with $\lambda_1 \geq \ldots \geq \lambda_n$. Find the unit vector $\mathbf{u}$ for which the variance of the coordinate $\langle \mathbf{Y} - \boldsymbol{\mu}, \mathbf{u}\rangle$ is maximized. (It's the coordinate of the centered random vector $\mathbf{Y} - \boldsymbol{\mu}$ in the direction of $\mathbf{u}$; this question is about determining in which direction $\mathbf{Y}$ "spreads out" most from its expectation $\boldsymbol{\mu}$.)

⋄

Solution: The expectation of $\langle \mathbf{u}, \mathbf{Y} - \boldsymbol{\mu}\rangle$ is zero, so its variance is simply its expected squared value. With some clever regrouping of matrix multiplications, we can bring $\mathbf{u}$ outside of the expectation.

$$\begin{aligned}\mathbb{E}[\mathbf{u}^T(\mathbf{Y} - \boldsymbol{\mu})]^2 &= \mathbb{E}\mathbf{u}^T(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{u} \\ &= \mathbf{u}^T[\underbrace{\mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})^T}_{\mathbb{C}}]\mathbf{u}\end{aligned}$$

We see that the variance is exactly the quadratic form for the covariance matrix. We know from Exercise 1.41, that it's maximized by the principal eigenvector $\mathbf{q}_1$ in which case its value is the principal eigenvalue λ_1. Notice that Exercise 1.54 on principal components maximizing variance is the special case of this result applied to an empirical distribution.

Exercise 3.16

Let $\mathbf{Y}$ have expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$, and let $\mathbb{C} = \lambda_1\mathbf{q}_1\mathbf{q}_1^T + \ldots + \lambda_n\mathbf{q}_n\mathbf{q}_n^T$ be a spectral decomposition with $\lambda_1 \geq \ldots \geq \lambda_n$. Find the variance of $\langle \mathbf{Y} - \boldsymbol{\mu}, \mathbf{q}_j\rangle$, the coordinate of the centered random vector $\mathbf{Y} - \boldsymbol{\mu}$ in the direction of the eigenvector $\mathbf{q}_j$.

⋄

Solution: The expectation of $\langle \mathbf{u}, \mathbf{Y} - \boldsymbol{\mu}\rangle$ is zero, so its variance is simply its expected squared value. As in Exercise 3.15,

$$\mathbb{E}\langle \mathbf{Y} - \boldsymbol{\mu}, \mathbf{q}_j\rangle^2 = \mathbf{q}_j\mathbb{C}\mathbf{q}_j.$$

By Exercise 1.40, this equals the corresponding eigenvalue λ_j.

❧ ❋ ❧

3.3. Standardizing

You may recall *standardizing* a random variable by subtracting its expectation then dividing by its standard deviation; the resulting random variable has expectation equal to zero and standard deviation equal to one. There's a generalization of this process for random vectors. The first step is to subtract the expectation, of course, but the second step is a little more complicated: multiply by the square root[2] of the inverse of the covariance matrix.[3] The resulting random vector has expectation equal to the zero vector and covariance equal to the identity matrix as we'll see in Exercise 3.18.

[2]The square root of a matrix was defined in Exercise 1.32.

[3]Notice that standardizing requires the covariance matrix to be invertible.

Exercise 3.17

Let $\mathbb{V}$ be a positive definite matrix. Based on Exercises 1.32 and 1.35, explain why the inverse of the square root of $\mathbb{V}$ is the same as the square root of the inverse of $\mathbb{V}$.

$\diamond$

Solution: To find the square root of a positive semidefinite matrix, you replace the eigenvalues by their square roots. To find the inverse of an invertible symmetric matrix, you replace the eigenvalues by their reciprocals. No matter which order you do these two operations in, you end up with the same matrix:

$$\frac{1}{\sqrt{\lambda_1}}\mathbf{q}_1\mathbf{q}_1^T + \ldots + \frac{1}{\sqrt{\lambda_n}}\mathbf{q}_n\mathbf{q}_n^T$$

where $\mathbf{q}_1, \ldots, \mathbf{q}_n$ are eigenvectors of $\mathbb{V}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$.

Exercise 3.18

Let $\mathbf{Y}$ have expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$. Find the expectation and covariance of $\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$.

$\diamond$

Solution: The random vector $\mathbf{Y} - \boldsymbol{\mu}$ has expectation zero, so based on Exercise 3.2, the random vector $\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})$ has expectation $\mathbb{C}^{-1/2}\mathbf{0} = \mathbf{0}$. For the covariance, we apply the formula from Exercise 3.13 to get

$$\begin{aligned}\operatorname{cov}[\mathbb{C}^{-1/2}(\mathbf{Y} - \boldsymbol{\mu})] &= \mathbb{C}^{-1/2}\mathbb{C}(\mathbb{C}^{-1/2})^T \\ &= \underbrace{\mathbb{C}^{-1/2}\mathbb{C}^{1/2}}_{\mathbb{I}}\underbrace{\mathbb{C}^{1/2}\mathbb{C}^{-1/2}}_{\mathbb{I}} \\ &= \mathbb{I}.\end{aligned}$$

Because $\mathbb{C}^{1/2}$ is symmetric, so is its inverse according to Exercise 1.38.

A concept from introductory statistics that is closely related to standardizing is the *z-score* of a number relative to a distribution; it tells you how many standard deviations above the mean that number is. A non-negative generalization of this concept for random vectors is **Mahalanobis distance**. The Mahalanobis distance from a point (vector) $\mathbf{v}$ to a distribution with expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$ is defined to be $\|\mathbb{C}^{-1/2}(\mathbf{v} - \boldsymbol{\mu})\|$. Squared Mahalanobis distance is also commonly expressed as a quadratic form:

$$\begin{aligned}\|\mathbb{C}^{-1/2}(\mathbf{v} - \boldsymbol{\mu})\|^2 &= [\mathbb{C}^{-1/2}(\mathbf{v} - \boldsymbol{\mu})]^T[\mathbb{C}^{-1/2}(\mathbf{v} - \boldsymbol{\mu})] \\ &= (\mathbf{v} - \boldsymbol{\mu})^T(\mathbb{C}^{-1/2})^T\mathbb{C}^{-1/2}(\mathbf{v} - \boldsymbol{\mu}) \\ &= (\mathbf{v} - \boldsymbol{\mu})^T\mathbb{C}^{-1}(\mathbf{v} - \boldsymbol{\mu}).\end{aligned}$$

The Mahalanobis distance from a distribution to its mean is zero, it increases monotonically as you move away from the mean in any direction, and its level curves are shaped as *ellipsoids* with the covariance eigenvector directions as axes.

Exercise 3.19

If $\mathbf{Y}$ is an $\mathbb{R}^n$-valued random vector with expectation $\boldsymbol{\mu}$ and covariance matrix $\mathbb{C}$, find the expected squared Mahalanobis distance from $\mathbf{Y}$ to its own distribution.

Exercise 3.20

Let P be a distribution on $\mathbb{R}^n$ with expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$. Let $\mathbb{C} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T$ be a spectral decomposition with $\lambda_1 \geq \ldots \geq \lambda_n$. Find the unit vector $\mathbf{u}$ for which the Mahalanobis distance from $\boldsymbol{\mu} + \mathbf{u}$ to P is minimized.

Exercise 3.21

Let P be a distribution on $\mathbb{R}^n$ with expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$. Let $\mathbb{C} = \lambda_1 \mathbf{q}_1 \mathbf{q}_1^T + \ldots + \lambda_n \mathbf{q}_n \mathbf{q}_n^T$ be a spectral decomposition. Find $z \geq 0$ for which the Mahalanobis distance from $\boldsymbol{\mu} + z\mathbf{q}_j$ to P equals a specified $\alpha \geq 0$.

Note that by comparing the reasoning in Exercises 3.16 and 3.20, we can observe that the squared Mahalanobis distance in any eigenvector direction $\mathbf{q}_1, \ldots, \mathbf{q}_n$ from $\boldsymbol{\mu}$ is exactly the reciprocal of the variance of the coordinate in that direction. This reveals a simple way of understanding Exercise 3.21: a point that is α standard deviations from the mean in the direction of $\mathbf{q}_j$ has Mahalanobis distance α from P.

3.4. Random quadratic forms

Recall that every orthogonal projection matrix is both idempotent (Exercise 1.67) and symmetric (Exercise 1.68). For any orthogonal projection matrix $\mathbb{H}$, these two properties together let us represent the squared norm of $\mathbb{H}\mathbf{v}$ conveniently as a quadratic form:

$$\begin{aligned} \|\mathbb{H}\mathbf{v}\|^2 &= (\mathbb{H}\mathbf{v})^T(\mathbb{H}\mathbf{v}) \\ &= \mathbf{v}^T\mathbb{H}^T\mathbb{H}\mathbf{v} \\ &= \mathbf{v}^T\mathbb{H}\mathbf{v}. \end{aligned}$$

The identity works for random vectors as well: the squared length of $\mathbb{H}\mathbf{Y}$ equals $\mathbf{Y}^T\mathbb{H}\mathbf{Y}$. This representation will help us find the expected squared length of the orthogonal projection of a random vector onto a subspace.

A clever trick for manipulating quadratic forms comes from realizing that the result of the matrix multiplications is just a number, which is a 1×1 matrix and is equal to its trace; this allows us to make use of the cyclic permutation property of the trace operator (Exercise 1.33). Let $\mathbf{Y}$ be a random vector *with mean zero* and covariance matrix $\mathbb{C} \in \mathbb{R}^{n \times n}$. For any matrix $\mathbb{M} \in \mathbb{R}^{n \times n}$,

$$\begin{aligned} \mathbb{E}\mathbf{Y}^T\mathbb{M}\mathbf{Y} &= \mathbb{E}\,\mathrm{tr}\,(\mathbf{Y}^T\mathbb{M}\mathbf{Y}) \\ &= \mathbb{E}\,\mathrm{tr}\,(\mathbb{M}\mathbf{Y}\mathbf{Y}^T) \\ &= \mathrm{tr}\,(\mathbb{M}\,\mathbb{E}\mathbf{Y}\mathbf{Y}^T) \\ &= \mathrm{tr}\,[\mathbb{M}\mathbb{C}]. \end{aligned}$$

Note that the trace operator commutes with the expectation operator because you get the same result if you take the expectations of the diagonals of a matrix before summing them or if you sum them before taking the expectation.

Exercise 3.22

Let $\mathbf{Y}$ be a random vector with expectation $\boldsymbol{\mu}$ and covariance $\mathbb{C}$. Derive a formula for $\mathbb{E}\mathbf{Y}^T\mathbb{M}\mathbf{Y}$ in this more general case using the formula from the mean-zero case.

◇

Solution: Notice that $\mathbf{Y} - \boldsymbol{\mu}$ has the same covariance that $\mathbf{Y}$ does. And because $\mathbf{Y} - \boldsymbol{\mu}$ has mean zero, we can apply our trace formula if it's the vector in a quadratic form.

$$\begin{aligned}\mathbb{E}\mathbf{Y}^T\mathbb{M}\mathbf{Y} &= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu} + \boldsymbol{\mu})^T\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu} + \boldsymbol{\mu}) \\ &= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})^T\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu}) + \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})^T\mathbb{M}\boldsymbol{\mu} + \mathbb{E}\boldsymbol{\mu}^T\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu}) + \mathbb{E}\boldsymbol{\mu}^T\mathbb{M}\boldsymbol{\mu} \\ &= \operatorname{tr}[\mathbb{M}\mathbb{C}] + \boldsymbol{\mu}^T\mathbb{M}\boldsymbol{\mu}\end{aligned}$$

The two middle terms are zero because $\mathbb{E}\mathbf{Y} = \boldsymbol{\mu}$.

Exercise 3.23

Let $\mathbb{H}$ be the orthogonal projection matrix for $\mathcal{S}$, and let $\mathbf{Y}$ be a random vector with covariance $\sigma^2\mathbb{I}$. Use the result of Exercise 3.22 along with the observation from Exercise 1.71 to relate $\mathbb{E}\|\mathbb{H}\mathbf{Y}\|^2$ to $\mathcal{S}$.

◇

Solution: Because $\mathbb{H}$ is an orthogonal projection matrix, the squared length of $\mathbb{H}\mathbf{Y}$ equals the quadratic form $\mathbf{Y}^T\mathbb{H}\mathbf{Y}$. Using our formula for expected quadratic forms, this becomes

$$\begin{aligned}\mathbb{E}\mathbf{Y}^T\mathbb{H}\mathbf{Y} &= \operatorname{tr}[\mathbb{H}\sigma^2\mathbb{I}] + \boldsymbol{\mu}^T\mathbb{H}\boldsymbol{\mu} \\ &= \sigma^2\operatorname{tr}\mathbb{H} + \boldsymbol{\mu}^T\mathbb{H}\boldsymbol{\mu} \\ &= \sigma^2\dim(\mathcal{S}) + \boldsymbol{\mu}^T\mathbb{H}\boldsymbol{\mu}\end{aligned}$$

because according to Exercise 1.71 the trace of an orthogonal projection matrix equals the dimension of the subspace that it projects onto.

In the result of Exercise 3.23, it can be more convenient to rewrite $\boldsymbol{\mu}^T\mathbb{H}\boldsymbol{\mu}$ as $\|\mathbb{H}\boldsymbol{\mu}\|^2$; in other words, when the random vector doesn't have mean zero, the additional term in its expected quadratic form is equal to the squared length of the orthogonal projection of its mean onto the subspace that $\mathbb{H}$ projects onto. Clearly if $\boldsymbol{\mu}$ is orthogonal to that space, then the second term vanishes. As another special case of interest, if $\boldsymbol{\mu}$ is in the subspace, then the second term simplifies to $\|\boldsymbol{\mu}\|^2$.

Homework 5: Simulating random vectors

In *random-vectors.Rmd*, you'll perform simulations to see some of this chapter's results in action.

Now that we understand random vectors, we're ready to use them to define the linear model in Chapter 4.

CHAPTER

4

THE LINEAR MODEL

A MODEL IS A MATHEMATICAL formulation that is proposed to approximately represent a real-world phenomenon. The model is a *simplified* version of reality that is typically imperfect but may still be useful.[1] Often the model is a set of statements that are all under consideration, and the task remains to select one statement in particular based on your observations of real-world data. Many real-world quantities cannot be predicted exactly, either due to limitations in our knowledge or due to the nature of physical reality. In those cases, we use *random variables* in the formulations and call them *probabilistic* (as opposed to *deterministic*) models. In this chapter, we will begin considering probability distributions that depend on fixed explanatory variables and randomly generate the response variables.

The terminology "linear model" can be misleading. It refers to the response variable and its expectation being related *linearly in the parameters*; it's not about linearity in the explanatory variables. Thus, for example, with $\epsilon_1, \ldots, \epsilon_n$ as mean-zero random variables (called "errors") and with the parameter θ ranging over $\mathbb{R}$,

$$Y_i = \theta e^{x_i} + \epsilon_i$$

is a linear model, but

$$Y_i = e^{\theta x_i} + \epsilon_i$$

isn't a linear model.

For simplicity, we'll focus on a special case in this text, the *multiple linear model*:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i$$

with $(\beta_0, \ldots, \beta_d)$ ranging over $\mathbb{R}^{d+1}$ and with $\epsilon_1, \ldots, \epsilon_n$ each having mean zero. Notice that the expectation of Y_i is $\beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_m x_i^{(m)}$.

In general, however, we'll use the term *linear model* for any model of the form

$$Y_i = \beta_0 f_0(x_i^{(1)}, \ldots, x_i^{(m)}) + \ldots + \beta_d f_d(x_i^{(1)}, \ldots, x_i^{(m)}) + \epsilon_i. \tag{4.1}$$

with $(\beta_0, \ldots, \beta_d)$ ranging over $\mathbb{R}^{d+1}$ and with $\epsilon_1, \ldots, \epsilon_n$ each having mean zero.

[1] In general, there is a trade-off between simplicity and accuracy.

4.1. Visualizing the observations

As we discuss modeling assumptions, we'll add them into the pictures we developed in Chapter 2. Paralleling our approach in that chapter, we'll begin by visualizing the *observations*.

