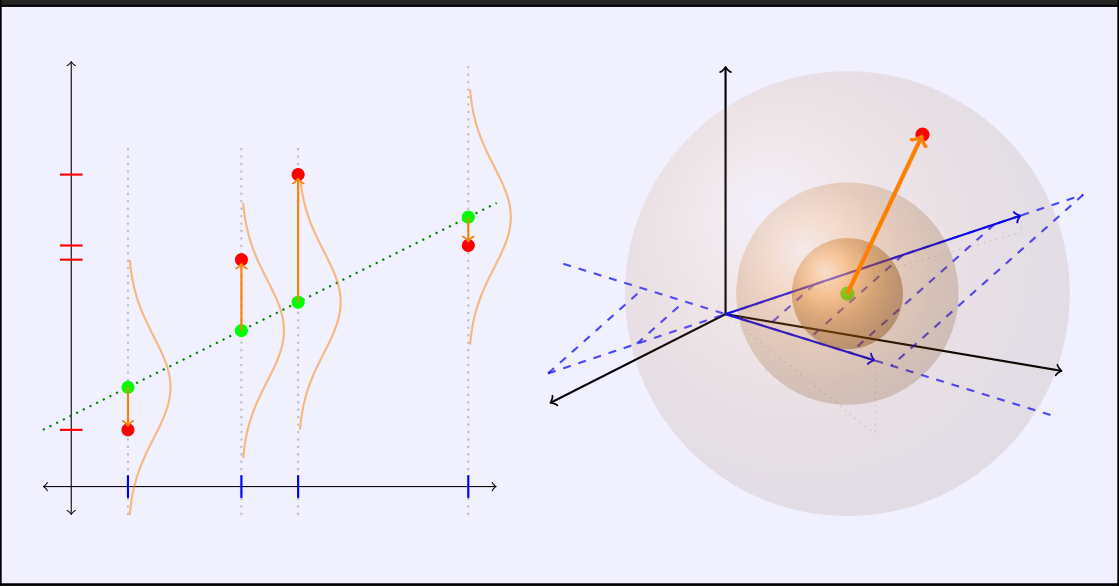
Visualizing Linear Models



W. D. Brinda, PhD

VISUALIZING LINEAR MODELS

W. D. BRINDA

Contents

Preface v

Acknowledgments vii

1 Least-squares Regression 1
- 1.1 Visualizing the observations 2
- 1.2 Visualizing the variables 8

2 The Linear Model 33
- 2.1 Visualizing the observations 35
- 2.2 Visualizing the variables 41

3 Normal Errors 57
- 3.1 Spherical symmetry 59
- 3.2 Inference 62
- 3.3 Categorical explanatory variables 69
- 3.4 Prediction 74

PREFACE

This book covers the main ideas for the theory portion of 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 exercises multiple times if needed. Eventually you'll achieve a mental breakthrough and the pictures will

all make sense.

For the serious student of probability or mathematical statistics, an analogous two-pictures approach to understanding random vectors is explained in exquisite detail in my follow-up book *Visualizing Random Vectors*.

ACKNOWLEDGMENTS

This book wouldn't have been possible without the countless teachers, friends, and family members who have helped along the way. There are too many people to list, so I'll only mention a few. Most importantly, my wife Sonya has supported me in every possible way and without any hesitation. I'm also grateful to my students for their helpful feedback and for spotting all my errors. Finally, I owe a great deal to Professor Joseph Chang who taught me this material so long ago and has continually encouraged my development as a teacher and writer over the years.

CHAPTER

1

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 and/or their sizes.

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 **3.3**. 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 the data*.

1.1. 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.

1.1.1. Least-squares point

As we learn about linear regression and modeling, one 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.

1.1 Assume that you only have $y_1, \ldots, y_n$ without any explanatory variables. Use calculus (or completing the square) to find the fitted value for the response variable that minimizes the sum of squared residuals. (We might call this the *least-squares*

point.)

1.2 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?

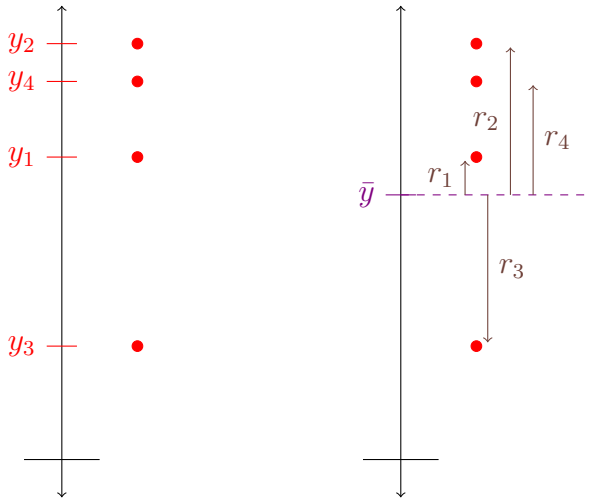


Figure 1.1: Left: Generic values of a single quantitative variable. Right: Arrows represent the residuals produced when the constant $\bar{y}$ is used to fit the variable.

1.1.2. 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 1.2.

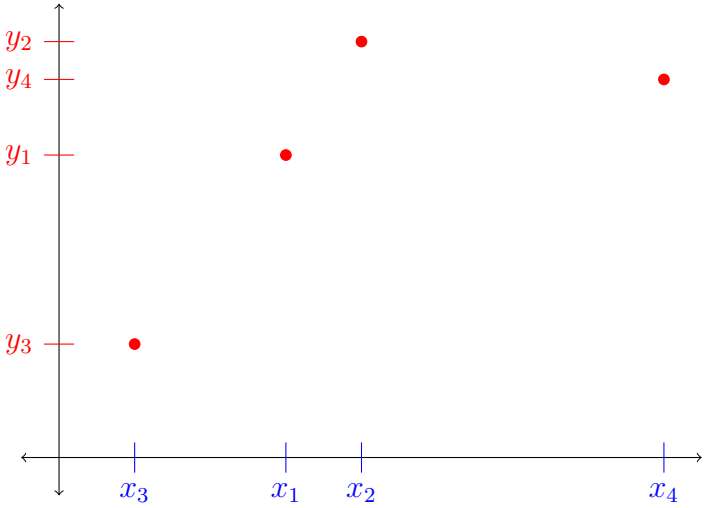


Figure 1.2: The response variable values are the same as in Figure 1.1, but now we also have an explanatory variable value for each observation which allows us to draw a scatterplot. In a scatterplot, every observation in the data is represented by a point.

The scatterplot could indicate any number of types of relationships between the data variables. Here, we're interested in summarizing the *affine*[1] relationship. 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

[1]Some readers may prefer the more familiar term *linear* here. But in serious mathematics, two variables are considered to be linearly related if they are proportional to each other. They are *affinely* related if they are related by a line. To clarify, the difference is that an affine relationship can have a non-zero intercept.

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 1.3.

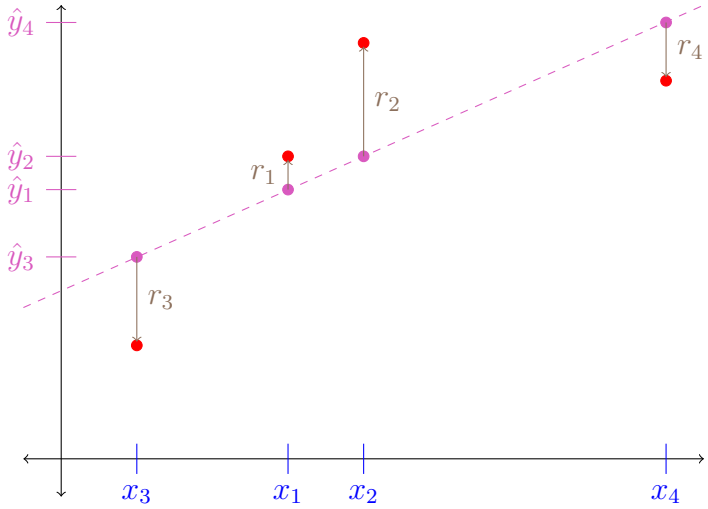


Figure 1.3: Arrows represent the residuals produced when the dotted line is used to fit the variable. In fact, the particular line shown is the least-squares line for this data, that is, the line that produces the smallest sum of squared residuals.

1.3 Suppose $(x_1, x_2) = (3, 4)$ and $(y_1, y_2) = (8.2, 4.6)$. Identify the least-squares line, and explain your answer. Sketch a scatterplot, and draw the least-squares line.