4.1.1. The location model

A random variable can always be expressed as its expectation plus a mean-zero *error* random variable whose distribution is simply a shifted version of the distribution of the original random variable. For example, $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$ is equivalent to $Y_i = \mu + \epsilon_i$ with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.[2] Models of the form $Y_i = \mu + \epsilon_i$ with mean-zero errors and with μ ranging over $\mathbb{R}$ are called *location models*.

We can easily include the new ingredients into our picture; see Figure 4.1 which depicts in particular an iid Normal distribution for the errors.

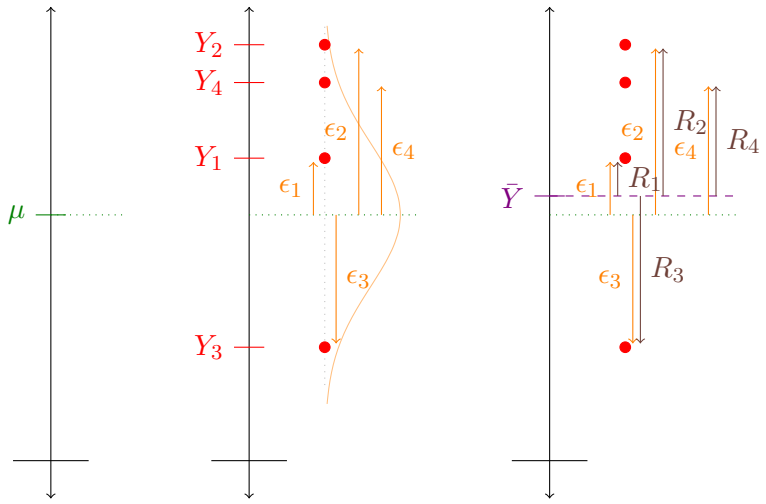


Figure 4.1: Left: According to the location model, the expected value of each response is μ. Middle: The response variable values would be *at μ except that they are "kicked" away by the random errors. This is a probabilistic explanation for the data in Figure 2.1. Right: The least-squares point provides a fitted value for the response variable. In Figure 2.1 it served to summarize the data, but now we also consider it an estimator of μ. Each residual is the difference between the response value and its fitted value, whereas each error is the difference between the response value and its expected value.*

Exercise 4.1

Suppose $Y_i = \mu + \epsilon_i$ for $i \in \{1, \ldots, n\}$ with mean-zero errors $\epsilon_1, \ldots, \epsilon_n$. Show that the average $\bar{Y}$ is an unbiased estimator for μ.

Exercise 4.2

Suppose $Y_i = \mu + \epsilon_i$ for $i \in \{1, \ldots, n\}$ with uncorrelated mean-zero errors $\epsilon_1, \ldots, \epsilon_n$ each having variance σ^2. Find the variance of $\bar{Y}$.

Exercise 4.3

Suppose $Y_i = \mu + \epsilon_i$ for $i \in \{1, \ldots, n\}$ with mean-zero errors $\epsilon_1, \ldots, \epsilon_n$. Find the expectation of the residual $Y_i - \bar{Y}$.

[2] Here the mean is a constant μ, but in upcoming sections, we will take the mean to be a function of explanatory variables.

Exercise 4.4

Suppose $Y_i = \mu + \epsilon_i$ for $i \in \{1, \ldots, n\}$ with uncorrelated mean-zero errors $\epsilon_1, \ldots, \epsilon_n$ each having variance σ^2. Find the variance of the residual $Y_i - \bar{Y}$.

4.1.2. The simple linear model

Next, suppose we have one explanatory variable. We'll consider the *simple linear model*:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

with (β_0, β_1) ranging over $\mathbb{R}^2$ and $\epsilon_1, \ldots, \epsilon_n$ each having expected value zero. Figures 4.2, 4.3, and 4.4 put the new ingredients into the scatterplot picture, depicting in particular an iid Normal distribution for the errors.

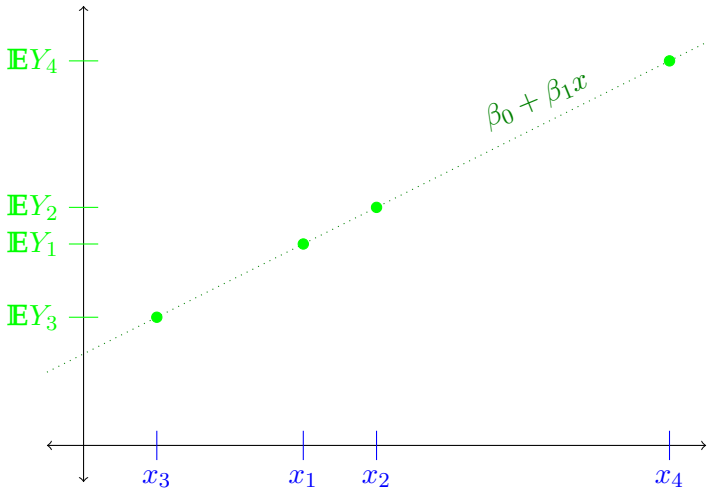


Figure 4.2: According to the simple linear model, there is a true line along which the expected values of the response variable lie.

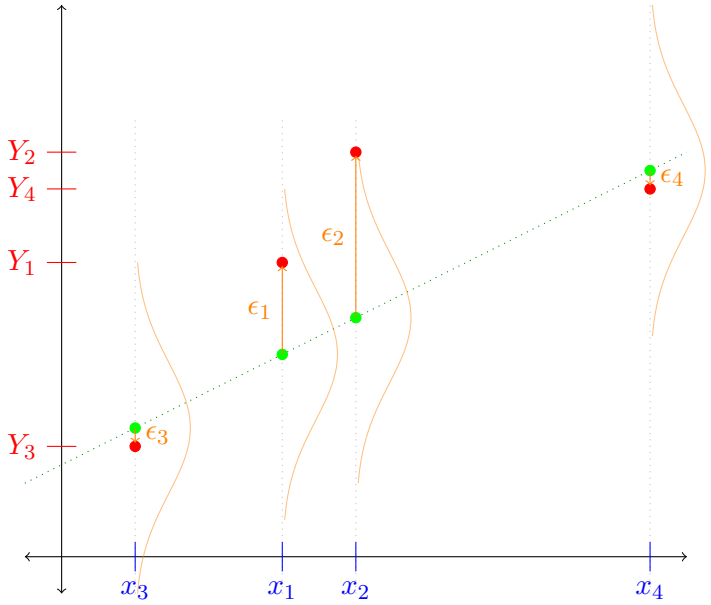


Figure 4.3: For each observation, an error kicks the response variable away from its expectation on the true line. This is a probabilistic explanation for the data in Figure 2.2.

Exercise 4.5

Which is larger: the sum of squared errors or the sum of squared residuals? Base your answer on the definition of the least-squares line, and explain.

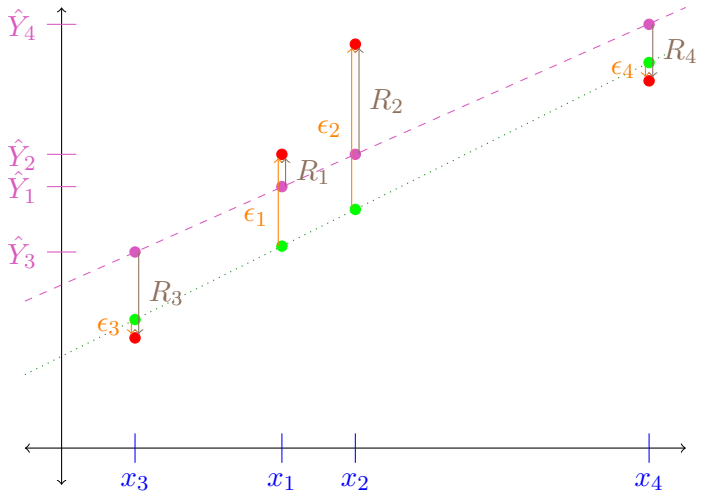


Figure 4.4: The least-squares line provides fitted values for the response variable. In Figure 2.3 it served to describe the data, but now we also consider it an estimator of the true line.

Suppose we want to estimate the "true" line, assuming the model is correct. We might consider using the least-squares line. Specifically, we can use the least-squares coefficients as estimates of the true coefficients. In this chapter, we'll analyze these estimators and related quantities without assuming a particular distribution for the errors. In Chapter 6, we'll assume specifically that the errors are Normal, which will allow us to devise hypothesis tests and confidence intervals.

4.1.3. The multiple linear model

A generalization of the simple linear model is the *multiple linear model* which has arbitrarily many explanatory variables:

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i \tag{4.2}$$

with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)$ ranging over $\mathbb{R}^{m+1}$ and with mean-zero errors.[3] We can also write the response values and errors as entries of random vectors $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots \epsilon_n)$.

We can also write this model with the vector equation

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1^{(1)} & \ldots & x_1^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n^{(1)} & \ldots & x_n^{(m)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

A more concise notation for the model is $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, defining the design matrix as in Chapter 2. Notice that the expectation of the response variable is

$$\begin{aligned} \mathbb{E}\mathbf{Y} &= \mathbb{E}(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \mathbb{E}\underbrace{\mathbb{X}\boldsymbol{\beta}}_{\text{non-random}} + \underbrace{\mathbb{E}\epsilon}_{\mathbf{0}} \\ &= \mathbb{X}\boldsymbol{\beta}. \end{aligned}$$

In Chapter 2, least-squares fitting was introduced as a sensible way of summarizing the relationships between data vectors or perhaps as a way to compress data. Now observe that the

[3] It is left to the reader to envision analogues of Figures 4.2, 4.3, and 4.4 with two explanatory variables.

random vector $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_n)$ of least-squares predicted values and the random vector of least-squares coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_m)$ are both linear transformations of $\mathbf{Y}$. This turns out to be incredibly convenient in allowing us to analyze a variety of properties of $\hat{\mathbf{Y}}$ considered as an estimator of $\mathbb{E}\mathbf{Y}$ as well as the vector of least-squares coefficients considered as estimators of the true coefficient vector when we assume that the data actually behaves according to the model. Letting $\mathbb{H}$ represent the orthogonal projection matrix onto $C(\mathbb{X})$, the expectation of the least-squares prediction vector is

$$\begin{aligned}\mathbb{E}\hat{\mathbf{Y}} &= \mathbb{E}\mathbb{H}\mathbf{Y} \\ &= \mathbb{E}\mathbb{H}[\mathbb{E}\mathbf{Y} + \boldsymbol{\epsilon}] \\ &= \mathbb{H}\mathbb{E}\mathbf{Y} + \mathbb{H}\underbrace{\mathbb{E}\boldsymbol{\epsilon}}_{\mathbf{0}} \\ &= \mathbb{H}\mathbb{E}\mathbf{Y}.\end{aligned}$$

If $\mathbb{E}\mathbf{Y} \in C(\mathbb{X})$, then $\mathbb{H}\mathbb{E}\mathbf{Y} = \mathbb{E}\mathbf{Y}$. In that case, $\hat{\mathbf{Y}}$ is indeed unbiased for $\mathbb{E}\mathbf{Y}$.

Exercise 4.6

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for a mean-zero error vector $\boldsymbol{\epsilon}$. Find the expectation of the residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$.

Exercise 4.7

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for a mean-zero error vector $\boldsymbol{\epsilon}$. Assuming the columns of the design matrix $\mathbb{X}$ are linearly independent, show that the least-squares coefficient vector $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for the true coefficients.

◇

Solution: Using our least-squares coefficients formula,

$$\begin{aligned}\mathbb{E}\hat{\boldsymbol{\beta}} &= \mathbb{E}[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}] \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underbrace{\mathbb{E}\mathbf{Y}}_{\mathbb{X}\boldsymbol{\beta}} \\ &= \boldsymbol{\beta}.\end{aligned}$$

Exercise 4.8

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and that the errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated and each have variance σ^2. Assuming the columns of the design matrix $\mathbb{X}$ are linearly independent, find the covariance matrix of the least-squares coefficient vector $\hat{\boldsymbol{\beta}}$.

◇

Solution: The error vector has covariance matrix $\sigma^2\mathbb{I}$, and therefore so does $\mathbf{Y}$. Using the formula from Exercise 3.13,

$$\begin{aligned}\mathrm{cov}\hat{\boldsymbol{\beta}} &= \mathrm{cov}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y} \\ &= [(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T](\underbrace{\mathrm{cov}\mathbf{Y}}_{\sigma^2\mathbb{I}})[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T]^T \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\underbrace{\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}}_{\mathbb{I}} \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}.\end{aligned}$$

Exercise 4.9

Suppose the multiple linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds and that the errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated and each have variance σ^2. Assume the columns of the design matrix $\mathbb{X}$ are linearly independent. Find the covariance matrix for $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$ in terms of σ^2, n, and $\boldsymbol{\Sigma}$, the empirical covariance matrix for the explanatory variables.

$\diamond$

Solution: In Chapter 2's coverage of least-squares fitting, we found an alternative expression for these estimators. Letting $\widetilde{\mathbb{X}}$ be the matrix whose columns are the centered versions of the explanatory vectors (excluding the $\mathbf{1}$ column), the least-squares coefficients (excluding the intercept) are $\boldsymbol{\Sigma}^{-1}(\frac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y})$. The covariance of this estimator is

$$\begin{aligned}
\text{cov}[\boldsymbol{\Sigma}^{-1}(\tfrac{1}{n}\widetilde{\mathbb{X}}^T\mathbf{Y})] &= [\boldsymbol{\Sigma}^{-1}(\tfrac{1}{n}\widetilde{\mathbb{X}}^T)](\underbrace{\text{cov}\mathbf{Y}}_{\sigma^2\mathbb{I}})[\boldsymbol{\Sigma}^{-1}(\tfrac{1}{n}\widetilde{\mathbb{X}}^T)]^T \\
&= \tfrac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}(\underbrace{\tfrac{1}{n}\widetilde{\mathbb{X}}^T\widetilde{\mathbb{X}}}_{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}^{-1} \\
&= \tfrac{\sigma^2}{n}\boldsymbol{\Sigma}^{-1}.
\end{aligned}$$

This result not only relates the covariance of $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$ to a familiar quantity, it also helps clarify the effect of sample size in the accuracy of estimating $(\beta_1, \ldots, \beta_m)$.

Exercise 4.10

Suppose the multiple linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ holds and that the errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated and each have variance σ^2. Assume the columns of the design matrix $\mathbb{X}$ are linearly independent. With $\hat{\alpha}_0$ representing the least-squares intercept when the explanatory variables have been centered, find the covariance matrix for $(\hat{\alpha}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)$ in terms of σ^2, n, and $\boldsymbol{\Sigma}$, the empirical covariance matrix for the explanatory variables.

Exercise 4.11

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and that the errors $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated and each have variance σ^2. What is the covariance matrix of the least-squares predictions $\hat{\mathbf{Y}}$.

4.2. Visualizing the variables

The new modeling ingredients can also be nicely incorporated in the variables picture. In each case, $\mathbb{E}\mathbf{Y}$ is a vector (which we'll represent as a point) in the column space of $\mathbb{X}$ showing where $\mathbf{Y}$ *would be*, except that the pesky error vector $\boldsymbol{\epsilon}$ kicks it off into space. Finally, least-squares regression projects $\mathbf{Y}$ back into the column space of $\mathbb{X}$.

4.2.1. The location model

Again we start with no explanatory variables. When we were focusing on the observations, we stated the location model assumption $Y_i = \mu + \epsilon_i$ with mean-zero errors. A vector version of this is $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$ with $\boldsymbol{\mu} := \mu\mathbf{1}$ and with each component of the error vector $\boldsymbol{\epsilon}$ having mean zero.

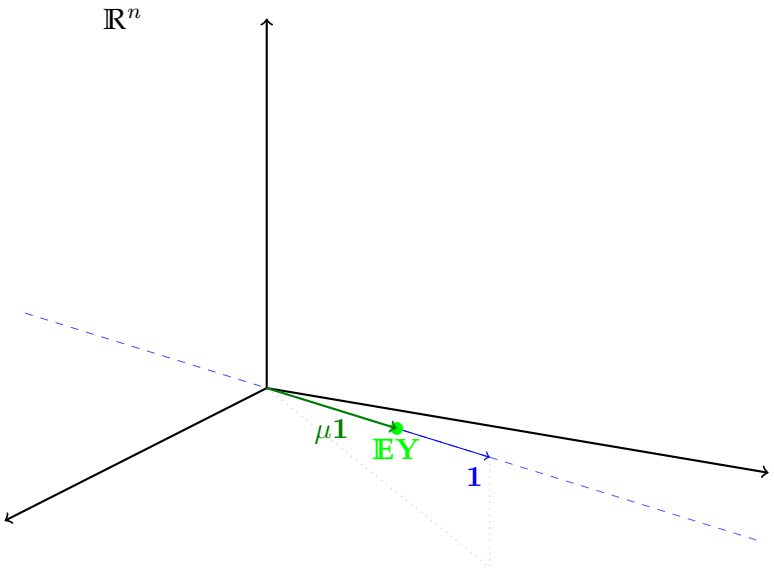


Figure 4.5: A generic picture of the constant vector $\mathbf{1}$*, its span, and the expectation of the response variable vector* $\mathbb{E}\mathbf{Y} = \mu\mathbf{1}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*.*

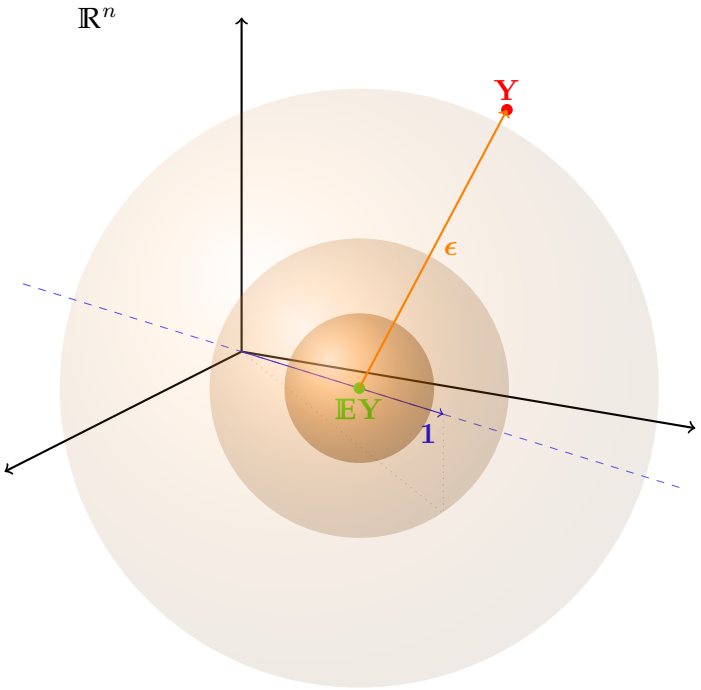


Figure 4.6: A generic picture of the constant vector $\mathbf{1}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, a density for the error vector, and a realization of that error* $\boldsymbol{\epsilon}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*. (The density depicted is spherically symmetric, bringing to mind the symmetric multivariate Normal density, though we won't specifically assume Normal errors until Chapter 6.) The error vector kicks* $\mathbf{Y}$ *out into space away from its expectation.*

4.2.2. The simple linear model

With one explanatory variable, we consider the simple linear model, written in terms of vectors as

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \boldsymbol{\epsilon}$$

with mean-zero errors. Compare Figures 4.8, 4.9, and 4.10 to our earlier Figures 2.8 and 2.9 to see what the model adds.

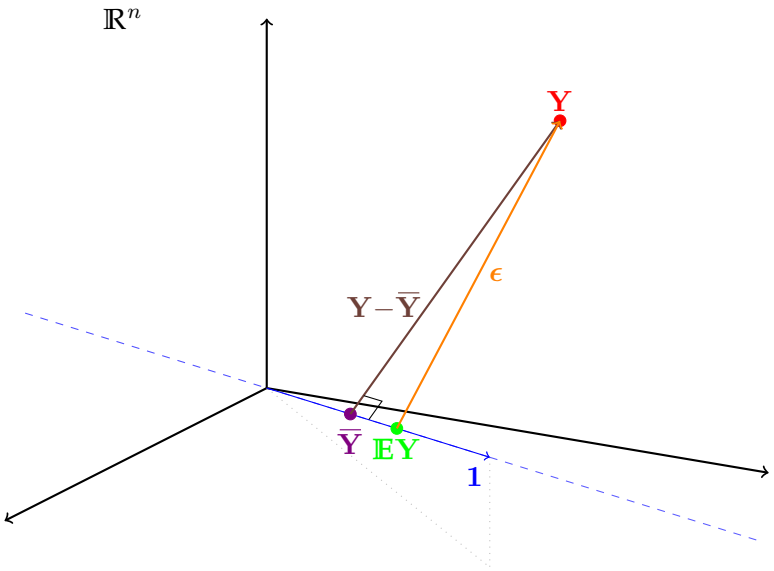


Figure 4.7: A generic picture of the constant vector $\mathbf{1}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, and a realization of the error* $\boldsymbol{\epsilon}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*. The orthogonal projection* $\hat{\mathbf{Y}}$ *of the response* $\mathbf{Y}$ *back onto span*$\{\mathbf{1}\}$ *can be thought of as an estimator for* $\mathbb{E}\mathbf{Y}$*.*

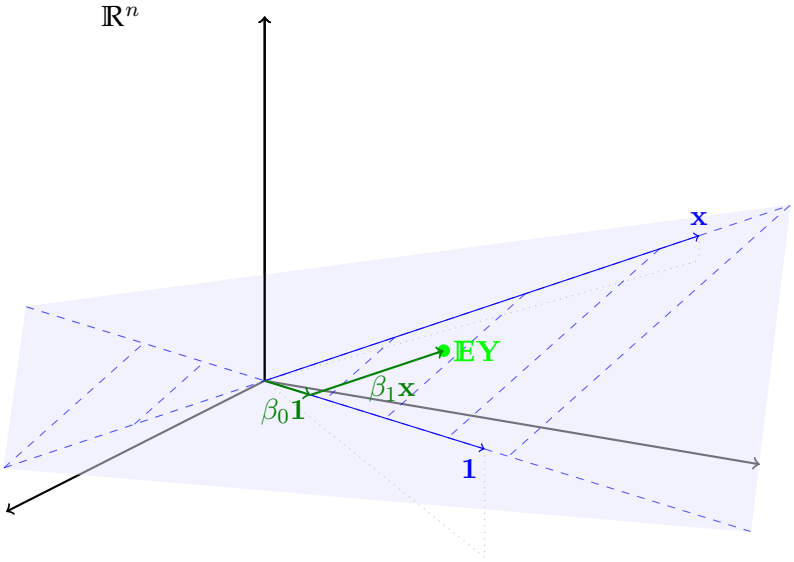


Figure 4.8: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and the expectation of the response variable vector* $\mathbb{E}\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}$ *in span*$\{\mathbf{1}, \mathbf{x}\}$*, assuming the simple linear model is true.*

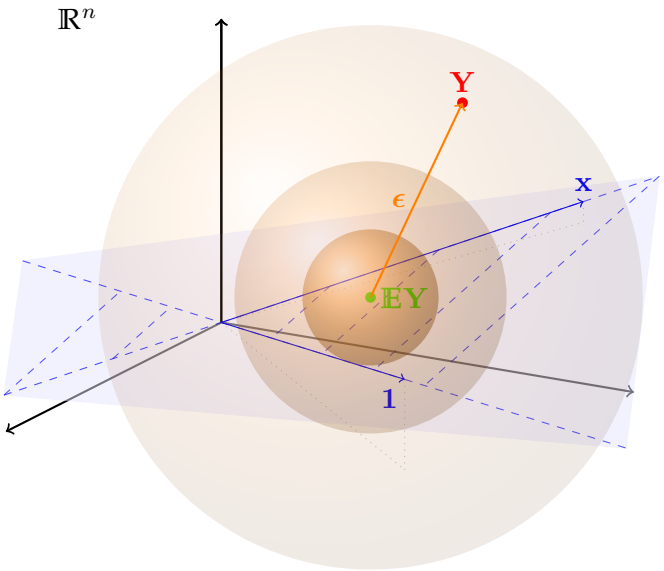


Figure 4.9: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, a density for the error vector, and a realization of that error* $\boldsymbol{\epsilon}$*, assuming the simple linear model is true. (The density depicted is spherically symmetric, bringing to mind the symmetric multivariate Normal density, though we won't specifically assume Normal errors until Chapter 6.) The error vector kicks* $\mathbf{Y}$ *out into space away from its expectation.*

In Chapter 2, we noted a potential pathology of the data. If every component of the $\mathbf{x}$ vector is the same, then the span of $\{\mathbf{1}, \mathbf{x}\}$ is one-dimensional rather than two-dimensional. In that case,

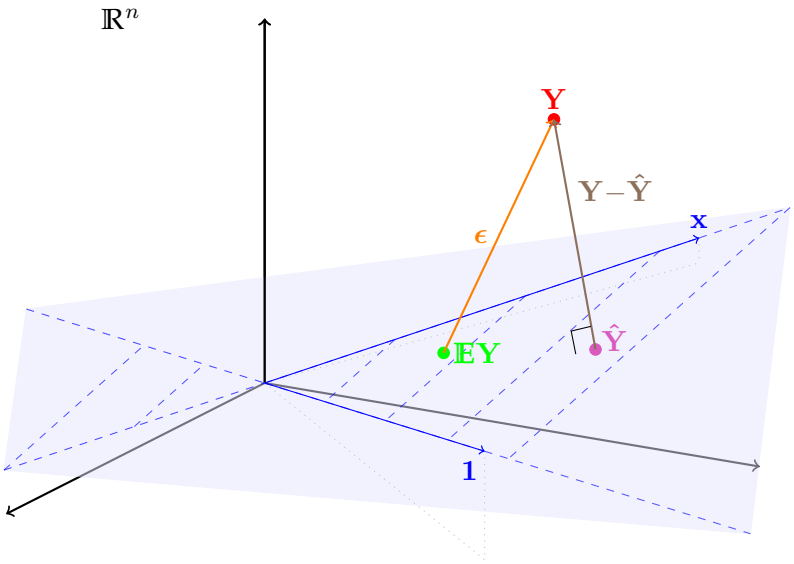


Figure 4.10: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, and a realization of the error* $\boldsymbol{\epsilon}$*, assuming the simple linear model is true. The orthogonal projection* $\hat{\mathbf{Y}}$ *of the response* $\mathbf{Y}$ *back onto span*$\{\mathbf{1}, \mathbf{x}\}$ *can be thought of as an estimator for* $\mathbb{E}\mathbf{Y}$*.*

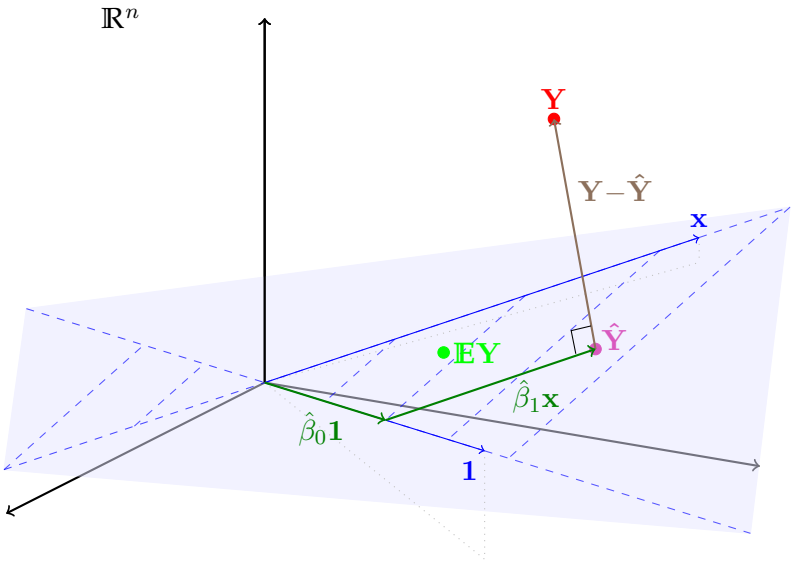


Figure 4.11: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the response* $\mathbf{Y}$*, and its orthogonal projection* $\hat{\mathbf{Y}}$ *back into the column space of design matrix. The least-squares coefficients of* $\mathbf{1}$ *and* $\mathbf{x}$ *can be considered estimators for the true coefficients* β_0 *and* β_1*.*

there are infinitely many linear combinations of $\mathbf{1}$ and $\mathbf{x}$ that result in the orthogonal projection $\overline{\mathbf{Y}}$. More generally, if the columns of $\mathbb{X}$ aren't linearly independent, then there are infinitely many coefficient vectors that minimize the sum of squared residuals.

A model is called **identifiable** if each possible value of the parameter vector results in a distinct statement about the data. If, on the other hand, there are two different values of the parameter vector that say exactly the same thing about the data's distribution, then we couldn't possibly hope to tell which one is the *true* parameter value no matter how much data we have.

In our context, the linear model is identifiable if and only if the columns of $\mathbb{X}$ are linearly independent. However, even if the full set of parameters isn't identifiable, certain linear combinations of the parameters will be identifiable.

Exercise 4.12

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the response variable, and assume $\mathbf{x} = c\mathbf{1}$ for some $c \in \mathbb{R}$. In the context of the simple linear model, argue that the derived parameter $a_0 := b_0 + cb_1$ is identifiable.

Exercise 4.13

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ be the response variable, and assume $\mathbf{x} = c\mathbf{1}$ for some $c \in \mathbb{R}$. In the context of the simple linear model, identify an unbiased estimator for the parameter $a_0 := b_0 + cb_1$.

4.2.3. The multiple linear model

The multiple linear model, written in terms of vectors, is

$$\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}^{(1)} + \ldots + \beta_m\mathbf{x}^{(m)} + \boldsymbol{\epsilon}$$

with mean-zero errors and $\boldsymbol{\beta}$ ranging over $\mathbb{R}^{m+1}$. It generalizes the location model and the simple linear model, and yet it is still a special case of the general form (Equation 4.1). If the model is true, then the set of possible expectation vectors is $C(\mathbb{X})$, which was exactly the subspace of possible fits under consideration for multiple linear regression.

It takes a bit more care to draw a picture of the general form of the linear model in $\mathbb{R}^n$ that includes $\mathbb{E}\mathbf{Y}$, $\mathbf{Y}$, and $\hat{\mathbf{Y}}$. The column space of $\mathbb{X}$ will be depicted as a generic two-dimensional subspace in the picture. In particular, the intersection of the subspace with our picture has to be the plane that includes the origin, $\mathbb{E}\mathbf{Y}$ and $\hat{\mathbf{Y}}$. As a result, we aren't at liberty to draw any specific column vector of $\mathbb{X}$ if we want the picture to remain accurate; see Figures 4.12, 4.13, and 4.14.

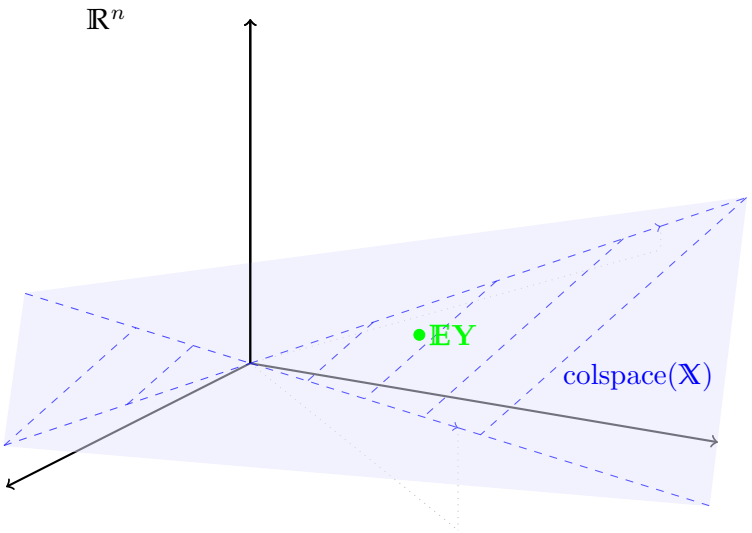


Figure 4.12: Assuming the linear model is true, $\mathbb{E}\mathbf{Y} = \mathbb{X}\boldsymbol{\beta}$ lies in the column space of the $\mathbb{X}$ matrix.

Exercise 4.14

The variables picture provides us with a more specific answer to Exercise 4.5. Use the Pythagorean identity to quantify the difference between the sum of squared residuals and the sum of squared errors if the linear model is true and the least-squares estimator $\hat{\mathbf{Y}}$ is used.