1.4 Suppose $(x_1, x_2, x_3, x_4) = (3, 3, 3, 3)$ and $(y_1, y_2, y_3, y_4) = (8.2, 4.6, 4.4, 5.7)$. Identify the least-squares line(s), and explain your answer.

Sketch a scatterplot, and draw a least-squares line.

1.5 Suppose $(x_1, x_2, x_3) = (3, 4, 3)$ and $(y_1, y_2, y_3) = (8.2, 4.6, 4.4)$. Identify the least-squares line, and explain your answer. Sketch a scatterplot, and draw the least-squares line.

1.6 Let the real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ be the explanatory and response values of n observations. Use calculus (or completing the square) to find the *line* that minimizes the sum of squared residuals. (Hint: first, find the critical intercept in terms of the slope.)

1.7 Suppose $(x_1, x_2, x_3, x_4) = (3, 4, 5, 6)$ and $(y_1, y_2, y_3, y_4) = (8.2, 4.6, 4.4, 5.7)$. Calculate the least-squares line. Sketch a scatterplot, and draw the least-squares line and the residuals.

1.1.3. 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 1.4.

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 1.5.

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 1.2.3.

1.8 Suppose

$$\left(x_1^{(1)}, x_2^{(1)}, x_3^{(1)}, x_4^{(1)}\right) = (0, 0, 1, 0),$$
$$\left(x_1^{(2)}, x_2^{(2)}, x_3^{(2)}, x_4^{(2)}\right) = (0, 0, 0, 1), \text{ and}$$
$$(y_1, y_2, y_3, y_4) = (-1, 1, -2, 3).$$

Sketch a 3D scatterplot of the data. Figure out

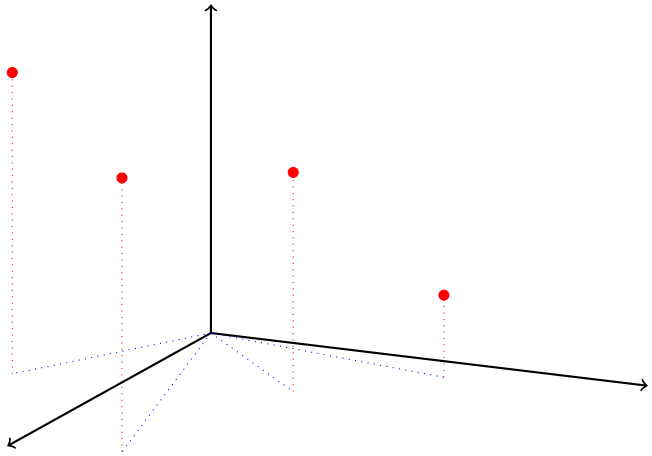


Figure 1.4: The response variable and the first explanatory variable are the same as in the previous plots, but now we also have a second explanatory variable value for each observation which allows us to draw a 3D scatterplot. The explanatory variables correspond to the horizontal plane while the response variable corresponds to the height that the point is placed above that plane.

the least-squares plane by inspection, and explain your reasoning.

1.2. 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.

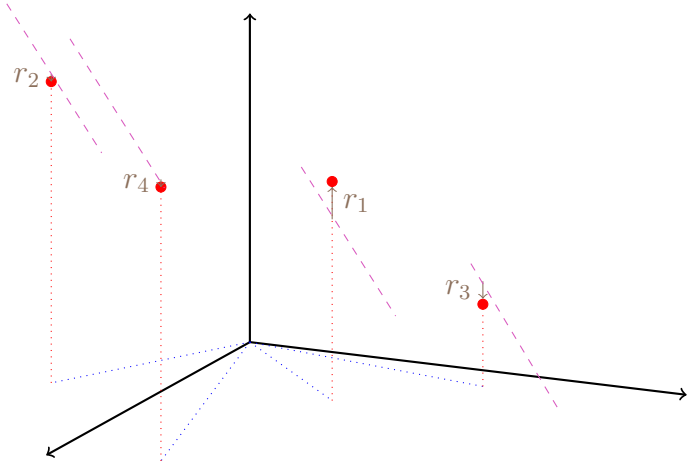


Figure 1.5: Arrows represent the residuals produced when the least-squares plane is used to fit the variable. Level curves of the least-squares plane are drawn at the locations of the fitted values.

1.2.1. 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 fit $y_1, \ldots, y_n$, 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.[2] 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 fit $y_1, \ldots, y_n$ is equivalent to selecting an a for which $a\mathbf{1}$ fits

[2] Although vectors are written out horizontally in this book's paragraph text, they should always be interpreted as column vectors in any mathematical expression.

$\mathbf{y}$. The span of $\mathbf{1}$, $\{a\mathbf{1} : a \in \mathbb{R}\}$, can be called *the constant subspace* because each vector in span$\{\mathbf{1}\}$ has a value that is constant across all of its components.

The residuals are also neatly represented by an n-dimensional Euclidean vector $\mathbf{r} := \mathbf{y} - a\mathbf{1} = (y_1 - a, \ldots, y_n - a)$.

1.9 Write the Euclidean distance between $\mathbf{y}$ and the vector of fitted values in terms of the residuals. How is minimizing this Euclidean distance related to minimizing the sum of squared residuals?

Recall that, given any vector $\mathbf{v} \in \mathbb{R}^n$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^n$, the vector in $\mathcal{S}$ that is closest (in Euclidean distance) to $\mathbf{v}$ is the *orthogonal projection* of $\mathbf{v}$ onto $\mathcal{S}$. The orthogonal projection's defining characteristic is that it is the unique $\mathbf{v}_\mathcal{S} \in \mathcal{S}$ such that $\mathbf{v} - \mathbf{v}_\mathcal{S}$ is orthogonal to every vector in $\mathcal{S}$.

1.10 Find the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$ by using the defining characteristic of orthogonal projections. In other words, your task is to find the value of the coefficient a for which $a\mathbf{1}$ is as close as possible to $\mathbf{y}$. We will call this vector $\bar{\mathbf{y}}$ or $\bar{y}\mathbf{1}$. Compare your answer to Exercise **1.1**, and interpret this based on the observation from Exercise **1.9**. (In this way of viewing things, one may prefer the term *least-squares constant* rather than *least-squares point*.)

1.11 Let $\mathbf{y} = (-2, 3, 11)$. Sketch a picture in $\mathbb{R}^3$ that includes $\mathbf{y}$, $\mathbf{1}$, $\bar{\mathbf{y}}$, and the least-squares residual

vector $\mathbf{y} - \bar{\mathbf{y}}$. Now consider a different constant vector $a\mathbf{1}$ for a generic a; draw a point on the span of $\mathbf{1}$ somewhere beyond $\bar{\mathbf{y}}$, and label the mark "$a\mathbf{1}$." In general, consider the triangle with corners at $\mathbf{y}$, $a\mathbf{1}$, and $\bar{\mathbf{y}}$. Use the Pythagorean theorem to relate the lengths of the sides of this triangle.

In Exercise **1.11**, it was easy 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.[3] Often all of the vectors of interest lie within a two- or three-dimensional slice of $\mathbb{R}^n$ and can therefore be accurately depicted in a sketch; see Figures 1.6 and 1.7.

1.2.2. 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 $a\mathbf{1} + b\mathbf{x}$ for $a, b \in \mathbb{R}$. That means that the set of possible fitted values is exactly the span of $\{\mathbf{1}, \mathbf{x}\}$, which forms a two-dimensional subspace[4] of $\mathbb{R}^n$.

[3] 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.

[4] This is true unless $\mathbf{x}$ is constant, in which case $\text{span}\{\mathbf{1}, \mathbf{x}\}$ is one-dimensional. In that case, the explanatory variable vector doesn't

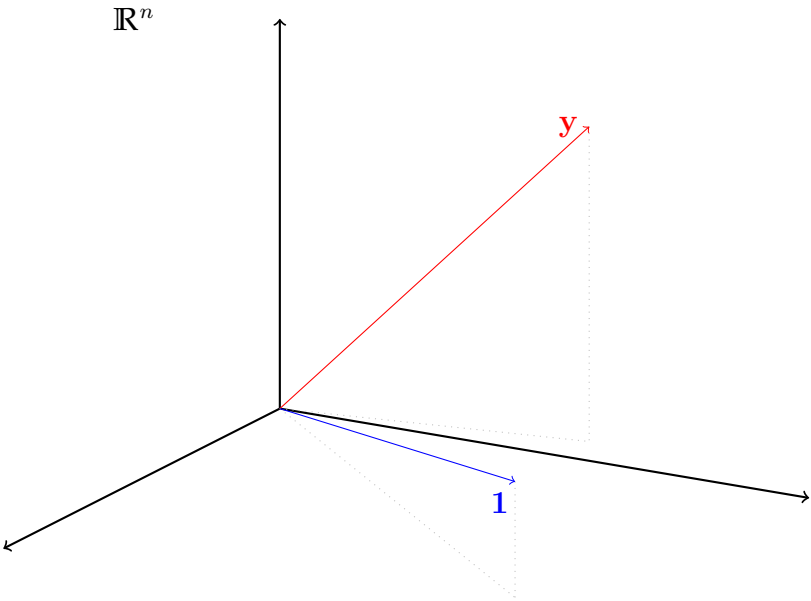


Figure 1.6: A generic picture of the constant vector $\mathbf{1}$ *and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. We know that there exist infinitely many three-dimensional subspaces that includes the origin along with these two vectors' endpoints, so we can assume that the view shown here is such a three-dimensional slice 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 $\mathbf{0}$, $\mathbf{1}$, $\mathbf{x}$ and $\mathbf{y}$; see Figures 1.8 and 1.9. All of the familiar rules of Euclidean geometry apply, of course, in this three-dimensional subspace.

add anything to the set of possible fits.