Solution: Because $\mathbb{E}\mathbf{Y} \in C(\mathbb{X})$, we have a right triangle with $\mathbb{E}\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\mathbf{Y}$ as its vertices.

$$\begin{aligned} \|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2 &= \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2 \\ \Rightarrow \|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 &= \|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2 \end{aligned}$$

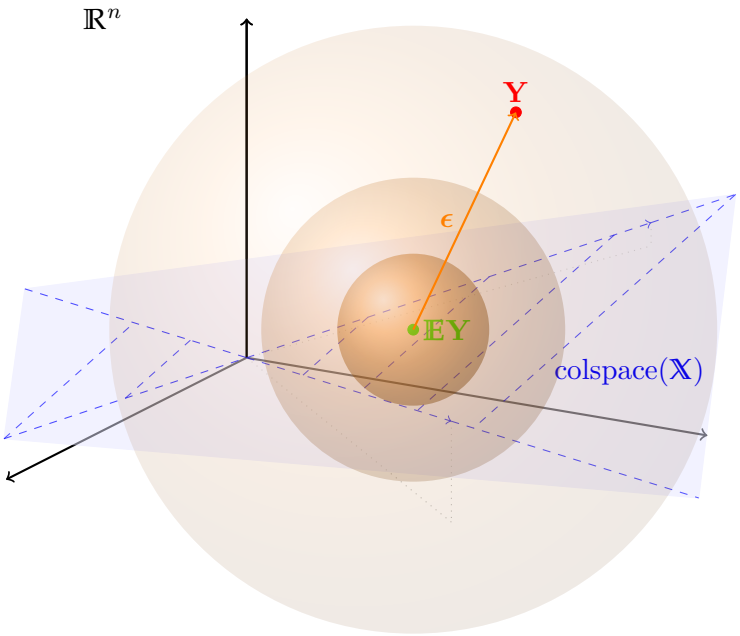


Figure 4.13: The error vector kicks the response $\mathbf{Y}$ *off into space away from its expectation.*

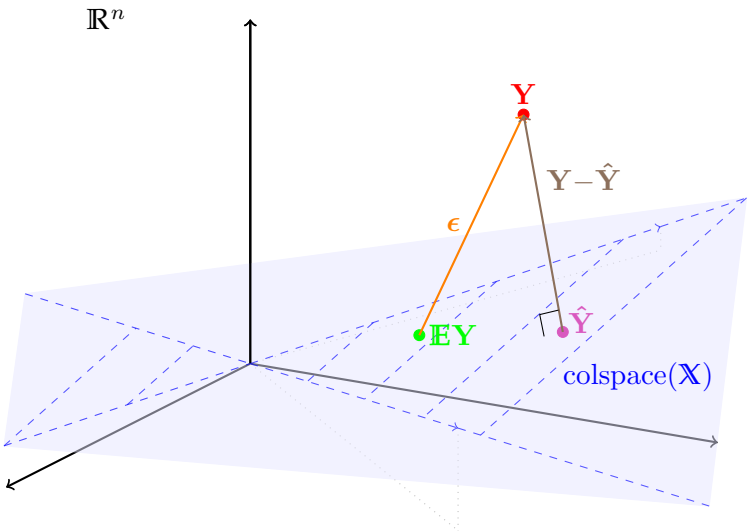


Figure 4.14: The orthogonal projection $\hat{\mathbf{Y}}$ *of the response* $\mathbf{Y}$ *back onto the column space of* $\mathbb{X}$ *can be considered an estimator for* $\mathbb{E}\mathbf{Y}$.

Exercise 4.15

With $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, express the least-squares residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$ as an orthogonal projection matrix times the error vector.

◇

Solution: Let $\mathbb{H}$ denote the orthogonal projection matrix onto $C(\mathbb{X})$. The residual vector is

$$\begin{aligned}\mathbf{Y} - \hat{\mathbf{Y}} &= \mathbf{Y} - \mathbb{H}\mathbf{Y} \\ &= (\mathbb{I} - \mathbb{H})\mathbf{Y} \\ &= (\mathbb{I} - \mathbb{H})(\mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\mathbb{I} - \mathbb{H})\boldsymbol{\epsilon}.\end{aligned}$$

By Exercise 1.73, $\mathbb{I} - \mathbb{H}$ is the orthogonal projection matrix onto the orthogonal complement of $C(\mathbb{X})$. (And because $\mathbb{X}\beta \in C(\mathbb{X})$, it gets mapped to $\mathbf{0}$ by $\mathbb{I} - \mathbb{H}$.)

Exercise 4.15 points out that the residual vector is the orthogonal projection of the error vector onto $C(\mathbb{X})^\perp$. A more direct way of seeing this is to simply express the error vector as

$$\begin{aligned}\boldsymbol{\epsilon} &= \mathbf{Y} - \mathbb{E}\mathbf{Y} \\ &= (\underbrace{\mathbf{Y} - \hat{\mathbf{Y}}}_{\perp C(\mathbb{X})}) - (\underbrace{\mathbb{E}\mathbf{Y} - \hat{\mathbf{Y}}}_{\in C(\mathbb{X})})\end{aligned}$$

which also tells us that $\mathbb{E}\mathbf{Y} - \hat{\mathbf{Y}}$ is the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{X})$.

Homework 6: Simulating linear models

In *linear-models.Rmd*, you'll simulate data according to a known linear model so that you can compare the least-squares estimates to the true values.

4.3. Expected sums of squares

For linear models, we found in Section 4.1.3[4] that the least-squares prediction $\hat{\mathbf{Y}}$ is an unbiased estimator for the expectation $\mathbb{E}\mathbf{Y}$ and that the least-squares coefficient vector $\hat{\boldsymbol{\beta}}$ is an unbiased estimator for the true coefficients $\boldsymbol{\beta}$. With the additional assumptions that the errors are uncorrelated and that each has variance σ^2, we found simple formulas for the covariance matrices of both $\hat{\mathbf{Y}}$ and $\hat{\boldsymbol{\beta}}$. With these simple assumptions, we can also obtain expected squared lengths of various random vectors using formulas from Chapter 3.

Exercise 4.16

If every error has variance σ^2, what is the expected squared length of the error vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$?

Exercise 4.17

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error covariance $\sigma^2\mathbb{I}$. Find $\mathbb{E}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$, the expected sum of squared residuals using the result of Exercise 3.23.

◇

Solution: In Exercise 4.15, we realized that the residual vector can be expressed as $(\mathbb{I}-\mathbb{H})\boldsymbol{\epsilon}$. By Exercise 3.23, the expectation of interest equals $\sigma^2(n - \text{rank}(\mathbb{X}))$.

Exercise 4.18

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error covariance $\sigma^2\mathbb{I}$. Use the result of Exercise 4.17 to devise an *unbiased* estimator for σ^2.

◇

Solution: We know that $\mathbb{E}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 = \sigma^2(n - \text{rank}(\mathbb{X}))$. Therefore,

$$\hat{\sigma}^2 := \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{n - \text{rank}(\mathbb{X})}$$

is an unbiased estimator for σ^2 because that's its expectation.

[4] This section was about the multiple linear model, but it is clear that the derivations work for linear modeling in general.

Exercise 4.19

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error covariance $\sigma^2\mathbb{I}$. Find $\mathbb{E}\|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$ by realizing that it is equals the trace of the covariance matrix of $\hat{\mathbf{Y}}$.

Exercise 4.20

Use the Pythagorean identity along with the results of Exercises 4.16 and 4.17 to verify your solution to Exercise 4.19.

$\diamond$

Solution: Because $\hat{\mathbf{Y}}$ and $\mathbb{E}\mathbf{Y}$ are both in $C(\mathbb{X})$, the vector $\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}$ is orthogonal to the residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$. By the Pythagorean identity,

$$\|\mathbf{Y} - \mathbb{E}\mathbf{Y}\|^2 = \|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2 + \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2.$$

Notice that $\mathbf{Y} - \mathbb{E}\mathbf{Y}$ is exactly the error vector $\boldsymbol{\epsilon}$. Therefore, after solving for $\|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$, we can find its expectation to be

$$\begin{aligned}
\mathbb{E}\|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2 &= \mathbb{E}[\|\boldsymbol{\epsilon}\|^2 - \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2] \\
&= \mathbb{E}\|\boldsymbol{\epsilon}\|^2 - \mathbb{E}\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \\
&= \sigma^2 n - \sigma^2(n - \text{rank}(\mathbb{X})) \\
&= \sigma^2 \text{rank}(\mathbb{X}).
\end{aligned}$$

Exercise 4.21

Use the results of Exercises 3.8 and 4.19 to identify a random vector that is proportional to $\hat{\mathbf{Y}}$ that has smaller expected squared loss for estimating $\mathbb{E}\mathbf{Y}$ than $\hat{\mathbf{Y}}$ does.

Exercise 4.22

A celebrated result called the *Gauss-Markov Theorem* (which the interested reader can find elsewhere) implies that the least-squares fit $\hat{\mathbf{Y}}$ has the smallest possible expected squared loss among all random vectors that are both linear functions of $\mathbf{Y}$ and unbiased for $\mathbb{E}\mathbf{Y}$. Exercise 4.21 identified a random variable that has smaller expected squared loss than $\hat{\mathbf{Y}}$; explain why this doesn't contradict the Gauss-Markov Theorem.

So far, we've considered the behavior of the least-squares estimators when the true model is being used. But what if, for example, the true data-generating mechanism involves a variable that is omitted from the least-squares estimation? We'll consider such questions in the next exercises, as well as the upcoming homework.

Exercise 4.23

Assume $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and assume that $\mathbf{1}$ and $\mathbf{x}$ are linearly independent. Let $\hat{\mathbf{Y}}$ be the least-squares prediction for the model $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}$. Find $\mathbb{E}\|\hat{\mathbf{Y}} - \mathbb{E}\hat{\mathbf{Y}}\|^2$.

Exercise 4.24

Assume $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$. Let $\hat{\mathbf{Y}}$ be the least-squares prediction for the model $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \boldsymbol{\epsilon}$. Find $\|\mathbb{E}\mathbf{Y} - \mathbb{E}\hat{\mathbf{Y}}\|^2$.

Exercise 4.25

Assume $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$. Let $\hat{\mathbf{Y}}$ be the least-squares prediction for the model $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}$. Use the bias-variance decomposition along with your results from Exercises 4.23 and 4.24 to find $\mathbb{E}\|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$.

Notice the bias-variance trade-off exhibited in Exercise 4.25. If the $\mathbf{z}$ variable were included, then the squared bias term would drop to zero, but the variance term would increase from $2\sigma^2$ to $3\sigma^2$. So as long as $\beta_2\|(\mathbb{I}-\mathbb{H})\mathbf{z}\|$ is smaller than σ, it's statistically better to use the simpler model even though it's not true. This is demonstrative of a key trade-off in estimation and prediction. The benefit of low variance of an unrealistically simple model can outweigh its bias. In general, increasingly complex models gradually become preferable as the sample size grows.

The preceding exercises considered the bias-variance trade-off in estimating $\mathbb{E}\mathbf{Y}$ by $\hat{\mathbf{Y}}$. Next we explore the bias-variance trade-off in estimating coefficients.

Exercise 4.26

Let $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and suppose that $\mathbf{1}, \mathbf{x}, \mathbf{z}$ are linearly independent. If the multiple linear model is correctly specified and β_0, β_1, and β_2 are estimated by least-squares, find the "variance" of $(\hat{\beta}_1, \hat{\beta}_2)$, that is $\mathbb{E}\|(\hat{\beta}_1, \hat{\beta}_2) - (\beta_1, \beta_2)\|^2$, in terms of σ^2, n, and the eigenvalues of the empirical covariance matrix.

Exercise 4.27

Let $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and suppose that $\mathbf{1}, \mathbf{x}, \mathbf{z}$ are linearly independent. If the multiple linear model is incorrectly specified, omitting a term for $\mathbf{z}$, and thus only β_0, β_1 are estimated by least-squares, show that $\tilde{\beta}_1$, the incorrect model's least-squares estimate of β_1 is equal to $\beta_1 + \rho_{\mathbf{x},\mathbf{z}}\frac{\sigma_\mathbf{z}}{\sigma_\mathbf{x}}\beta_2 + \frac{\langle\boldsymbol{\epsilon}, \mathbf{x}-\bar{\mathbf{x}}\rangle}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}$.

Exercise 4.28

Let $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and suppose that $\mathbf{1}, \mathbf{x}, \mathbf{z}$ are linearly independent. If the multiple linear model is incorrectly specified, omitting a term for $\mathbf{z}$, and thus only β_0, β_1 are estimated by least-squares, find the "variance" of the resulting estimator $(\tilde{\beta}_1, \tilde{\beta}_2)$. Here $\tilde{\beta}_1$ represents the incorrect model's least-squares estimate of β_1 while $\tilde{\beta}_2$ is fixed at 0.

Exercise 4.29

Let $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and suppose that $\mathbf{1}, \mathbf{x}, \mathbf{z}$ are linearly independent. If the multiple linear model is incorrectly specified, omitting a term for $\mathbf{z}$, and thus only β_0, β_1 are estimated by simple linear regression, find the expectation of $\tilde{\beta}_1$, the incorrect model's least-squares estimate of β_1.

Exercise 4.30

Let $\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x} + \beta_2\mathbf{z} + \boldsymbol{\epsilon}$ with $\text{cov}\boldsymbol{\epsilon} = \sigma^2\mathbb{I}$, and suppose that $\mathbf{1}, \mathbf{x}, \mathbf{z}$ are linearly independent. If the multiple linear model is incorrectly specified, omitting a term for $\mathbf{z}$, and thus only β_0, β_1 are estimated by simple linear regression, find the "squared bias" of the resulting estimator $(\tilde{\beta}_1, \tilde{\beta}_2)$, as defined in Exercise 4.28.

Homework 7: Bias-variance tradeoff

In *bias-variance-tradeoff.Rmd*, you'll work out the exact bias-variance quantities for the scenario in Homework 6 and compare them to your earlier simulation results.

Next, we'll study the Normal distribution in Chapter 5 so that in Chapter 6 we can add a modeling assumption that the errors are Normal. As you'll see, this assumption opens the door to a tremendous variety of new opportunities for statistical inference.

CHAPTER

5

REVIEW: NORMALITY

The family of Normal distributions plays a key role in the theory of probability and statistics. According to the familiar Central Limit Theorem, the distribution of an average of iid random variables (with finite variance) tends toward Normality. In fact, more advanced versions of the theorem don't require the random variables to be iid, as long as they aren't *too* dependent or *too* disparate in their scales. We see this Central Limit phenomenon play out in the real world when we observe "bell-shaped" histograms of measurements in a wide range of contexts. The prevalence of approximate Normality in the world makes Normal distributions a natural part of statistical modeling. Fortunately, the Normal family is also particularly convenient for analyzing estimation procedures for these models.

The density of the *multivariate* Normal distribution with expectation $\boldsymbol{\mu} \in \mathbb{R}^n$ and positive definite covariance $\mathbb{C} \in \mathbb{R}^{n \times n}$ is

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d \det(\mathbb{C})}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbb{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{5.1}$$

In this chapter, we'll work out properties of the Normal family that will gain us much when we finally assume that the linear model's errors are Normal in Chapter 6.

Exercise 5.1

Let f be the density of a multivariate Normal distribution on $\mathbb{R}^d$. Let m denote the squared Mahalanobis distance from $\mathbf{x}$ to that distribution. If the squared Mahalanobis distance from $\mathbf{y}$ to the distribution is $m + 2$, find $\log \frac{f(\mathbf{x})}{f(\mathbf{y})}$.

5.1. Affine transformations

Equation 5.1 provided the general form of [multivariate] Normal densities on $\mathbb{R}^n$, but there's an interesting alternative characterization of Normality: *a random vector is [multivariate] Normal if and only if every non-zero linear combination of its entries is univariate Normal.* (We won't

prove this equivalence, but the interested reader can easily find justifications elsewhere.) This characterization makes it easy to verify the tremendously useful fact that *Normality is preserved by affine transformation.*

Exercise 5.2

Let $\mathbf{Y}$ be a Normally distributed random vector. For a non-random vector $\mathbf{v}$, show that $\mathbf{Y} + \mathbf{v}$ is also Normal.

Exercise 5.3

Let $\mathbf{Y}$ be a Normally distributed random vector. If $\mathbb{V}$ is a matrix with linearly independent rows, show that $\mathbb{V}\mathbf{Y}$ is also Normal.

❧ ❋ ❧

If the $n \times m$ matrix of the affine transformation has linearly dependent rows, then the resulting distribution is called *degenerate Normal*, meaning that it's multivariate Normal on some proper hyperplane within $\mathbb{R}^n$. In particular, this hyperplane is the range of the affine transformation. In the degenerate case, the usual density formula (Equation 5.1) doesn't quite work because the covariance matrix isn't invertible.

Exercise 5.4

If two random variables are *multivariate* Normal and are uncorrelated with each other, then they are independent; one can verify that their joint density factors into a product of their marginal densities. However, without *multivariate* Normality, uncorrelated doesn't necessarily imply independent. Construct a pair of marginally Normal random variables that are uncorrelated but not independent.

We learned in Homework 5 how a standardized random vector can be transformed to have any desired expectation and covariance by using just the right affine mapping. Because Normality is preserved by affine transformation, we have a simple method for producing draws from any multivariate Normal distribution by using only standard Normal draws: if $\mathbf{Z} \sim N(\mathbf{0}, \mathbb{I})$, then $\mathbb{C}^{1/2}\mathbf{Z} + \boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \mathbb{C})$. The idea of standardizing and "unstandardizing" while retaining Normality will be a recurring trick in a few of this chapter's exercises.

❧ ❋ ❧

5.2. Spherical symmetry

Let the distribution of $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ be defined by having iid $N(0, \sigma^2)$ draws as its entries. Alternatively, the entries $\epsilon_1, \ldots, \epsilon_n$ can be thought of as the coordinates of a *single draw* from the *multivariate Normal* distribution $N(\mathbf{0}, \sigma^2\mathbb{I})$.

Exercise 5.5

Let $\epsilon_1, \ldots, \epsilon_n$ be iid $N(0, \sigma^2)$; each has density $g(z) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z^2}{2\sigma^2}}$. Derive the density f of $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$, and observe that $f(z_1, \ldots, z_n)$ only depends on the length of its argument vector $(z_1, \ldots, z_n)$.

◇

Solution: By independence, the joint density equals the product of the individual densities.

$$\begin{aligned} f(z_1, \ldots, z_n) &= \prod_i g(z_i) \\ &= \prod_i \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{z_i^2}{2\sigma^2}} \\ &= (\tfrac{1}{2\pi\sigma^2})^{n/2} e^{-\frac{\sum_i z_i^2}{2\sigma^2}} \\ &= (\tfrac{1}{2\pi\sigma^2})^{n/2} e^{-\frac{\|\mathbf{z}\|^2}{2\sigma^2}} \end{aligned}$$

with $\mathbf{z} := (z_1, \ldots, z_n)$.

❧ ❋ ❧

Exercise 5.5 reveals that the density of $N(\mathbf{0}, \sigma^2\mathbb{I})$ at $\mathbf{z} = (z_1, \ldots, z_n)$ only depends on its distance from the origin and that the density decreases monotonically as the distance from the origin increases. Spherical level sets for its density are drawn in Figure 5.1 from a generic three-dimensional perspective. Much of the analysis in Chapter 6 will stem from this *spherical symmetry* of the $N(\mathbf{0}, \sigma^2\mathbb{I})$ distribution.

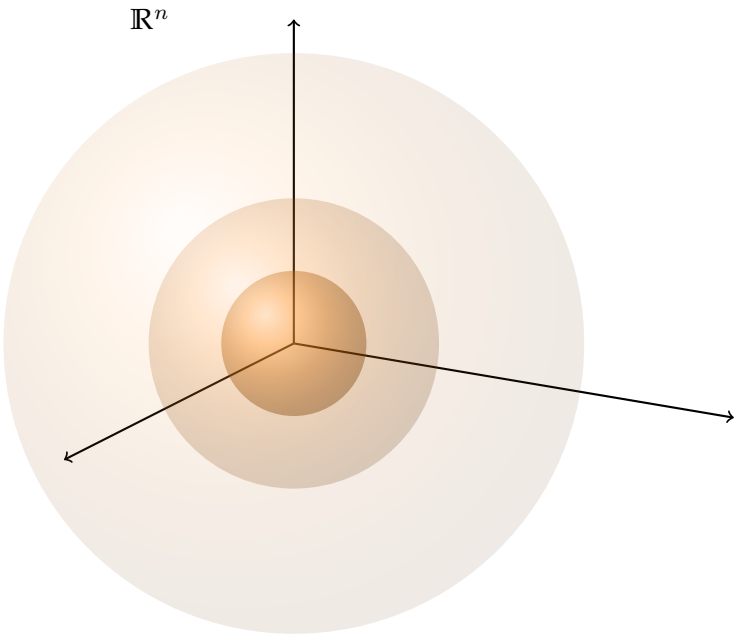


Figure 5.1: If $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$, then $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. The density of the $N(\mathbf{0}, \sigma^2\mathbb{I})$ distribution is spherically symmetric about $\mathbf{0}$.

Now suppose that we want to use another set of orthogonal coordinate axes for $\mathbb{R}^n$ that are rotated relative to the original axes. Let $(\delta_1, \ldots, \delta_n)$ be the coordinates of $\boldsymbol{\epsilon}$ with respect to the new coordinate axes. *By spherical symmetry, the joint distribution of the new coordinates $\delta_1, \ldots, \delta_n$ is exactly the same as the joint distribution of the original coordinates $\epsilon_1, \ldots, \epsilon_n$; in other words, $\delta_1, \ldots, \delta_n$ are also iid $N(0, \sigma^2)$.*

With $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$, let's think about the orthogonal projection of $\boldsymbol{\epsilon}$ onto a subspace $\mathcal{S}$. Consider an alternative orthonormal basis $\mathbf{u}_1, \ldots, \mathbf{u}_n$ which has its first $\dim(\mathcal{S})$ vectors spanning $\mathcal{S}$ and its remaining $n-\dim(\mathcal{S})$ vectors (necessarily) orthogonal to $\mathcal{S}$. With $(\delta_1, \ldots, \delta_n)$ denoting

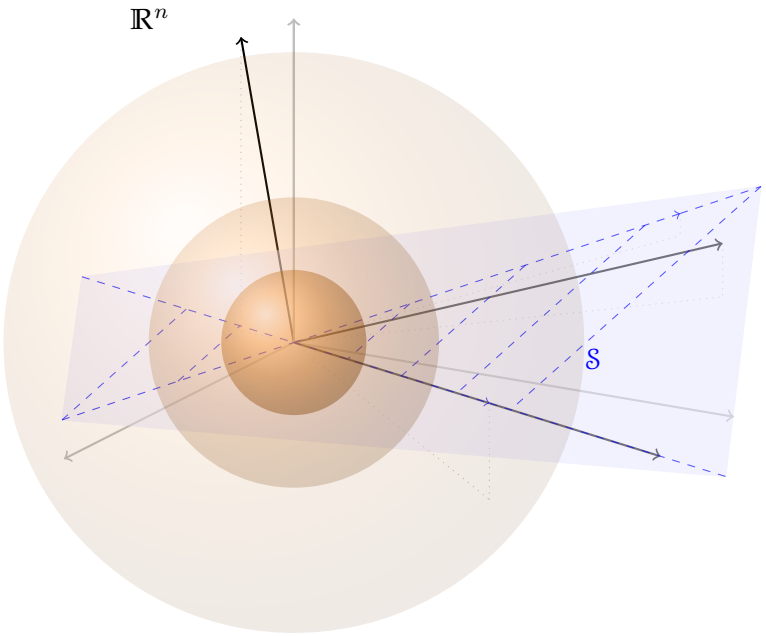


Figure 5.2: What if $\boldsymbol{\epsilon}$ is represented by another choice of orthogonal coordinate axes? By spherical symmetry, we can see that the joint distribution of the coordinate random variables must be the same regardless of the choice of orthogonal coordinate axes. In particular, we consider a choice in which the first $\dim(\mathcal{S})$ axes are chosen from $\mathcal{S}$, and the remaining $n - \dim(\mathcal{S})$ are (necessarily) chosen orthogonally to $\mathcal{S}$. In the figure, the original axes have been grayed out, and new axes are drawn.

the coordinates, the representation

$$\boldsymbol{\epsilon} = \underbrace{\delta_1 \mathbf{u}_1 + \ldots + \delta_{\dim(\mathcal{S})} \mathbf{u}_{\dim(\mathcal{S})}}_{\in \mathcal{S}} + \underbrace{\delta_{\dim(\mathcal{S})+1} \mathbf{u}_{\dim(\mathcal{S})+1} + \ldots + \delta_n \mathbf{u}_n}_{\perp \mathcal{S}}$$

shows that $\delta_1 \mathbf{u}_1 + \ldots + \delta_{\dim(\mathcal{S})} \mathbf{u}_{\dim(\mathcal{S})}$ is the orthogonal projection of $\boldsymbol{\epsilon}$ onto $\mathcal{S}$.

If there are two orthogonal subspaces of interest $\mathcal{S}_1 \perp \mathcal{S}_2$, we can let $\delta_1, \ldots, \delta_{\dim(\mathcal{S}_1)}$ be the coordinates of $\boldsymbol{\epsilon}$ along orthogonal axes spanning $\mathcal{S}_1$, and let $\delta_{\dim(\mathcal{S}_1)+1}, \ldots, \delta_{\dim(\mathcal{S}_1)+\dim(\mathcal{S}_2)}$ be the coordinates for orthogonal axes spanning $\mathcal{S}_2$. All of the coordinates are independent of each other, so the orthogonal projection of $\boldsymbol{\epsilon}$ onto $\mathcal{S}_1$ (which is a function of $\delta_1, \ldots, \delta_{\dim(\mathcal{S}_1)}$) is independent of the orthogonal projection onto $\mathcal{S}_2$ (which is a function of $\delta_{\dim(\mathcal{S}_1)+1}, \ldots, \delta_{\dim(\mathcal{S}_1)+\dim(\mathcal{S}_2)}$). This observation generalizes to any number of orthogonal subspaces: *the orthogonal projections of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$ onto orthogonal subspaces are independent of each other.*

In general, we don't need to explicitly construct an alternative coordinate system. The alternative coordinate system is merely a conceptual tool that enables us to understand the distributions of a variety of statistics in Chapter 6.

5.3. Other transformations

Other transformations of Normal random vectors also result in distributions of interest. Here, we'll review the χ^2, t, and f families along with their "non-central" generalizations. Many of the associated exercises will demonstrate how these distributions can naturally arise in the context of spherically symmetric Normal draws.

5.3.1. χ^2-distributions

For independent standard Normal random variables $Z_1, \ldots, Z_k$, the distribution of $Z_1^2 + \ldots + Z_k^2$ is called the *χ^2-distribution with k degrees of freedom* and is denoted χ_k^2. Realize that the sum of squared iid standard Normal draws can also be thought of directly as the squared length of a draw from the multivariate $N(\mathbf{0}, \mathbb{I})$.

Exercise 5.6

Find the expectation of $W \sim \chi^2_n$.

Exercise 5.7

If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbb{C})$ is an $\mathbb{R}^n$-valued random vector, what's the distribution of the squared Mahalanobis distance of $\mathbf{Y}$ from its own distribution?

Exercise 5.8

Let $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ with iid $N(0, \sigma^2)$ entries, and let $\mathbb{H}$ be the orthogonal projection matrix onto a subspace $\mathcal{S}$. Find the distribution of $\frac{\|\mathbb{H}\boldsymbol{\epsilon}\|^2}{\sigma^2}$.

◇

Solution: The squared length of $\mathbb{H}\boldsymbol{\epsilon}$ is simply the sum of the squared coordinates; with $\delta_1, \ldots, \delta_{\dim(\mathcal{S})}$ representing the coordinates with respect to orthonormal axes spanning $\mathcal{S}$,

$$\|\mathbb{H}\boldsymbol{\epsilon}\|^2 = \delta_1^2 + \ldots + \delta^2_{\dim(\mathcal{S})}.$$

Dividing by σ^2 produces the random variable

$$\begin{aligned}\frac{\|\mathbb{H}\boldsymbol{\epsilon}\|^2}{\sigma^2} &= \frac{\delta_1^2 + \ldots + \delta^2_{\dim(\mathcal{S})}}{\sigma^2} \\ &= \left(\frac{\delta_1}{\sigma}\right)^2 + \ldots + \left(\frac{\delta_{\dim(\mathcal{S})}}{\sigma}\right)^2 \\ &\sim \chi^2_{\dim(\mathcal{S})}\end{aligned}$$

as $\frac{\delta_1}{\sigma}, \ldots, \frac{\delta_{\dim(\mathcal{S})}}{\sigma}$ are iid standard Normal.

❧ ❋ ❧

Now consider $\mathbf{Y} := (Y_1, \ldots, Y_k)$ with independent coordinates $Y_i \sim N(\mu_i, 1)$. Then $\mathbf{Y} \sim N(\boldsymbol{\mu}, \mathbb{I})$ with $\boldsymbol{\mu} := (\mu_1, \ldots, \mu_k)$. Let's think about the distribution of $\|\mathbf{Y}\|^2 = Y_1^2 + \ldots + Y_k^2$. The distribution of $\mathbf{Y}$ looks like the $N(\mathbf{0}, \mathbb{I})$ distribution, only translated by $\boldsymbol{\mu}$. Based on spherical symmetry, we can reason that the distribution of $\|\mathbf{Y}\|^2$ shouldn't depend on the specific direction of $\boldsymbol{\mu}$ but only on its length $\|\boldsymbol{\mu}\|$. Indeed, any sum of k squared independent Normal draws with arbitrary means and each with variance 1 has the [*non-central*] χ^2-distribution with k degrees of freedom and with *non-centrality parameter* equal to the sum of the squared means. For example, the distribution of $\mathbf{Y}$ as defined above can be denoted $\chi^2_{k,\|\boldsymbol{\mu}\|^2}$.

Exercise 5.9

If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbb{C})$ is an $\mathbb{R}^n$-valued random vector, what's the distribution of $\|\mathbb{C}^{-1/2}(\mathbf{Y} - \mathbf{v})\|^2/\sigma^2$?

❧ ❋ ❧

5.3.2. t-distributions

If $Z \sim N(0,1)$ and $V \sim \chi^2_k$ are independent of each other, then the ratio $\frac{Z}{\sqrt{V/k}}$ has the *t-distribution with k degrees of freedom*[1], denoted t_k. As we'll see, many statistics depend on an unknown scaling factor. However, we can often devise a ratio of two independent statistics that are both proportional to that factor, making it cancel out; the distribution of the resulting ratio doesn't depend on the unknown scaling factor, allowing us to know its distribution exactly.

Exercise 5.10

Let $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ with iid $N(0, \sigma^2)$ entries. Let $\mathbb{H}$ be the orthogonal projection matrix onto the subspace $\mathcal{S}$, and let $\mathbf{v} \perp \mathcal{S}$. Find the distribution of $\frac{\langle \boldsymbol{\epsilon}, \mathbf{v}\rangle / \|\mathbf{v}\|}{\|\mathbb{H}\boldsymbol{\epsilon}\| / \sqrt{\dim(\mathcal{S})}}$.

⬦

Solution: Notice that the numerator is simply the coordinate of $\boldsymbol{\epsilon}$ in the direction of the unit vector $\frac{\mathbf{v}}{\|\mathbf{v}\|}$. We can let δ_1 be the coordinate in that direction, and let $\delta_2, \ldots, \delta_{\dim(\mathcal{S})+1}$ be the coordinates with respect to orthogonal axes spanning $\mathcal{S}$. These coordinate random variables are iid $N(0, \sigma^2)$. In Exercise 5.8, we found that $\|\mathbb{H}\boldsymbol{\epsilon}\|^2/\sigma^2 \sim \chi^2_k$. We can produce this random variable if we divide both the numerator and the denominator by σ.

$$\begin{aligned}\frac{\langle \boldsymbol{\epsilon}, \mathbf{v}\rangle / \|\mathbf{v}\|}{\|\mathbb{H}\boldsymbol{\epsilon}\| / \sqrt{\dim(\mathcal{S})}} &= \frac{(\langle \boldsymbol{\epsilon}, \mathbf{v}\rangle / \|\mathbf{v}\|)/\sigma}{(\|\mathbb{H}\boldsymbol{\epsilon}\| / \sqrt{\dim(\mathcal{S})})/\sigma} \\ &= \frac{\delta_1/\sigma}{\sqrt{(\|\mathbb{H}\boldsymbol{\epsilon}\|^2/\sigma^2)/\dim(\mathcal{S})}}\end{aligned}$$

Because $\frac{\delta_1}{\sigma} \sim N(0,1)$ is independent of the denominator, this expression matches the definition of the $t_{\dim(\mathcal{S})}$-distribution.