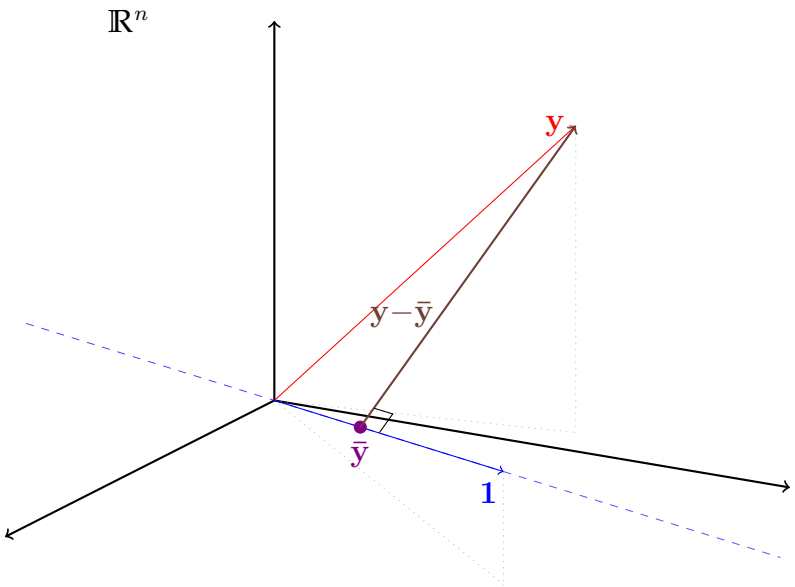


Figure 1.7: A generic picture of the constant vector $\mathbf{1}$ *and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dotted line follows a portion of the span of* $\mathbf{1}$*.* $\bar{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto that subspace, and* $\mathbf{y}-\bar{\mathbf{y}}$ *is orthogonal to it.*

1.12 Find the orthogonal projection of $\mathbf{y}$ onto the span of $\{\mathbf{1}, \mathbf{x}\}$ by using the defining characteristic of orthogonal projections. In other words, your task is to find the values of a and b for which $a\mathbf{1}+b\mathbf{x}$ is as close as possible to $\mathbf{y}$. (Hint: It suffices to find the pair of coefficients (a, b) for which the residual vector $\mathbf{y}-(a\mathbf{1}+b\mathbf{x})$ is orthogonal to both $\mathbf{1}$ and $\mathbf{x}$). We will call this vector $\hat{\mathbf{y}}$. Compare your answer to Exercise **1.6**, and interpret this based on the observation from Exercise **1.9**.

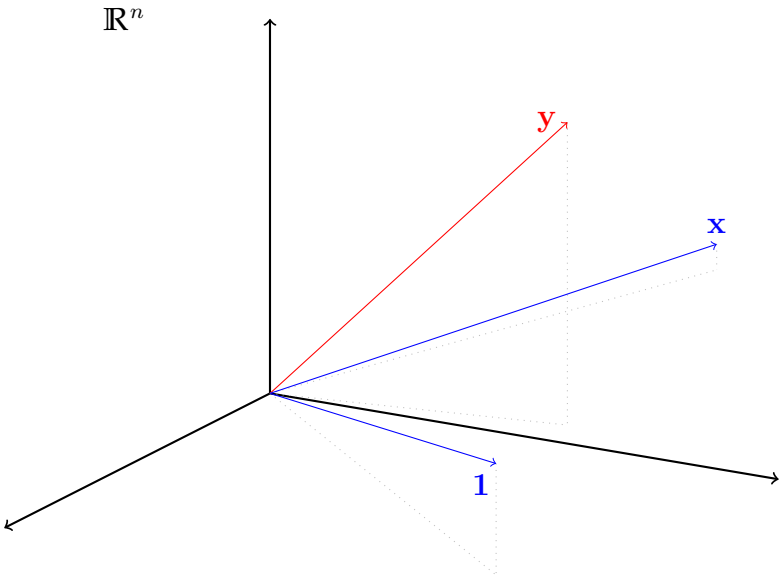


Figure 1.8: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. We know that there exists a three-dimensional subspace that includes the origin along with these three vectors' endpoints, so we can assume that the view shown here is such a three-dimensional slice of* $\mathbb{R}^n$*.*

1.13 Let $\mathbf{y} = (-2, 3, 11)$ and $\mathbf{x} = (-4, 3, 3)$. Sketch a picture in $\mathbb{R}^3$ that includes $\mathbf{y}$, $\mathbf{1}$, $\mathbf{x}$, $\text{span}\{\mathbf{1}, \mathbf{x}\}$, $\bar{\mathbf{y}}$, $\hat{\mathbf{y}}$, and the least-squares residual vectors $\mathbf{y} - \hat{\mathbf{y}}$ and $\mathbf{y} - \bar{\mathbf{y}}$. In general, consider the triangle with corners at $\mathbf{y}$, $\hat{\mathbf{y}}$, and $\bar{\mathbf{y}}$. Use the Pythagorean theorem to relate the lengths of the sides of this triangle.

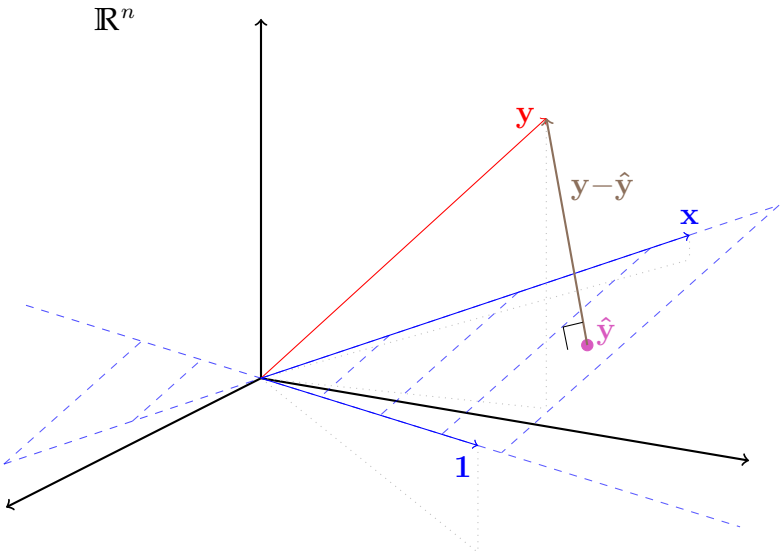


Figure 1.9: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and a response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dashed lines outline a portion of the span of* $\mathbf{1}$ *and* $\mathbf{x}$*.* $\hat{\mathbf{y}}$ *is the orthogonal projection of* $\mathbf{y}$ *onto that subspace, and* $\mathbf{y} - \hat{\mathbf{y}}$ *is orthogonal to it.*

1.14 Is it possible for the least-squares line's sum of squared residuals to be strictly greater than the least-squares point's sum of squared residuals? Is it possible for them to be equal? First, answer these questions with the observations-picture in mind, then answer them based on the variables-picture. Observe that the Pythagorean theorem result in Exercise **1.13** leads to a more specific answer.