❧ ❋ ❧

If the independent numerator is instead $Y \sim N(\mu, 1)$, then the distribution of the ratio $\frac{Y}{\sqrt{V/k}}$ is called *non-central t* with k degrees of freedom and *non-centrality parameter* μ, denoted $t_{k,\mu}$.

Exercise 5.11

Suppose $Y \sim N(\mu, \sigma^2 s^2)$ and $W/\sigma^2 \sim \chi^2_k$ are independent of each other, with $\sigma \in \mathbb{R}$ non-random. For $a \in \mathbb{R}$, what's the distribution of $\frac{(Y-a)/s}{\sqrt{W/k}}$?

❧ ❋ ❧

5.3.3. f-distributions

If $V \sim \chi^2_k$ and $W \sim \chi^2_m$ are independent of each other, then the ratio $\frac{V/k}{W/m}$ has the *f-distribution with k numerator degrees of freedom and m denominator degrees of freedom*, denoted $f_{k,m}$.

[1] Notice that V/k can be represented as an average $\frac{1}{k}(Z_1^2 + \ldots + Z_k^2)$ with $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$. By the law of large numbers, V/k becomes increasingly concentrated near its expectation of 1 as k increases, so the distribution of $\frac{Z}{\sqrt{V/k}}$ increasingly resembles the distribution of $Z \sim N(0,1)$.

Exercise 5.12

Let $T \sim t_k$. What's the distribution of T^2?

Exercise 5.13

Let $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ with iid $N(0, \sigma^2)$ entries. Let $\mathbb{H}_1$ and $\mathbb{H}_2$ be the orthogonal projection matrices onto subspaces $\mathcal{S}_1 \perp \mathcal{S}_2$. Find the distribution of $\frac{\|\mathbb{H}_1\boldsymbol{\epsilon}\|^2/\dim(\mathcal{S}_1)}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\dim(\mathcal{S}_2)}$.

$\diamond$

Solution: Divide both the numerator and the denominator by σ^2 to produce random variables whose distributions we know from Exercise 5.8.

$$\frac{\|\mathbb{H}_1\boldsymbol{\epsilon}\|^2/\dim(\mathcal{S}_1)}{\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\dim(\mathcal{S}_2)} = \frac{(\|\mathbb{H}_1\boldsymbol{\epsilon}\|^2/\sigma^2)/\dim(\mathcal{S}_1)}{(\|\mathbb{H}_2\boldsymbol{\epsilon}\|^2/\sigma^2)/\dim(\mathcal{S}_2)}$$

The numerator is a $\chi^2_{\dim(\mathcal{S}_1)}$-distributed random variable divided by its degrees of freedom, while the denominator is a $\chi^2_{\dim(\mathcal{S}_2)}$-distributed random variable divided by its degrees of freedom. Because $\mathcal{S}_1 \perp \mathcal{S}_2$, we know that the two orthogonal projections are independent of each other, allowing us to conclude that the ratio matches the definition of $f_{\dim(\mathcal{S}_1),\dim(\mathcal{S}_2)}$.

If the numerator is instead *non-central* $V \sim \chi^2_{k,\|\boldsymbol{\mu}\|^2}$, then the distribution of the ratio $\frac{V/k}{W/m}$ is called *non-central* f with k numerator degrees of freedom, m denominator degrees of freedom, and *non-centrality parameter* $\|\boldsymbol{\mu}\|^2$, denoted $f_{k,m,\|\boldsymbol{\mu}\|^2}$.

Exercise 5.14

If $\mathbf{Y} \sim N(\boldsymbol{\mu}, \sigma^2\mathbb{C})$ is an $\mathbb{R}^n$-valued random vector and $W/\sigma^2 \sim \chi^2_k$ is independent of $\mathbf{Y}$, what is the distribution of $\frac{\|\mathbb{C}^{-1/2}(\mathbf{Y}-\mathbf{v})\|^2/n}{W/k}$?

Homework 8: Simulating Normality

Work through *simulating-normality.Rmd* to verify some of the properties of Normal random vectors and their relationships to other important distributions.

Now that we've developed intuition and tools for the multivariate Normal distribution, we can put them into practice in the context of linear modeling with the assumption that the errors are Normal.

CHAPTER

6

NORMAL ERRORS

THROUGHOUT THIS CHAPTER, we'll continue analyzing the linear model from Chapter 4, but we'll assume in particular that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$ for an unknown σ^2. The figures in Chapter 4 already indicated iid Normal errors, although the results we derived in that chapter didn't require such strong assumptions on the distribution of the errors. With the new Normality assumption, our earlier results remain valid of course, but we'll also be able to do a good deal more in terms of inference, including hypothesis tests and confidence sets.

Let's begin with a few warm-up exercises to start connecting ideas from Chapters 4 and 5.

Exercise 6.1

Let m be the number of explanatory variables, and let $x_i^{(j)}$ represent the value of the ith observation of the jth explanatory variable. Consider modeling the response variables by

$$Y_i = f_\theta(x_i^{(1)}, \ldots, x_i^{(m)}) + \epsilon_i$$

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ and $\theta \in \Theta$ indexing a set of possible functions. (Notice that this form is far more general than the linear model with iid Normal errors.) Show that the maximum likelihood estimator for θ is precisely the parameter that minimizes the sum of squared residuals.

$\diamond$

Solution: The response values have distribution $Y_i \sim N(f_\theta(x_i^{(1)}, \ldots, x_i^{(m)}), \sigma^2)$ and are independent of each other. Because of independence, the overall likelihood $L(\theta; \mathbf{Y})$ is the product of the individual observations' likelihoods.

$$\begin{aligned} L(\theta; \mathbf{Y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(Y_i - f_\theta(x_i^{(1)}, \ldots, x_i^{(m)}))^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - f_\theta(x_i^{(1)}, \ldots, x_i^{(m)}))^2} \end{aligned}$$

The parameter θ only appears in the sum of squared residuals $\sum_{i=1}^{n}(Y_i - f_\theta(x_i^{(1)}, \ldots, x_i^{(m)}))^2$. The smaller the sum of squared residuals is, the larger the likelihood is, so the "least-squares

parameter" is exactly the maximum likelihood estimator. Notice that this equivalence doesn't depend on the value of σ and that it holds even if σ is unknown.

Exercise 6.2

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbb{C})$. What's the distribution of $\mathbf{Y}$?

◇

Solution: Normality is preserved by the translation, so $\mathbf{Y} \sim N(\mathbb{X}\boldsymbol{\beta}, \mathbb{C})$.

Exercise 6.3

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. What's the distribution of the least-squares predictions $\hat{\mathbf{Y}}$?

Exercise 6.4

Assuming $\mathbb{X}$ has d linearly independent columns, explain how we can tell that $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ has linearly independent rows.

Exercise 6.5

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. Assuming the design matrix $\mathbb{X}$ has d linearly independent columns, what's the distribution of the least-squares coefficient vector $\hat{\boldsymbol{\beta}}$?

◇

Solution: Least least-squares coefficient vector $\hat{\boldsymbol{\beta}} := (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbf{Y}$ is a linear transformation of $\mathbf{Y}$ which is Normal. Exercise 6.4 showed that this transformation has full row rank, so we know that $\hat{\boldsymbol{\beta}}$ is Normal on $\mathbb{R}^d$. The expectation and covariance, were worked out in Exercises 4.7 and 4.8, so we see in particular that $\hat{\boldsymbol{\beta}} \sim N(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$.

Exercise 6.6

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. If $\hat{\mathbf{Y}}$ is the least-squares prediction vector, what is the distribution of $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/\sigma^2$?

◇

Solution: We saw in Chapter 4 that the residual vector $\mathbf{Y} - \hat{\mathbf{Y}}$ is the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{X})^\perp$ which has dimension $n - \text{rank}(\mathbb{X})$. Therefore, according to Exercise 5.8, $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/\sigma^2 \sim \chi^2_{n-\text{rank}(\mathbb{X})}$.

6.1. Independence of least-squares coefficients and residuals

A crucial observation that we'll need throughout this chapter is that the least-squares coefficients are independent of the residuals when the errors are iid Normal. To verify this fact, we'll first establish that $\mathbf{Y} - \hat{\mathbf{Y}}$ is independent of $\hat{\mathbf{Y}}$, then we'll see that the least-squares coefficients can be written as a (non-random) function of $\hat{\mathbf{Y}}$ so that they inherit the independence.

To prove that $\hat{\mathbf{Y}}$ is independent of $\mathbf{Y} - \hat{\mathbf{Y}}$, we can make reference to the fact that orthogonal

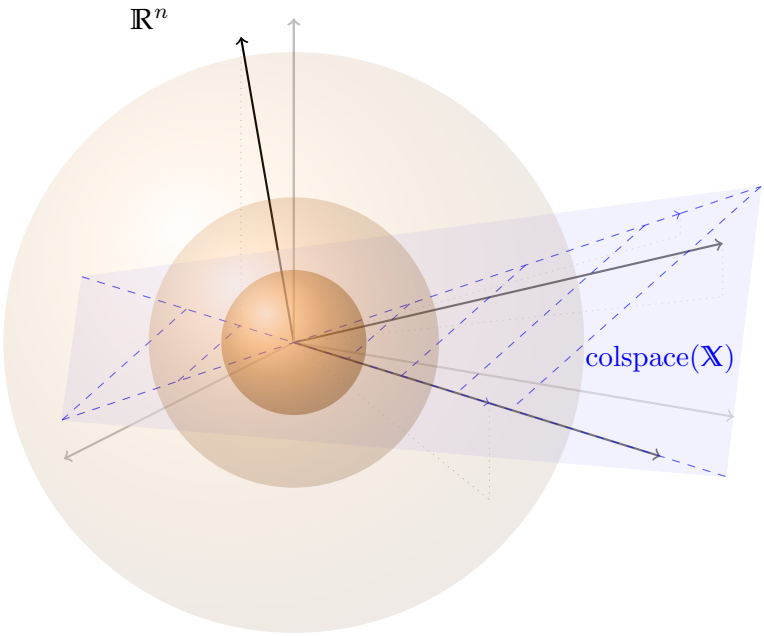


Figure 6.1: What if $\boldsymbol{\epsilon}$ is represented by another choice of orthogonal coordinate axes? By spherical symmetry, we can see that the joint distribution of the coordinate random variables must be the same regardless of the choice of orthogonal coordinate axes. In particular, one can consider a choice in which the first rank($\mathbb{X}$) axes are chosen from the column space of $\mathbb{X}$, and the remaining $n - $ rank($\mathbb{X}$) are (necessarily) chosen orthogonally to the column space of $\mathbb{X}$. In the figure, the original axes have been grayed out, and new axes are drawn.

projections of $\boldsymbol{\epsilon}$ onto orthogonal subspaces are independent of each other (see Section 5.2). We saw in Chapter 4 that the residual vector is the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{X})^\perp$. With $\mathbb{H}$ representing the orthogonal projection matrix onto $C(\mathbb{X})$, vector of least-squares predictions is

$$\begin{aligned}\hat{\mathbf{Y}} &= \mathbb{H}\mathbf{Y} \\ &= \mathbb{H}(\mathbb{E}\mathbf{Y} + \boldsymbol{\epsilon}) \\ &= \mathbb{H}\mathbb{E}\mathbf{Y} + \mathbb{H}\boldsymbol{\epsilon}.\end{aligned}$$

Its randomness only depends on the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{X})$, so it must be independent of the orthogonal projection of $\boldsymbol{\epsilon}$ onto $C(\mathbb{X})^\perp$. From Exercise 2.15, we can conclude that the randomness in $\hat{\beta}$ only depends on $\hat{\mathbf{Y}}$, so it inherits independence with the residual vector and in particular with $\hat{\sigma}^2$.

6.2. Inference for coefficients

The context for this section will be multiple linear regression with m explanatory variables, a full-rank design matrix (rank $m+1$), and iid Normal errors. Practitioners are generally more interested in the coefficients of the explanatory variables than in the intercept, so while the model and least-squares procedure do fit an intercept, we'll focus our study on inference tasks for the other coefficients. Recalling from Chapter 4 the expectation and covariance of the explanatory variables' least-squares coefficients, we conclude that with iid Normal errors, $(\hat{\beta}_1, \ldots, \hat{\beta}_m) \sim N((\beta_1, \ldots, \beta_m), \frac{\sigma^2}{n}\Sigma^{-1})$.

6.2.1. A single coefficient

The marginal distribution of any single least-squares coefficient can be read off from their joint distribution. For $j \in \{1, \ldots, m\}$,

$$\hat{\beta}_j \sim N(\beta_j, \frac{\sigma^2}{n}\Sigma^{-1}_{jj}).$$

The standardized version is

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{\Sigma^{-1}_{jj}/n}} \sim N(0,1).$$

The appearance of the unknown error variance σ prevents us from using this random variable for testing hypotheses or constructing confidence intervals. However, if we substitute $\hat{\sigma}$ (which by Section 6.1 is independent of $\hat{\beta}_j$) for σ,

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{\Sigma^{-1}_{jj}/n}} = \frac{(\hat{\beta}_j - \beta_j)/(\sigma\sqrt{\Sigma^{-1}_{jj}/n})}{\hat{\sigma}/\sigma}$$
$$\sim t_{n-m-1}$$

because $\hat{\sigma}/\sigma = \sqrt{\frac{\|\mathbf{Y}-\hat{\mathbf{Y}}\|^2/\sigma^2}{n-m-1}}$ is the square root of a χ^2-distributed random variable (Exercise 6.6) divided by its degrees of freedom.

Given a hypothesized value for β_j (usually 0), let T_j denote the resulting t-statistic; the significance probability is $2G(-|T_j|)$, where G is the cdf of the t_{n-m-1} distribution.[1]

Alternatively, we can use our t_{n-m-1}-distributed random variable[2] to derive confidence intervals for the true coefficient's value. For example, with $t^*_{.975}$ representing the .975-quantile of the t_{n-m-1} distribution,

$$\mathbb{P}\left\{-t^*_{.975} \le \frac{\beta_j - \hat{\beta}_j}{\hat{\sigma}\sqrt{\Sigma^{-1}_{jj}/n}} \le t^*_{.975}\right\} = .95$$
$$\Updownarrow$$
$$\mathbb{P}\left\{\hat{\beta}_j - t^*_{.975}\hat{\sigma}\sqrt{\Sigma^{-1}_{jj}/n} \le \beta_j \le \hat{\beta}_j + t^*_{.975}\hat{\sigma}\sqrt{\Sigma^{-1}_{jj}/n}\right\} = .95$$

Thus $\hat{\beta}_j \pm t^*_{.975}\hat{\sigma}\sqrt{\Sigma^{-1}_{jj}/n}$ is a 95% confidence interval for β_j.

6.2.2. All coefficients together

Next, let's consider all of the explanatory variables' coefficients together. We'll parallel the logic of the previous section by first standardizing the random vector:

$$\tfrac{\sqrt{n}}{\sigma}\Sigma^{1/2}\left(\begin{bmatrix}\hat{\beta}_1\\ \vdots \\ \hat{\beta}_m\end{bmatrix} - \begin{bmatrix}\beta_1\\ \vdots \\ \beta_m\end{bmatrix}\right) \sim N(\mathbf{0}, \mathbb{I}).$$

The squared norm of this random vector is χ^2_m-distributed. As before, the unknown error variance prevents us from making direct use of these quantities for inference. Also as before, when we use $\hat{\sigma}$ rather than σ, things work out neatly because the least-squares coefficients are independent of $\hat{\sigma}^2$.

$$\frac{n\left\|\Sigma^{1/2}\left(\begin{bmatrix}\hat{\beta}_1\\ \vdots \\ \hat{\beta}_m\end{bmatrix} - \begin{bmatrix}\beta_1\\ \vdots \\ \beta_m\end{bmatrix}\right)\right\|^2/m}{\hat{\sigma}^2} = \frac{\left\|\frac{\sqrt{n}}{\sigma}\Sigma^{1/2}\left(\begin{bmatrix}\hat{\beta}_1\\ \vdots \\ \hat{\beta}_m\end{bmatrix} - \begin{bmatrix}\beta_1\\ \vdots \\ \beta_m\end{bmatrix}\right)\right\|^2/m}{\hat{\sigma}^2/\sigma^2}$$
$$\sim f_{m,n-m-1}$$

[1] One-sided tests can similarly be performed.

[2] The confidence interval derivation actually replaces the original random variable with its negative, which is also t_{n-m-1}-distributed due to the symmetry of t distributions.

Given a hypothesized value for $(\beta_1, \ldots, \beta_m)$ (usually $\mathbf{0}$), let F denote the resulting f-statistic; the significance probability is $1 - G(F)$, where G is the cdf of the $f_{m,n-m-1}$ distribution.[3]

Alternatively, we can use our $f_{m,n-m-1}$-distributed random variable[4] to derive confidence sets for the true coefficient vector. For example, with $f^*_{.95}$ representing the .95-quantile of the $f_{m,n-m-1}$ distribution,

$$\mathbb{P}\left\{\left(n\left\|\Sigma^{1/2}\left(\begin{bmatrix}\beta_1\\ \vdots \\ \beta_m\end{bmatrix} - \begin{bmatrix}\hat{\beta}_1\\ \vdots \\ \hat{\beta}_m\end{bmatrix}\right)\right\|^2 /m\right) /\hat{\sigma}^2 \leq f^*_{.95}\right\} = .95$$

$$\Updownarrow$$

$$\mathbb{P}\left\{\left\|\frac{\sqrt{n}}{\hat{\sigma}}\Sigma^{1/2}\left(\begin{bmatrix}\beta_1\\ \vdots \\ \beta_m\end{bmatrix} - \begin{bmatrix}\hat{\beta}_1\\ \vdots \\ \hat{\beta}_m\end{bmatrix}\right)\right\|^2 \leq m f^*_{.95}\right\} = .95$$

Thus, we have a 95% confidence ellipsoid for $(\beta_1, \ldots, \beta_m)$ comprising the set of vectors whose squared Mahalanobis distance from $N((\hat{\beta}_1, \ldots, \hat{\beta}_m), \frac{\hat{\sigma}^2}{n}\Sigma^{-1})$ is no greater than $m f^*_{.95}$. (Notice that $N((\hat{\beta}_1, \ldots, \hat{\beta}_m), \frac{\hat{\sigma}^2}{n}\Sigma^{-1})$ is a sensible estimate for the distribution of $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$; it comes from simply substituting our estimates for the true distribution's unknown parameters.)

There's a dual relationship between the hypothesis tests and confidence ellipsoids here. A hypothesized coefficient vector $\boldsymbol{\beta}_{\text{null}}$ will be rejected by the .05-level test if and only if it lies outside of the 95% confidence ellipsoid.

Homework 9: Inference basics

Try out the basic inference tasks on a real dataset by working through *inference.Rmd*. In this exercise, you'll also see why using the confidence ellipsoids is preferable to simply looking at the coefficients' confidence intervals separately.

6.2.3. Prediction

It may not be the coefficients themselves that are important. Rather, the purpose of linear modeling is often to *predict* future response values using future explanatory values. Consider the $(n+1)$st observation, generated according to the same mechanism as the original data:

$$Y_{n+1} = \beta_0 + \beta_1 x^{(1)}_{n+1} + \ldots + \beta_m x^{(m)}_{n+1} + \epsilon_{n+1}$$

where $\epsilon_{n+1} \sim N(0, \sigma^2)$ is independent of the previous errors $\epsilon_1, \ldots, \epsilon_n$.

An obvious choice for predicting Y_{n+1} is to use the least-squares coefficients $\hat{Y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 x^{(1)}_{n+1} + \ldots + \hat{\beta}_m x^{(m)}_{n+1}$.

[3] This single test of the coefficients simultaneously is not the same as testing all of the coefficients separately. Performing multiple different tests results in an overall false positive probability that is larger than the false positive rates of the individual tests.

[4] The confidence set derivation actually replaces the original random vector with its negative which has the exact same squared length.

Exercise 6.7

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is an $\mathbb{R}^{n+1}$-valued random vector with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. If $\hat{\boldsymbol{\beta}}$ denotes the vector of least-squares coefficients based on the first n observations. With $\hat{Y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}^{(1)} + \ldots + \hat{\beta}_m x_{n+1}^{(m)}$, show that the expected squared prediction error $\mathbb{E}(Y_{n+1} - \hat{Y}_{n+1})^2$ decomposes into the risk of estimating $\mathbb{E}Y_{n+1}$ plus the variance of $\hat{Y}_{n+1}$ (which is an inherently unpredictable portion).

Exercise 6.8

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is an $\mathbb{R}^{n+1}$-valued random vector with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. If $\hat{\boldsymbol{\beta}}$ denotes the vector of least-squares coefficients based on the first n observations, show that $\hat{Y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}^{(1)} + \ldots + \hat{\beta}_m x_{n+1}^{(m)}$ is an unbiased estimator for $\mathbb{E}Y_{n+1}$.

From Exercises 6.7 and 6.8, we see that the expected squared difference between Y_{n+1} and its prediction $\hat{Y}_{n+1}$ equals the sum of the variance of Y_{n+1}, which is σ^2, and the variance of $\hat{Y}_{n+1}$. To derive a meaningful expression for the variance of $\hat{Y}_{n+1}$, it's advantageous to use an alternative parameterization for $C(\mathbb{X})$ that involves the centered versions of the explanatory variables (see Section 2.2.3). With this parameterization,

$$\hat{Y}_{n+1} = \hat{\alpha}_0 + \hat{\beta}_1(x_{n+1}^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x_{n+1}^{(m)} - \bar{x}^{(m)})$$

with $\hat{\alpha}_0 = \hat{\beta}_0 + \hat{\beta}_1\bar{x}^{(1)} + \ldots + \hat{\beta}_m\bar{x}^{(m)}$. Exercise 4.10 shows that the covariance of $(\hat{\alpha}_0, \hat{\beta}_1, \ldots, \hat{\beta}_m)$ is the block-diagonal matrix $\frac{\sigma^2}{n}\begin{bmatrix} 1 & \\ & \Sigma^{-1} \end{bmatrix}$. In particular, $\hat{\alpha}_0$ is uncorrelated with $(\hat{\beta}_1, \ldots, \hat{\beta}_m)$, so

$$\begin{aligned} \operatorname{var}\hat{Y}_{n+1} &= \operatorname{var}[\hat{\alpha}_0 + \hat{\beta}_1(x_{n+1}^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x_{n+1}^{(m)} - \bar{x}^{(m)})] \\ &= \underbrace{\operatorname{var}\hat{\alpha}_0}_{\sigma^2/n} + \operatorname{var}[\hat{\beta}_1(x_{n+1}^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x_{n+1}^{(m)} - \bar{x}^{(m)})]. \end{aligned}$$

The variance of a random variable can equivalently be considered the 1×1 covariance matrix of an $\mathbb{R}^1$-valued random vector. When we write the variance of the second term as a covariance, we can apply our formula to pull out a non-random $1 \times m$ matrix.

$$\begin{aligned} &\operatorname{var}[\hat{\beta}_1(x_{n+1}^{(1)} - \bar{x}^{(1)}) + \ldots + \hat{\beta}_m(x_{n+1}^{(m)} - \bar{x}^{(m)})] \\ &\quad = \operatorname{cov}\begin{bmatrix} x_{n+1}^{(1)} - \bar{x}^{(1)} & \ldots & x_{n+1}^{(m)} - \bar{x}^{(m)} \end{bmatrix}\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} \\ &\quad = \begin{bmatrix} x_{n+1}^{(1)} - \bar{x}^{(1)} & \ldots & x_{n+1}^{(m)} - \bar{x}^{(m)} \end{bmatrix}\left(\operatorname{cov}\begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix}\right)\begin{bmatrix} x_{n+1}^{(1)} - \bar{x}^{(1)} \\ \vdots \\ x_{n+1}^{(m)} - \bar{x}^{(m)} \end{bmatrix} \\ &\quad = \underbrace{\frac{\sigma^2}{n}\begin{bmatrix} x_{n+1}^{(1)} - \bar{x}^{(1)} & \ldots & x_{n+1}^{(m)} - \bar{x}^{(m)} \end{bmatrix}\Sigma^{-1}\begin{bmatrix} x_{n+1}^{(1)} - \bar{x}^{(1)} \\ \vdots \\ x_{n+1}^{(m)} - \bar{x}^{(m)} \end{bmatrix}}_{\text{“}d_{m+1}\text{”}} \end{aligned}$$

Notice that the part of the expression labeled “d_{m+1}” is precisely the squared Mahalanobis distance from the $(n+1)$st observation's vector of explanatory values to the empirical distribution

of the explanatory variables of the original n observations. This has a beautifully intuitive interpretation: *the less the* $(n+1)$*st observation resembles the original data, the more challenging it will be to predict its response value.*

Putting the terms of our derivation together,

$$\begin{aligned}\mathbb{E}(Y_{n+1} - \hat{Y}_{n+1})^2 &= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{n} d_{m+1} \\ &= \sigma^2[1 + \frac{1}{n}(1 + d_{m+1})].\end{aligned}$$

Notice that there is a limit to how well a new response can be predicted; no matter how large the sample size is, this quantity remains larger than σ^2.

In fact, $Y_{n+1} - \hat{Y}_{n+1}$ is a difference of independent Normal random variables, so we know more specifically that it's distribution is $N(0, \sigma^2[1+\frac{1}{n}(1+d_{m+1})])$. Using steps analogous to our earlier confidence interval derivations, we can actually formulate *prediction intervals* for $\hat{Y}_{n+1}$.

The standardized version of prediction error is

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sigma\sqrt{1 + \frac{1}{n}(1 + d_{m+1})}} \sim N(0, 1).$$

Because $\hat{\sigma^2}$ is independent of ϵ_{n+1} and of the least-squares coefficients, we can follow the same pattern as before, substituting $\hat{\sigma}$ for σ.

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{1 + \frac{1}{n}(1 + d_{m+1})}} \sim t_{n-m-1}$$

With $t^*_{.975}$ representing the .975-quantile of the t_{n-m-1} distribution,

$$\mathbb{P}\left\{-t^*_{.975} \leq \frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{1 + \frac{1}{n}(1 + d_{m+1})}} \leq t^*_{.975}\right\} = .95$$

$$\Updownarrow$$

$$\mathbb{P}\left\{\hat{Y}_{n+1} - t^*_{.975}\hat{\sigma}\sqrt{1 + \frac{1}{n}(1 + d_{m+1})} \leq Y_{n+1} \leq \hat{Y}_{n+1} + t^*_{.975}\hat{\sigma}\sqrt{1 + \frac{1}{n}(1 + d_{m+1})}\right\} = .95$$

Thus $\hat{Y}_n + 1 \pm t^*_{.975}\hat{\sigma}\sqrt{1 + \frac{1}{n}(1 + d_{m+1})}$ is a 95% prediction interval for Y_{n+1}.

Exercise 6.9

Suppose $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is an $\mathbb{R}^{n+1}$-valued random vector with error vector $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbb{I})$. Derive a 95% confidence interval for $\mathbb{E}Y_{n+1}$ as a function of $Y_1, \ldots, Y_n$.

$\diamond$

Solution: We've determined that $\hat{Y}_{n+1}$ is Normal and found its expectation and variance. Standardizing, we have

$$\frac{\hat{Y}_{n+1} - \mathbb{E}Y_{n+1}}{\sigma\sqrt{\frac{1}{n}(1 + d_{m+1})}} \sim N(0, 1).$$

Because $\hat{\sigma}^2$ is independent of this random variable,

$$\frac{\hat{Y}_{n+1} - \mathbb{E}Y_{n+1}}{\hat{\sigma}\sqrt{\frac{1}{n}(1+d_{m+1})}} = \frac{(\hat{Y}_{n+1} - \mathbb{E}Y_{n+1})/(\sigma\sqrt{\frac{1}{n}(1+d_{m+1})})}{\hat{\sigma}/\sigma}$$
$$\sim t_{n-m-1}$$

where d_{m+1} is the squared Mahalanobis distance from the new observations explanatory values to the empirical distribution of the first n observations' explanatory values. Therefore, with $t^*_{.975}$ representing the .975-quantile of the t_{n-m-1} distribution,

$$\mathbb{P}\left\{-t^*_{.975} \leq \frac{\mathbb{E}Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{1+\frac{1}{n}(1+d_{m+1})}} \leq t^*_{.975}\right\} = .95$$
$$\Updownarrow$$
$$\mathbb{P}\left\{\hat{Y}_{n+1} - t^*_{.975}\hat{\sigma}\sqrt{\frac{1}{n}(1+d_{m+1})} \leq \mathbb{E}Y_{n+1} \leq \hat{Y}_{n+1} + t^*_{.975}\hat{\sigma}\sqrt{\frac{1}{n}(1+d_{m+1})}\right\} = .95$$

shows that $\hat{Y}_{n+1} \pm t^*_{.975}\hat{\sigma}\sqrt{\frac{1}{n}(1+d_{m+1})}$ constitutes a 95% confidence interval for $\mathbb{E}Y_{n+1}$.

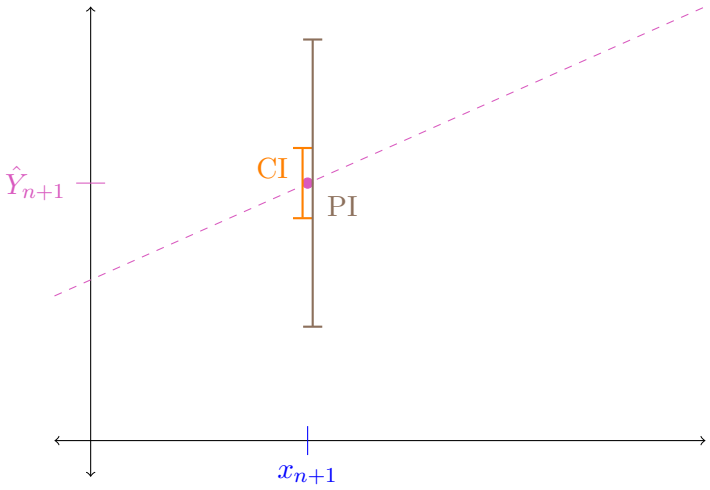


Figure 6.2: A comparison of a confidence interval (CI) and a prediction interval (PI) is drawn in the context of simple linear modeling. The confidence interval for $\mathbb{E}\mathbf{Y}_{n+1}$ and prediction interval for Y_{n+1} are both centered at the least-squares line's value at the new explanatory value x_{n+1}. The prediction interval necessarily extends further, and its width doesn't decrease to zero.

6.3. General testing of subspaces

The hypothesis that some coefficient β_j is exactly the same as the hypothesis that $\mathbb{E}\mathbf{Y}$ lies in the span of the other columns of the design matrix. Similarly, the hypothesis that $\beta_1, \ldots, \beta_m$ are all 0 is the same as the hypothesis that $\mathbb{E}\mathbf{Y}$ lies in the span of $\mathbf{1}$. While these two hypotheses can easily be incorporated into our earlier derivations from Sections 6.2.1 and 6.2.2, there's a general formula for testing the hypothesis that $\mathbb{E}\mathbf{Y}$ lies in any particular subspace of $C(\mathbb{X})$.

Let $\mathcal{S}$ be a subspace of $C(\mathbb{X})$, and let $\mathbb{H}$ and $\mathbb{H}_\mathcal{S}$ be the orthogonal projection matrices onto $C(\mathbb{X})$ and $\mathcal{S}$. Let $\widetilde{\mathbf{Y}}$ denote the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{S}$. The difference between $\hat{\mathbf{Y}}$

and $\widetilde{\mathbf{Y}}$ is

$$\begin{aligned}\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}} &= \mathbb{H}\mathbf{Y} - \mathbb{H}_{\mathcal{S}}\mathbf{Y} \\ &= (\mathbb{H} - \mathbb{H}_{\mathcal{S}})\mathbf{Y} \\ &= (\mathbb{H} - \mathbb{H}_{\mathcal{S}})(\mathbb{E}\mathbf{Y} + \boldsymbol{\epsilon}) \\ &= (\mathbb{H} - \mathbb{H}_{\mathcal{S}})\mathbb{E}\mathbf{Y} + (\mathbb{H} - \mathbb{H}_{\mathcal{S}})\boldsymbol{\epsilon}. \end{aligned} \tag{6.1}$$

From Exercise 1.73, $\mathbb{H} - \mathbb{H}_{\mathcal{S}}$ is the orthogonal projection matrix onto the orthogonal complement of $\mathcal{S}$ within $C(\mathbb{X})$. This subspace is in $C(\mathbb{X})$, so, assuming iid Normal errors, we know that $(\mathbb{H} - \mathbb{H}_{\mathcal{S}})\boldsymbol{\epsilon}$ must be independent of $\hat{\sigma}^2$.