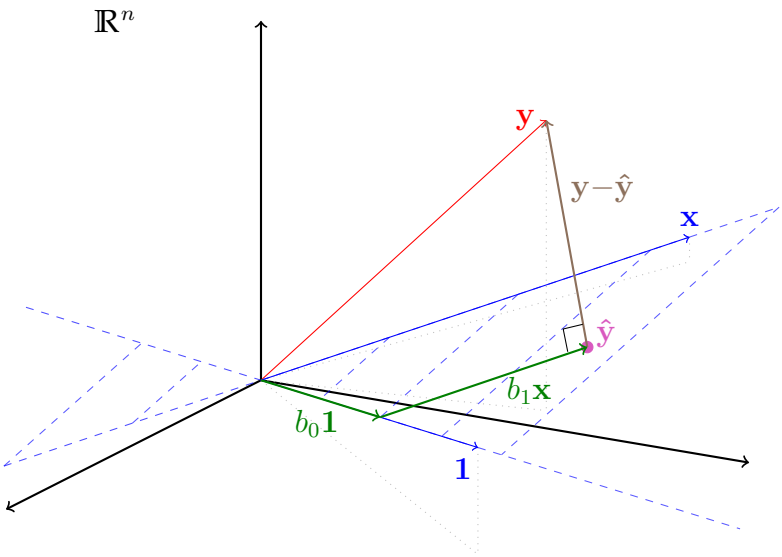


Figure 1.10: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, a response variable vector* $\mathbf{y}$*, and its projection onto the span of* $\mathbf{1}$ *and* $\mathbf{x}$*. The least-squares coefficients* b_0 *and* b_1 *are the unique coefficients of* $\mathbf{1}$ *and* $\mathbf{x}$ *for which* $b_0\mathbf{1} + b_1\mathbf{x} = \hat{\mathbf{y}}$*.*

1.15 Let $\bar{\mathbf{x}}$ be the orthogonal projection of $\mathbf{x}$ onto the span of $\mathbf{1}$. (Recall specifically how this works out in Exercise **1.10**.) Using the result of Exercise **1.12**, express $\hat{\mathbf{y}}$ as a linear combination of $\mathbf{1}$ and $\mathbf{x} - \bar{\mathbf{x}}$. Use this expression to figure out the orthogonal projection of $\hat{\mathbf{y}}$ onto the span of $\mathbf{1}$.

Moving forward, it will be useful to understand that various familiar statistics can be interpreted as Euclidean space quantities. For any vector $\mathbf{x}$, we'll use the term *cen-*

tered vector to refer to $\mathbf{x} - \bar{\mathbf{x}}$ the vector of deviations from the mean.

The statistic $\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})$ is sometimes called the *covariance* between the constant vectors $\mathbf{x}$ and $\mathbf{y}$.[5] More accurately, that statistic is an unbiased estimate of the covariance between the random variables X and Y, assuming $(x_1, y_1), \ldots, (x_n, y_n)$ were iid draws from the distribution of (X, Y). Alternatively, if one takes the empirical distribution of the data to be the distribution of interest, then the covariance is $\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})$ which differs from the *covariance estimate* by having n rather than $n-1$ in the denominator.

Throughout the remainder of this section, let the distribution of (X, Y) be uniform on $(x_1, y_1), \ldots (x_n, y_n)$, which is the empirical joint distribution defined by the data vectors $\mathbf{x}$ and $\mathbf{y}$. Then the expectation of X is $1/n$ times the dot product between $\mathbf{1}$ and $\mathbf{x}$:

$$\begin{aligned}\mathbb{E}X &:= \frac{1}{n}\sum x_i \\ &= \frac{1}{n}\mathbf{1}'\mathbf{x} \\ &= \bar{x}.\end{aligned}$$

The covariance between X and Y is $1/n$ times the dot product between the centered vectors:

$$\begin{aligned}\operatorname{cov}(X, Y) &= \tfrac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) \\ &= \tfrac{1}{n}(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{y} - \bar{\mathbf{y}}).\end{aligned}$$

[5] We will be careful to reserve *covariance* to mean either the covariance of a pair of random variables or the covariance matrix of a random vector.

The *variance* of X is $1/n$ times the squared length of the centered vector:

$$\begin{aligned}\mathrm{var}(X) &:= \mathrm{cov}(X, X) \\ &= \tfrac{1}{n}(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{x} - \bar{\mathbf{x}}) \\ &= \tfrac{1}{n}\|\mathbf{x} - \bar{\mathbf{x}}\|^2.\end{aligned}$$

The *standard deviation* is the square root of the variance, so it's $1/\sqrt{n}$ times the length of the centered vector:

$$\begin{aligned}\mathrm{sd}(X) &:= \sqrt{\mathrm{var}(X)} \\ &= \tfrac{1}{\sqrt{n}}\|\mathbf{x} - \bar{\mathbf{x}}\|.\end{aligned}$$

The *correlation* between X and Y is the the dot product of the centered vectors divided by the product of their lengths:

$$\begin{aligned}\mathrm{cor}(X, Y) &:= \frac{\mathrm{cov}(X, Y)}{\mathrm{sd}(X)\mathrm{sd}(Y)} \\ &= \frac{\frac{1}{n}(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{y} - \bar{\mathbf{y}})}{\frac{1}{\sqrt{n}}\|\mathbf{x} - \bar{\mathbf{x}}\|\frac{1}{\sqrt{n}}\|\mathbf{y} - \bar{\mathbf{y}}\|} \\ &= \frac{(\mathbf{x} - \bar{\mathbf{x}})'(\mathbf{y} - \bar{\mathbf{y}})}{\|\mathbf{x} - \bar{\mathbf{x}}\|\|\mathbf{y} - \bar{\mathbf{y}}\|}.\end{aligned}$$

This statistic can also be called the *sample correlation* between $\mathbf{x}$ and $\mathbf{y}$ and defined in terms of the covariance estimate and standard deviation estimates (all involving $n - 1$ rather than n) because the constant cancels out either way.[6]

[6]By the geometric definition of dot products, this correlation is exactly equal to the cosine of the angle between the centered vectors. Yet another equivalent expression for the correlation is the dot product between the normalized centered vectors $(\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|})'(\frac{\mathbf{y}-\bar{\mathbf{y}}}{\|\mathbf{y}-\bar{\mathbf{y}}\|})$.

Note that the dot product of $\mathbf{x}-\bar{\mathbf{x}}$ with $\mathbf{y}-\bar{\mathbf{y}}$ is exactly the same as the dot product of $\mathbf{x}$ with $\mathbf{y}-\bar{\mathbf{y}}$. Because $\mathbf{y}-\bar{\mathbf{y}}$ is orthogonal to $\mathbf{1}$, it "kills off" the portion of $\mathbf{x}$ that is in the direction of $\mathbf{1}$.

$$\begin{aligned}\mathbf{x}'(\mathbf{y}-\bar{\mathbf{y}}) &= (\bar{\mathbf{x}}+[\mathbf{x}-\bar{\mathbf{x}}])'(\mathbf{y}-\bar{\mathbf{y}}) \\ &= \underbrace{\bar{\mathbf{x}}'(\mathbf{y}-\bar{\mathbf{y}})}_{\mathbf{0}}+(\mathbf{x}-\bar{\mathbf{x}})'(\mathbf{y}-\bar{\mathbf{y}})\end{aligned}$$

So when expressing covariance or correlation, only one of the vectors needs to be centered.

Now let's review some additional general aspects of orthogonal projection before working out a formula for the least-squares line. Consider a subspace $\mathcal{S} \subseteq \mathbb{R}^n$. We know that any vector $\mathbf{v}$ can be uniquely represented as the sum of a vector $\mathbf{v}_{\mathcal{S}}$ in $\mathcal{S}$ and a vector $\mathbf{v}_{\mathcal{S}^\perp}$ in its orthogonal complement. Orthogonal projection is a linear operator and can thus be represented as a matrix multiplication. Let $\mathbb{M}_{\mathcal{S}}$ be the matrix that performs orthogonal projection onto $\mathcal{S}$. Notice that the orthogonal projection of $\mathbf{v}$ onto $\mathcal{S}$ is exactly $\mathbf{v}_{\mathcal{S}}$ from the unique decomposition.

$$\begin{aligned}\mathbb{M}_{\mathcal{S}}\mathbf{v} &= \mathbb{M}_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}+\mathbf{v}_{\mathcal{S}^\perp}) \\ &= \underbrace{\mathbb{M}_{\mathcal{S}}\mathbf{v}_{\mathcal{S}}}_{\mathbf{v}_{\mathcal{S}}}+\underbrace{\mathbb{M}_{\mathcal{S}}\mathbf{v}_{\mathcal{S}^\perp}}_{\mathbf{0}} \\ &= \mathbf{v}_{\mathcal{S}}\end{aligned}$$

The orthogonal projection leaves $\mathbf{v}_{\mathcal{S}}$ alone because it is in $\mathcal{S}$, and it eliminates $\mathbf{v}_{\mathcal{S}^\perp}$ because it is orthogonal to $\mathcal{S}$.[7]

[7]This property of orthogonal projection matrices is discussed in Section 2.2.3.