Now assume that $\mathbb{E}\mathbf{Y} \in \mathcal{S}$. Then the orthogonal projection of $\mathbb{E}\mathbf{Y}$ onto the orthogonal complement of $\mathcal{S}$ within $C(\mathbb{X})$ is $\mathbf{0}$, so

$$\begin{aligned}\frac{\|\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}}\|^2}{\sigma^2} &= \frac{\|(\mathbb{H} - \mathbb{H}_{\mathcal{S}})\mathbb{E}\mathbf{Y} + (\mathbb{H} - \mathbb{H}_{\mathcal{S}})\boldsymbol{\epsilon}\|^2}{\sigma^2} \\ &= \frac{\|(\mathbb{H} - \mathbb{H}_{\mathcal{S}})\boldsymbol{\epsilon}\|^2}{\sigma^2} \\ &\sim \chi^2_{\dim C(\mathbb{X}) - \dim \mathcal{S}}\end{aligned}$$

Because σ^2 is unknown, we need to use the same *ratio trick* as before to obtain a useful test statistic.

$$\frac{\|\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}}\|^2/(\dim C(\mathbb{X}) - \dim \mathcal{S})}{\hat{\sigma}^2} \sim f_{\dim C(\mathbb{X}) - \dim \mathcal{S}, n - \dim C(\mathbb{X})}$$

This provides a formula for testing any null hypothesis of the form $\mathbb{E}\mathbf{Y} \in \mathcal{S}$. (Notice that it doesn't require the design matrix to be full-rank or to include span$\{\mathbf{1}\}$ in its column space.) Most often it is applied to test the hypothesis that some particular subset of the coefficients are all zero.

Exercise 6.10

Our two approaches to linear model hypothesis testing can both be used to test that the coefficients $\beta_1, \ldots, \beta_m$ are all zero. Verify that the two test statistics

$$\frac{n \left\| \Sigma^{1/2} \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_m \end{bmatrix} \right\|^2 / m}{\hat{\sigma}^2} \qquad \text{and} \qquad \frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2/m}{\hat{\sigma}^2}$$

are equal to each other.

6.4. Analysis of variance

The one-way analysis of variance (ANOVA) test can be applied when the model involves only one explanatory variable and it is categorical.[5] We'll write the model

$$Y_i = \mu_1 \mathbb{I}\{x_i = 1\} + \ldots + \mu_k \mathbb{I}\{x_i = k\} + \epsilon_i$$

[5] With two categorical explanatory variables, two *two-way ANOVA* is analogous but a bit more complicated. There are several variations of two-way ANOVA, and each can be understood geometrically using the approach you've learned in this text. Likewise, inference on combinations of categorical and quantitative explanatory variables can be understood in our geometric framework, but for now at least, the specific details are beyond this book's scope.

with $\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} N(0, \sigma^2)$. The design matrix $\mathbb{X}$ has k columns indicating group membership as in Chapter 2.

The null hypothesis of ANOVA is that the groups all share the same expectation $\mu_1 = \ldots = \mu_k$. If we recognize that this null hypothesis is equivalent to the claim that $\mathbb{E}\mathbf{Y}$ is in the span of $\mathbf{1}$, then we can understand it as an instance of the form that was just described in Section 6.3. With k groups, the dimension of the design space is k, so the f-statistic is

$$\frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2/(k-1)}{\hat{\sigma}^2} \sim f_{k-1,n-k}.$$

This statistic is often written as $\frac{\mathrm{SS}_{\mathrm{reg}}/(k-1)}{\mathrm{SS}_{\mathrm{res}}/(n-k)}$ where $\mathrm{SS}_{\mathrm{reg}}$ means *regression sum of squares* and $\mathrm{SS}_{\mathrm{res}}$ means *residual sum of squares*. Recall that summing over groups rather than observations (see Section 2.3) provides alternative expressions for these quantities:

$$\mathrm{SS}_{\mathrm{reg}} = \sum_{j=1}^{k} n_j(\bar{Y}_j - \bar{Y})^2 \qquad \text{and} \qquad \mathrm{SS}_{\mathrm{res}} = \sum_{j=1}^{k} \sum_{i:x_i=j} (Y_i - \bar{Y}_j)^2.$$

When the ANOVA test rejects its null hypothesis, you conclude that the groups *don't* all share the same expectation. However, simply establishing that the groups are different isn't necessarily very useful. You'll likely want to make more specific comparisons among the groups. It's easy to test for equality of expectation for every pair of groups. This would produce $\binom{k}{2} = \frac{k^2-k)}{2}$ significance probabilities, which grows quadratically with the number of groups. Whenever multiple different tests are performed, the *overall* false positive rate is larger than the false positive rate of any of the individual tests. To keep the overall false positive rate under control and limit the likely number of false positives, practitioners use a combination of sequential testing (i.e. some of the tests are only performed if a preceding test rejects its null hypothesis), limiting the total number of tests, and using extra small false positive probabilities for the individual tests.

To this end, innumerable such schemes have been suggested. The most conservative approach is the *Bonferonni correction* method: the sum of the individual tests' false positive probabilities should equal the largest overall false positive rate that you're willing to tolerate. For example, if you perform m tests, each with false positive probability α/m, then the overall false positive rate is no greater than α. On the other hand, if the m test statistics are *independent* of each other, then *Sidak's correction* can be applied: a false positive rate of $1 - (1-\alpha)^{1/m}$ is used for each of the tests, resulting in an overall false positive rate of exactly α.

With one categorical explanatory variable and balanced groups, a preferred method for multiple testing is to use *orthogonal contrasts*. Let $\mathbf{c} := (c_1, \ldots, c_k) \in \mathbb{R}^k$, and let $\mathbf{c}_\mathbf{x} := \mathbb{X}\mathbf{c} \in \mathbb{R}^n$ as in Chapter 2. Consider the null hypothesis that $\mathbb{E}\mathbf{Y} \perp \mathbf{c}_\mathbf{x}$. To better interpret this hypothesis, we'll rewrite the inner product between $\mathbb{E}\mathbf{Y}$ and $\mathbf{c}_\mathbf{x}$ by summing over the groups:

$$\begin{aligned}\langle \mathbf{c}_\mathbf{x}, \mathbb{E}\mathbf{Y} \rangle &= \sum_{i=1}^{n} c_{x_i}\mu_{x_i} \\ &= \sum_{j=1}^{k} (n/k) c_j \mu_j.\end{aligned}$$

So the null hypothesis $\mathbb{E}\mathbf{Y} \perp \mathbf{c}_\mathbf{x}$ is equivalent to $\sum_{j=1}^{k} c_j\mu_j = 0$. As a specific example, suppose there are three groups, and consider the two orthogonal contrasts $\mathbf{c}^{(1)} := (1, -1, 0)$ and $\mathbf{c}^{(2)} := (\frac{1}{2}, \frac{1}{2}, -1)$. The hypothesis $\mathbb{E}\mathbf{Y} \perp \mathbf{c}^{(1)}_\mathbf{x}$ is equivalent to $\mu_1 = \mu_2$, while the hypothesis $\mathbb{E}\mathbf{Y} \perp \mathbf{c}^{(2)}_\mathbf{x}$ is equivalent to $\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2 = \mu_3$.

Next, let's derive a test statistic for any such *contrast hypothesis*. The coordinate of $\mathbf{Y}$ in the direction of $\mathbf{c_x}$ is

$$\begin{aligned}\left\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \mathbf{Y} \right\rangle &= \left\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \mathbb{E}\mathbf{Y} + \boldsymbol{\epsilon} \right\rangle \\ &= \left\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \mathbb{E}\mathbf{Y} \right\rangle + \left\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \boldsymbol{\epsilon} \right\rangle\end{aligned}$$

Furthermore, because $\mathbf{c_x} \in C(\mathbb{X})$, this coordinate is independent of $\hat{\sigma}^2$. If $\mathbb{E}\mathbf{Y} \perp \mathbf{c}_\mathcal{X}$, then this coordinate simplifies to $\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \boldsymbol{\epsilon} \rangle$, the coordinate of the error vector in the direction of $\mathbf{c_x}$. Because $\boldsymbol{\epsilon}$ is a spherically symmetric Normal random vector, its coordinate in any direction is a $N(0, \sigma^2)$-distributed random variable, as explained in Section 5.2. This squared coordinate divided by $\hat{\sigma}^2$ is

$$\begin{aligned}\frac{\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \boldsymbol{\epsilon} \rangle^2}{\hat{\sigma}^2} &= \frac{\langle \frac{\mathbf{c_x}}{\|\mathbf{c_x}\|}, \frac{\boldsymbol{\epsilon}}{\sigma} \rangle^2 / 1}{\frac{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2}{\sigma^2} / (n-k)} \\ &\sim f_{1,n-k}.\end{aligned}$$

Using a simplified expression for the squared contrast coefficient (see Section 2.3), we have the following test statistic for the null hypothesis that $\mathbb{E}\mathbf{Y} \perp \mathbf{c_x}$, i.e. that $\sum_{j=1}^k c_j \mu_j = 0$:

$$\frac{(n/k)[\sum_{j=1}^k c_j \bar{Y}_j]^2 / \sum_{j=1}^k c_j^2}{\hat{\sigma}^2} \sim f_{1,n-k}.$$

With $k-1$ orthogonal contrasts, we've seen that the squared contrast coefficients add up to the regression sum of squares. Thus the numerator of the ANOVA test statistic is the average of the numerators of the contrast test statistics. Additionally, because they involve projections of $\boldsymbol{\epsilon}$ onto orthogonal subspaces, the numerators of the contrast test statistics are independent of each other. The contrast test statistics aren't entirely independent because of their shared dependence on $\hat{\sigma}^2$, but they can be considered nearly independent especially if the sample size is large. As a result, statisticians commonly use Sidak's correction when testing *planned* orthogonal contrasts.[6]

Orthogonal contrast testing can be difficult to understand in the abstract; it might make more sense after you've worked through an example in Homework 10.

6.5. Power

The *power* of a test for a particular alternative hypothesis is the probability that the test will reject the null hypothesis when that alternative is true. As an example, let's consider the multiple linear model's f-test for the null hypothesis that all of the explanatory variables' coefficients are zero. An equivalent statement of the null hypothesis is that $\mathbb{E}\mathbf{Y} \in \text{span}\{\mathbf{1}\}$, which puts us in the context of Section 6.3. With $\mathbb{J}$ representing the orthogonal projection matrix onto the span of $\mathbf{1}$, Equation 6.1 in this case becomes

$$\begin{aligned}\hat{\mathbf{Y}} - \overline{\mathbf{Y}} &= (\mathbb{H} - \mathbb{J})\mathbb{E}\mathbf{Y} + (\mathbb{H} - \mathbb{J})\boldsymbol{\epsilon} \\ &= (\mathbb{H} - \mathbb{J})(\beta_0 \mathbf{1} + \beta_1 \mathbf{x}^{(1)} + \ldots + \beta_m \mathbf{x}^{(m)}) + (\mathbb{H} - \mathbb{J})\boldsymbol{\epsilon} \\ &= \beta_1(\mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}) + \ldots + \beta_m(\mathbf{x}^{(m)} - \bar{\mathbf{x}}^{(m)}) + (\mathbb{H} - \mathbb{J})\boldsymbol{\epsilon}.\end{aligned}$$

[6] If the contrasts were designed after the response values were seen (called *post hoc* testing), then Sidak's correction doesn't produce the specified overall false positive rate. In that case, sequential testing (after the ANOVA test) is a reasonable alternative approach.

The first part, $\beta_1(\mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}) + \ldots + \beta_m(\mathbf{x}^{(m)} - \bar{\mathbf{x}}^{(m)})$ is a non-random translation within the orthogonal complement of span$\mathbf{1}$ in $C(\mathbb{X})$, while the second part, $(\mathbb{H}-\mathbb{J})\boldsymbol{\epsilon}$ is a spherical Normal in that same (rank$\mathbb{X}-1$)-dimensional subspace. From our discussion of non-central χ^2 distributions in Section 5.3.1, we can infer that

$$\frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2}{\sigma^2} = \left\| \frac{\beta_1(\mathbf{x}^{(1)} - \bar{\mathbf{x}}^{(1)}) + \ldots + \beta_m(\mathbf{x}^{(m)} - \bar{\mathbf{x}}^{(m)})}{\sigma} + (\mathbb{H} - \mathbb{J})\frac{\boldsymbol{\epsilon}}{\sigma} \right\|^2$$
$$\sim \chi^2_{\text{rank}\mathbb{X}-1, \frac{1}{\sigma^2}\|\beta_1(\mathbf{x}^{(1)}-\bar{\mathbf{x}}^{(1)})+\ldots+\beta_m(\mathbf{x}^{(m)}-\bar{\mathbf{x}}^{(m)})\|^2}.$$

Recalling the definition of non-central f-distributions in Section 5.3.3, the f-statistic is

$$\frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2/(\text{rank}\mathbb{X} - 1)}{\hat{\sigma}^2} \sim f_{\text{rank}\mathbb{X}-1, n-\text{rank}\mathbb{X}, \frac{1}{\sigma^2}\|\beta_1(\mathbf{x}^{(1)}-\bar{\mathbf{x}}^{(1)})+\ldots+\beta_m(\mathbf{x}^{(m)}-\bar{\mathbf{x}}^{(m)})\|^2}$$

If we were testing with false positive rate .05, we would reject the null hypothesis precisely when the f-statistic is greater than $f^*_{.95}$. Therefore, the power of the test for an alternative specification of $\beta_1, \ldots, \beta_m, \sigma^2$ is $1 - G(f^*_{.95})$ where G is the cdf of[7]

$$f_{\text{rank}\mathbb{X}-1, n-\text{rank}\mathbb{X}, \frac{1}{\sigma^2}\|\beta_1(\mathbf{x}^{(1)}-\bar{\mathbf{x}}^{(1)})+\ldots+\beta_m(\mathbf{x}^{(m)}-\bar{\mathbf{x}}^{(m)})\|^2}.$$

Homework 10: ANOVA

In *anova.Rmd*, you'll build on the basics and practice a few more advanced inference tasks including analysis of variance and contrast tests.

This completes our coverage of the core theory of linear models. The focus was on developing the reader's ability to visualize data in two important ways: as *observations* and as *variables*. The *observations* picture is more natural and intuitive, while understanding the *variables* picture involves something of a mental breakthrough. Both approaches are tremendously valuable in understanding linear model theory.

Remarkably, random variables can also be profitably understood with two pictures that are perfectly analogous to those we've been studying. A random variable defined on a probability space has a distribution on $\mathbb{R}$ (the observation picture), but it can also be thought of as a vector in the space of all possible random variables on that probability space (the variable picture). Understanding this can revolutionize the way you think about probability theory. Additionally, it generalizes many of the results that we've developed in this book; these results are easily seen as special cases. So if you're bold enough to pursue the next mental breakthrough, continue your journey with *Visualizing Random Vectors* (currently still in preparation, as of this book's printing).

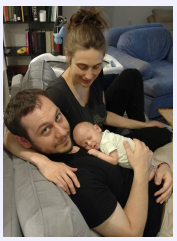
[7] While σ^2 doesn't generally need to be specified by a null hypothesis, it does need to be specified by an alternative hypothesis for the power calculation.

INDEX

About the book

Visualizing Linear Models develops the reader's understanding of the core aspects of least-squares regression and linear model theory by emphasizing two invaluable and complementary ways of visualizing the data and model: the *observations* picture and the *variables* picture. This intuitive and visual approach to the material makes it more accessible to students who aren't used to formal mathematics.

About the author



W. D. Brinda is a lecturer and researcher in the Department of Statistics and Data Science at Yale University where he also completed his doctorate. He lives in New Haven with his wife Sonya and their son Theodore.