One way to do orthogonal projection onto $\mathcal{S}$ is to use smaller orthogonal subspaces. Let $\mathcal{S}_1, \ldots, \mathcal{S}_k$ be subspaces of $\mathcal{S}$ that are orthogonal to each other and that together span $\mathcal{S}$. Any vector $\mathbf{v}$ can be uniquely represented as the sum of vectors $\mathbf{v}_1 \in \mathcal{S}_1, \ldots, \mathbf{v}_k \in \mathcal{S}_k$ plus a vector $\mathbf{v}_{\mathcal{S}^\perp}$ orthogonal to $\mathcal{S}$. Let $\mathbb{M}_j$ be the orthogonal projection onto $\mathcal{S}_j$.

$$\begin{aligned}\mathbb{M}_j\mathbf{v} &= \underbrace{\mathbb{M}_j\mathbf{v}_1}_{\mathbf{0}} + \ldots + \underbrace{\mathbb{M}_j\mathbf{v}_j}_{\mathbf{v}_j} + \ldots + \underbrace{\mathbb{M}_j\mathbf{v}_k}_{\mathbf{0}} + \underbrace{\mathbb{M}_j\mathbf{v}_{\mathcal{S}^\perp}}_{\mathbf{0}} \\ &= \mathbf{v}_j\end{aligned}$$

This allows us to see that orthogonal projection of $\mathbf{v}$ onto $\mathcal{S}$ equals the sum of the orthogonal projections onto $\mathcal{S}_1, \ldots, \mathcal{S}_k$ which we can use to make the closely related observation that the orthogonal projection matrix onto $\mathcal{S}$ is equal to the sum of the orthogonal projection matrices onto $\mathcal{S}_1, \ldots, \mathcal{S}_k$.

$$\begin{aligned}(\mathbb{M}_1 + \ldots + \mathbb{M}_k)\mathbf{v} &= \mathbb{M}_1\mathbf{v} + \ldots + \mathbb{M}_k\mathbf{v} \\ &= \mathbf{v}_1 + \ldots + \mathbf{v}_k \\ &= \mathbf{v}_{\mathcal{S}}\end{aligned}$$

Orthogonal projection onto the span of a unit vector is particularly convenient. Let $\mathbf{u}$ be a unit vector, and let $y_{\mathbf{u}}\mathbf{u}$ denote the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{u}$. Let's find the coefficient $y_{\mathbf{u}}$ that satisfies the equation defining the orthogonal projection.

$$\begin{aligned}&\mathbf{u}'(\mathbf{y} - y_{\mathbf{u}}\mathbf{u}) = 0 \\ &\Rightarrow \underbrace{\mathbf{u}'\mathbf{u}}_{1}\, y_{\mathbf{u}} = \mathbf{u}'\mathbf{y} \\ &\Rightarrow y_{\mathbf{u}} = \mathbf{u}'\mathbf{y}\end{aligned} \tag{1.1}$$

The orthogonal projection of any vector $\mathbf{y}$ onto the span of a unit vector $\mathbf{u}$ is simply $(\mathbf{y}'\mathbf{u})\mathbf{u}$.

Let's use this approach to find the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$ and $\mathbf{x}$, assuming they are linearly independent. First, we need an orthonormal basis for the subspace of interest. We'll start with the vector $\frac{\mathbf{1}}{\|\mathbf{1}\|}$ which is $\frac{\mathbf{1}}{\sqrt{n}}$. We need a second basis vector for $\text{span}\{\mathbf{1}, \mathbf{x}\}$ that is orthogonal to $\mathbf{1}$. Consider $\mathbf{x}$ minus it's orthogonal projection onto $\mathbf{1}$; indeed $\mathbf{x} - \bar{\mathbf{x}}$ is orthogonal to $\mathbf{1}$ by the defining property of orthogonal projections. So for our second basis vector, we can take the unit vector $\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}$. (If there were more vectors, we can continue obtaining orthonormalized versions by subtracting their projection onto the span of the previous vectors then dividing the resulting vector by its length, a procedure known as the *Gram-Schmidt algorithm.*)

Per our discussion above, the orthogonal projection $\hat{\mathbf{y}}$ onto $\text{span}\{\mathbf{1}, \mathbf{x}\}$ is the sum of the orthogonal projections onto the spans[8] of $\frac{\mathbf{1}}{\sqrt{n}}$ and $\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}$, which are simple by (1.1).

$$\begin{aligned}\hat{\mathbf{y}} &= \left(\mathbf{y}'\frac{\mathbf{1}}{\sqrt{n}}\right)\frac{\mathbf{1}}{\sqrt{n}} + \left(\mathbf{y}'\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\right)\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|} \\ &= \bar{\mathbf{y}} + \frac{\mathbf{y}'(\mathbf{x}-\bar{\mathbf{x}})}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\frac{\mathbf{x}-\bar{\mathbf{x}}}{\|\mathbf{x}-\bar{\mathbf{x}}\|}\end{aligned}$$

If we subtract $\bar{\mathbf{y}}$ then divide by the length of $\mathbf{y} - \bar{\mathbf{y}}$, we see a particularly enlightening form for the least-squares line

[8] Orthogonal projection onto the span of $\mathbf{1}$ is the same as orthogonal projection onto the span of the unit vector $\frac{\mathbf{1}}{\sqrt{n}}$. Check for yourself that the formula we've derived here produces the same $\bar{\mathbf{y}}$ vector that you found in Exercise **1.10**.

equation.

$$\frac{\hat{\mathbf{y}} - \bar{\mathbf{y}}}{\|\mathbf{y} - \bar{\mathbf{y}}\|} = \underbrace{\frac{\mathbf{y}'(\mathbf{x} - \bar{\mathbf{x}})}{\|\mathbf{y} - \bar{\mathbf{y}}\|\|\mathbf{x} - \bar{\mathbf{x}}\|}}_{\text{cor}(X,Y)} \frac{\mathbf{x} - \bar{\mathbf{x}}}{\|\mathbf{x} - \bar{\mathbf{x}}\|} \tag{1.2}$$

According to (1.2), the correlation is exactly the the coefficient of the least-squares line for the standardized versions of the variables; in other words, if the explanatory variable is z standard deviations above its mean, the least-squares line predicts the response variable to be $\text{cor}(X, Y)$ times z standard deviations above its mean. This phenomenon was originally called "regression toward mediocrity" by Francis Galton, an eminent British intellectual who pioneered least-squares fitting in the late nineteenth century; the term *regression* caught on and is now used broadly for fitting quantitative data.

1.2.3. 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 fitted vectors of the form

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

with $b_0, \ldots, b_m \in \mathbb{R}$. A more compact expression for these fitted vectors is $\mathbb{X}\mathbf{b}$ where[9]

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

With the vector of coefficients $\mathbf{b}$ ranging over $\mathbb{R}^{m+1}$, the set of possible fitted vectors is precisely the span of the columns $\mathbf{1}$, $\mathbf{x}^{(1)}$, ..., $\mathbf{x}^{(m)}$ (also known as the *column space* of $\mathbb{X}$), which is a subspace of $\mathbb{R}^n$. The fitted vector minimizing the sum of squared residuals is exactly the one that is closest to $\mathbf{y}$ in Euclidean distance. To generalize our earlier way of thinking from Section 1.1, if we envision the observations as data points in $\mathbb{R}^{m+1}$, each possible fit function is an m-dimensional hyperplane defined by $f_{\mathbf{b}}(\mathbf{x}) = b_0 + b_1 x^{(1)} + \ldots + b_m x^{(m)}$ for any explanatory observation $(x^{(1)}, \ldots, x^{(m)}) \in \mathbb{R}^m$. To be consistent with our earlier terminology, we might use the term *least-squares hyperplane*, but it is much more common to say simply *least-squares fit*.

As before, the least-squares fit is the orthogonal projection of $\mathbf{y}$ onto the span of $\{\mathbf{1}, \mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$. If we can find a vector $\mathbf{b}^* \in \mathbb{R}^{m+1}$ for which

$$\mathbb{X}'(\mathbf{y} - \mathbb{X}\mathbf{b}^*) = \mathbf{0}, \tag{1.3}$$

then $\mathbb{X}\mathbf{b}^*$ is the orthogonal projection of $\mathbf{y}$ onto the column span of $\mathbb{X}$. Assuming the columns of $\mathbb{X}$ are linearly

[9] $\mathbb{X}$ is called the **design matrix**.

independent,[10] there is a unique $\mathbf{b}^*$ that leads to the orthogonal projection. In that case, we can solve (1.3) for $\mathbf{b}^*$ to get

$$\begin{aligned} \mathbb{X}'\mathbb{X}\mathbf{b}^* &= \mathbb{X}'\mathbf{y} \qquad (1.4)\\ \Rightarrow \qquad \mathbf{b}^* &= (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{y}. \end{aligned}$$

(Notice that if the columns of $\mathbb{X}$ are an orthonormalized version of $\mathbf{1}$ and the explanatory variables, as described in Section 1.2.2, the formula simplifies to $\mathbb{X}'\mathbf{y}$.)

1.16 If $\mathbb{X}\mathbf{b}^*$ is the orthogonal projection of $\mathbf{y}$ onto the column span of $\mathbb{X}$, then $\mathbf{y} - \mathbb{X}\mathbf{b}^*$ should be orthogonal to *every vector* in the column space of $\mathbb{X}$. Why is it sufficient to check that $\mathbf{y}-\mathbb{X}\mathbf{b}^*$ is orthogonal to the columns of $\mathbb{X}$, as in (1.3)? (Hint: An arbitrary vector in the column space can be represented as a linear combination of the column vectors.)

Finally, the orthogonal projection of $\mathbf{y}$ onto the span of $\mathbf{1}$, $\mathbf{x}^{(1)}$, ..., $\mathbf{x}^{(m)}$, which we will again call $\hat{\mathbf{y}}$, is

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

We will continue to draw pictures to represent the variable vectors in $\mathbb{R}^n$, but at this point we need to think more

[10] If the columns of $\mathbb{X}$ are linearly dependent, then there are infinitely many coefficient vectors that lead to the orthogonal projection. We'll return to that case later in this section.

carefully about how to depict the vectors and subspaces accurately. Figures 1.11 and 1.12 are the prime examples of pictures with potentially more than one explanatory variable, and we will consider each figure in turn.

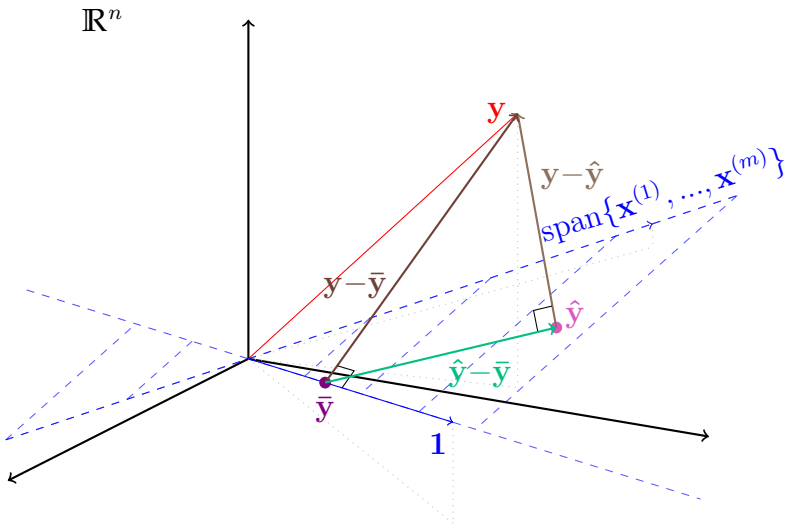


Figure 1.11: A generic picture showing the constant subspace (the span of $\mathbf{1}$*) and the response variable vector* $\mathbf{y}$ *in* $\mathbb{R}^n$*. The dashed lines represent a portion of the span of* $\mathbf{1}$ *and the explanatory variables together.* $\hat{\mathbf{y}}$ *is the orthogonal 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's first look at Figure 1.11 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

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

Let $\mathbf{v} := b_1\mathbf{x}^{(1)} + \ldots + b_m\mathbf{x}^{(m)}$, and note that it's in $\mathcal{S}$. Because the $b_0\mathbf{1}$ and $\hat{\mathbf{y}}$ are both in our three-dimensional picture, so is $\mathbf{v} = \hat{\mathbf{y}} - b_0\mathbf{1}$. 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.

Next, how accurate is Figure 1.12? 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 $\mathcal{S} := \text{span}(\mathcal{S}_0 \cup \mathcal{S}_1)$ 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 dimen-

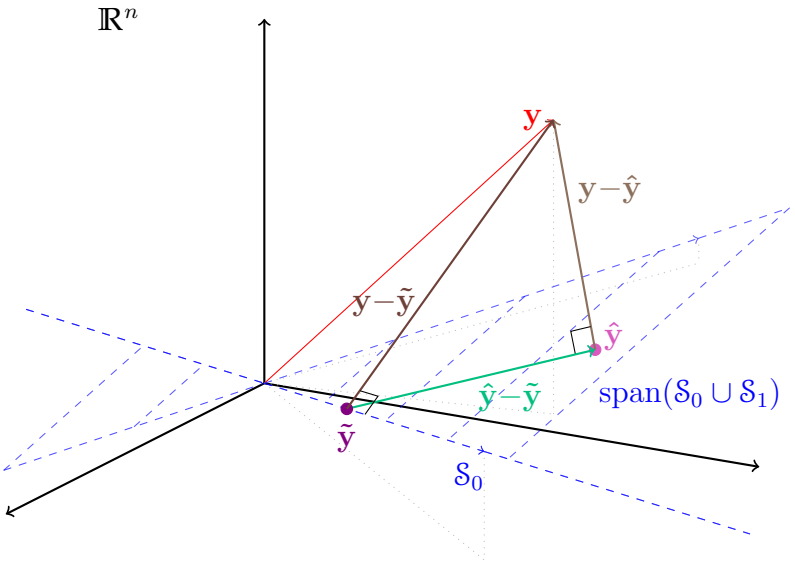


Figure 1.12: A generic picture showing a response variable vector $\mathbf{y}$ *in* $\mathbb{R}^n$*, its orthogonal projection* $\tilde{\mathbf{y}}$ *onto a subspace* $\mathcal{S}_0$*, and its orthogonal projection* $\hat{\mathbf{y}}$ *onto the span of* $\mathcal{S}_0 \cup \mathcal{S}_1$*. In general,* $\mathcal{S}_0$ *intersects this three-dimensional perspective in a line, and* $\mathcal{S}_0 \cup \mathcal{S}_1$ *intersects it in a plane. The dashed lines represent a portion of the span of* $\mathcal{S}_0 \cup \mathcal{S}_1$*.*

sion (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 $\mathbf{y} \in \mathcal{S}_1$, 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.

1.17 In the context of Figure 1.12, explain how we know that $\mathbf{y} - \hat{\mathbf{y}}$ is orthogonal to $\hat{\mathbf{y}} - \tilde{\mathbf{y}}$? Use the Pythagorean theorem to relate their squared lengths.

Let's finally think about the messier case in which the columns of $\mathbb{X}$ are linearly dependent, then there are infinitely many coefficient vectors that lead to the orthogonal projection $\hat{\mathbf{y}}$. In that case, the inverse of $\mathbb{X}'\mathbb{X}$ doesn't exist. Fortunately, mathematicians have studied a generalization of the inverse called the *Moore-Penrose inverse* that satisfies a set of defining conditions.[11] Every matrix has a unique Moore-Penrose inverse, and it has the following property (among many others): with $\mathbb{M}^-$ denoting the Moore-Penrose inverse of the matrix $\mathbb{M}$,

$$\mathbb{M}\mathbb{M}^-\mathbf{v} = \mathbf{v} \tag{1.5}$$

for every $\mathbf{v}$ in the column space of $\mathbb{M}$.

1.18 Explain how we know that $\mathbb{X}'\mathbf{y}$ is in the column space of $\mathbb{X}'\mathbb{X}$. Then use (1.5) to verify that $\mathbf{b}^* = (\mathbb{X}'\mathbb{X})^-\mathbb{X}'\mathbf{y}$ satisfies (1.4).

[11] We won't bother going through the definition of Moore-Penrose inverses here, but the interested reader is encouraged to learn about them elsewhere.

Moore-Penrose inverse matrices are readily calculated by mathematical software packages.

If an inverse matrix exists, then it is also the Moore-Penrose inverse. Thus, regardless of whether the columns of $\mathbb{X}$ are linearly independent or not, an expression for the least-squares fit is[12]

$$\hat{\mathbf{y}} = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-}\mathbb{X}'\mathbf{y}. \tag{1.6}$$

1.19 Let $\mathbb{M}$ be a real matrix. Write a general expression for a matrix that maps any vector to its orthogonal projection onto the column space of $\mathbb{M}$. This is called an *orthogonal projection matrix* for $\mathbb{M}$. (Hint: think about (1.6).) Use this formula to write out the orthogonal projection matrix onto the span of $\mathbf{1}$.

Based on the intuition you've developed by working through this chapter, hopefully it's 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.*

1.20 Let $\mathbf{y}$ be a response variable vector and $\mathbf{x}$ be an explanatory variable vector. Consider fitting the response variable by using quadratic functions of

[12] For the remainder of this book, $(\mathbb{X}'\mathbb{X})^{-}$ should be understood to denote the Moore-Penrose inverse of $\mathbb{X}'\mathbb{X}$.

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 $\mathbb{R}^n$. Explain how to find the fit that minimizes the sum of squared residuals.

Now let's make an important observation based on Figure 1.12. Because the vector $\hat{\mathbf{y}} - \tilde{\mathbf{y}}$ is in the span of $\mathcal{S}_0 \cup S_1$, it must be orthogonal to $\mathbf{y} - \hat{\mathbf{y}}$. Thus we observe a right triangle, and by the Pythagorean theorem[13]

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

The reduction in sum of squared residuals by adding in the new variables spanning $\mathcal{S}_1$ is equal to squared distance from the original fit $\tilde{\mathbf{y}}$ to the new fit $\hat{\mathbf{y}}$.

In particular, letting $\mathcal{S}_0$ be the span of $\mathbf{1}$ and $\mathcal{S}_1$ be the span of the explanatory variables (as in Figure 1.11),

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

The terms in this equation are called *total sum of squares*, *residual sum of squares*, and *regression sum of squares*. 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 $\mathbf{y} - \bar{\mathbf{y}}$

[13] In certain cases of linear dependence, one side of the triangle can have length zero, but this result still holds as stated here.

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 (1.2). 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 1.2.2, means that the R^2 is the squared correlation between $\mathbf{y}$ and $\mathbf{x}$.[14]

In the next chapter, we'll make certain modeling assumptions (assumptions about probability distributions generating the data), and see how to incorporate them into our visualizations.

[14]Galton used the letter R for the "regression coefficient" in (1.2) 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.

CHAPTER

2

THE LINEAR MODEL

A MODEL IS A MATHEMATICAL formulation that is supposed 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

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

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. We will continue to denote explanatory variables with lower-case letters, but we will now use capital letters for the random variable quantities. Results from Chapter 1 that were true for arbitrary $y_1, \ldots, y_n$ also hold for random variables $Y_1, \ldots, Y_n$. For example, the sum of squared differences from a point a is now a random variable $\sum(Y_i - a)^2$; furthermore, this sum of squares is 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 $\bar{Y} := \frac{1}{n}\sum Y_i$.

The terminology "linear model" can be misleading. It refers to the fact that the response variable and its expectation are related *linearly in the parameters*, not that they're linear 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 called a linear model, but

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

isn't called a linear model. In general, we'll use the term *linear model* for any model of the form

$$Y_i = b_0 f_0(x_i^{(1)}, \ldots, x_i^{(m)}) + \ldots + b_k f_k(x_i^{(1)}, \ldots, x_i^{(m)}) + \epsilon_i \tag{2.1}$$

with $(b_0, \ldots, b_k)$ ranging over $\mathbb{R}^{k+1}$ and with $\epsilon_1, \ldots, \epsilon_n$ each having mean zero.

2.1 Consider the linear model as defined in (2.1). Each $\mathbf{b} := (b_0, \ldots, b_k) \in \mathbb{R}^{k+1}$ can be understood to provide an estimate of the expectation of $\mathbf{Y} := (Y_1, \ldots, Y_n)$. Show that the set of possible estimates of $\mathbb{E}\mathbf{Y}$ is a subspace of $\mathbb{R}^n$; we can also think of these estimates as fitted values and thus define *residuals* just as in Chapter 1. Explain how to find the estimate that minimizes the sum of squared residuals.

2.1. Visualizing the observations

As we discuss modeling assumptions, we'll add them into the pictures we developed in Chapter 1. Paralleling our approach in that chapter, we'll begin by visualizing the *observations*.

2.1.1. 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 sometimes called *location models*.

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

2.2 Suppose $Y_i = \mu + \epsilon_i$ with mean-zero errors $\epsilon_1, \ldots, \epsilon_n$. What is the expected value of $\bar{Y} := \frac{1}{n}\sum Y_i$? What is the expected value of the residual $Y_i - \bar{Y}$? Now also assume that the errors are uncorrelated and each one has variance σ_0^2. What is the variance of $\bar{Y}$? What is the variance of $Y_i - \bar{Y}$? Show your work.

We can easily include the new ingredients into our picture; see Figure 2.1 which depicts in particular an iid Normal distribution for the errors.

2.1.2. 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 2.2, 2.3, and 2.4 put the new ingredients into the scatterplot picture, depicting in particular an iid Normal distribution for the errors.

2.3 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.

Suppose we want to estimate the "true" line, assuming the model is correct. We might consider using the least-

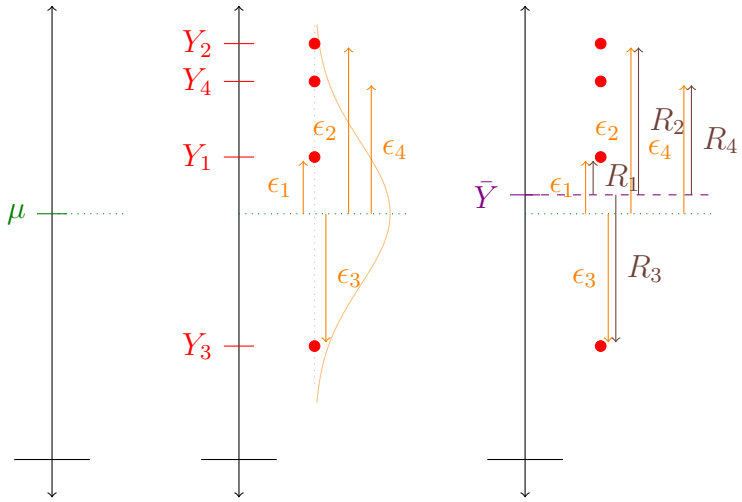


Figure 2.1: Left: According to the location model, the expected value of each response is μ. Middle: The response variable values would be *at μ except that they are "kicked" away by the random errors. This is a probabilistic explanation for the data in Figure 1.1. Right: The least-squares point provides a fitted value for the response variable. In Figure 1.1 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.*

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 3, we'll assume specifically that the errors are Normal, which will allow us to devise hypothesis tests and confidence intervals.

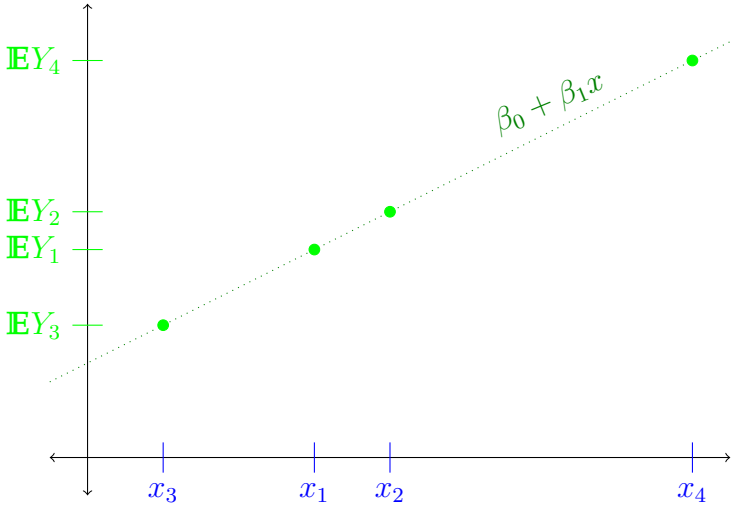


Figure 2.2: According to the simple linear model, there is a true line along which the expected values of the response variable lie.

2.1.3. The multiple linear model

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

$$Y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i \tag{2.2}$$

with $(\beta_0, \ldots, \beta_m)$ ranging over $\mathbb{R}^{m+1}$ and with mean-zero errors.[3]

In Chapter 1, 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

[3]It is left to the reader to envision analogues of Figures 2.2, 2.3, and 2.4 with two explanatory variables.

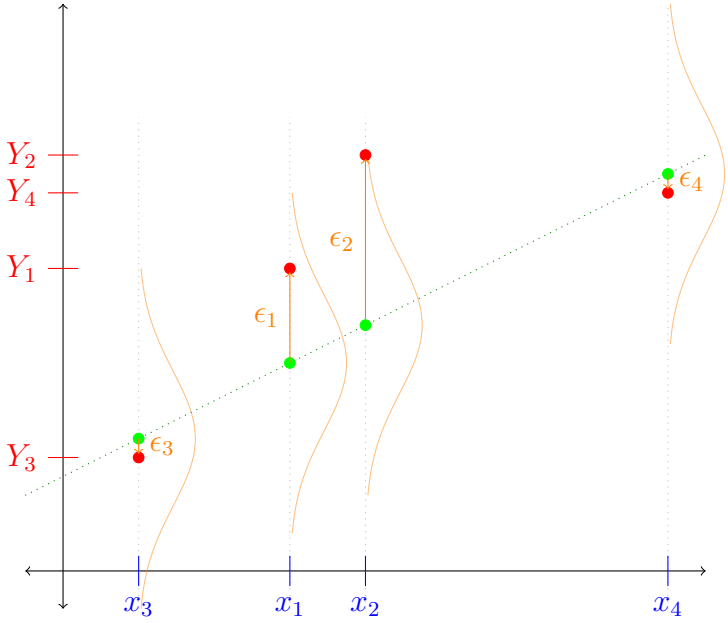


Figure 2.3: For each observation, an error kicks the response variable away from its expectation on the true line. This is a probabilistic explanation for the data in Figure 1.2.

observe that the vector $\hat{\mathbf{Y}}$ of fitted values and the vector of least-squares coefficients 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 β when we assume that the data actually behaves according to the model.

2.4 Let $\mathbf{Y}$ be an $\mathbb{R}^n$-valued random vector with covariance matrix $\mathbf{C}$. Let $\mathbb{M}$ be a real matrix with

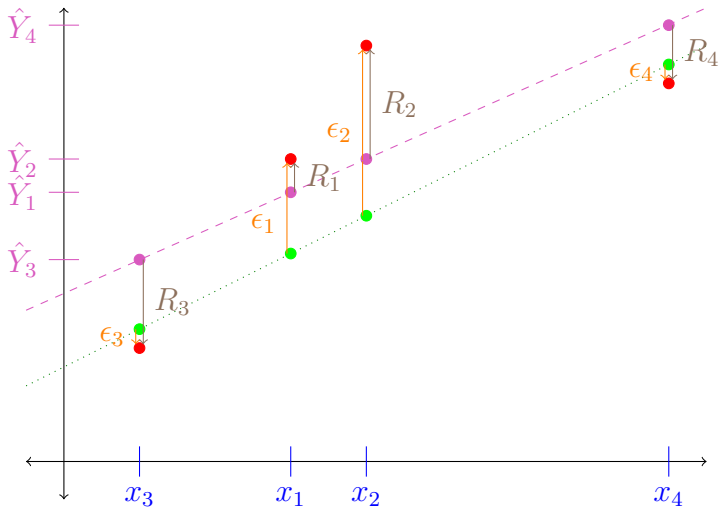


Figure 2.4: The least-squares line provides fitted values for the response variable. In Figure 1.3 it served to describe the data, but now we also consider it an estimator of the true line.

n columns. Use the definition of the covariance matrix to find a formula for $\text{cov}(\mathbb{M}\mathbf{Y})$ in terms of $\mathbb{C}$.

2.5 Suppose the response variable is truly governed by the assumptions of (2.2). Find the expected values of $\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\mathbf{R} := \mathbf{Y} - \hat{\mathbf{Y}}$. Assume also that all of the errors have the same variance σ^2. What is the expected value of the squared length of $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$? Assuming further that the errors are uncorrelated, find the covariance matrix of $\hat{\mathbf{Y}}$. What does it simplify to if $\mathbb{X}$ is orthonormal?

(Useful facts for finding $\text{cov}(\hat{\mathbf{Y}})$: 1. If $\mathbb{M}^-$ is the Moore-Penrose inverse of $\mathbb{M}$, then $\mathbb{M}^-\mathbb{M}\mathbb{M}^- = \mathbb{M}^-$. 2. The Moore-Penrose inverse of a symmetric matrix is itself symmetric.)

2.6 Suppose the response variable is truly governed by the assumptions of (2.2) and that the columns of $\mathbb{X}$ are linearly independent, find the expected value of the least-squares coefficient vector, which we will call $\hat{\boldsymbol{\beta}}$ in this context. Assuming also that the errors are uncorrelated and that they all have the same variance σ^2, find the covariance matrix of $\hat{\boldsymbol{\beta}}$. What does it simplify to if $\mathbb{X}$ is orthonormal?

2.2. 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 $\mathbb{X}$.

2.2.1. 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.

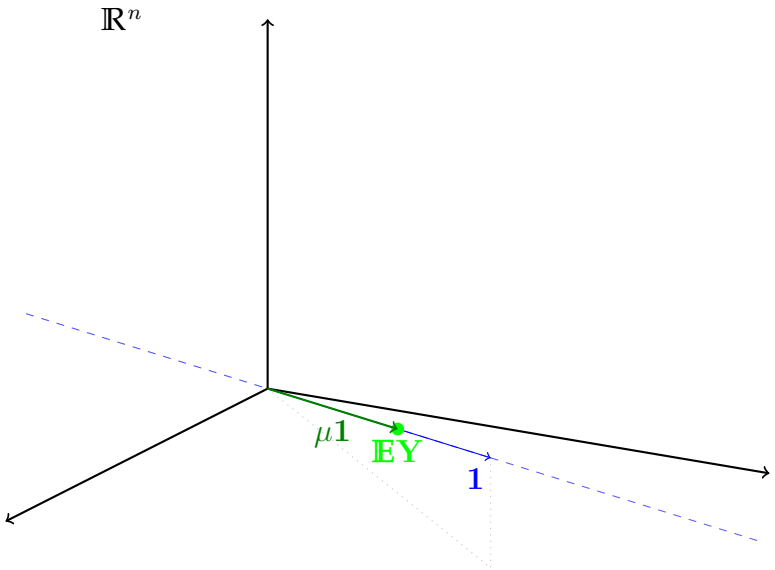


Figure 2.5: A generic picture of the constant vector $\mathbf{1}$*, its span, and the expectation of the response variable vector* $\mathbb{E}\mathbf{Y} = \mu\mathbf{1}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*.*

2.2.2. 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 2.8, 2.9, and 2.10 to our earlier Figures 1.8 and 1.9 to see what the model adds.

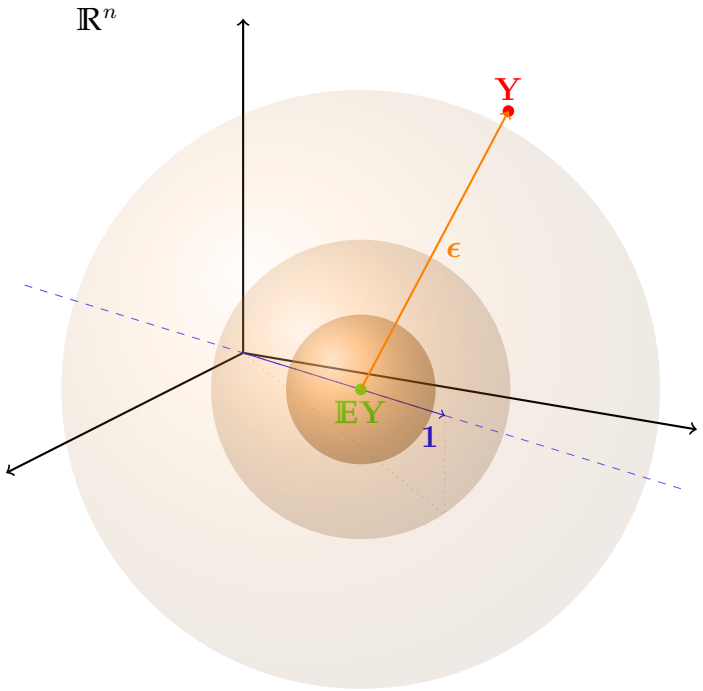


Figure 2.6: A generic picture of the constant vector $\mathbf{1}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, a density for the error vector, and a realization of that error* $\boldsymbol{\epsilon}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*. (The density depicted is spherically symmetric, bringing to mind the symmetric multivariate Normal density, though we won't specifically assume Normal errors until Chapter 3.) The error vector kicks* $\mathbf{Y}$ *out into space away from its expectation.*

In Chapter 1, 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, there are infinitely many linear combinations of $\mathbf{1}$ and $\mathbf{x}$ that result in the orthog-

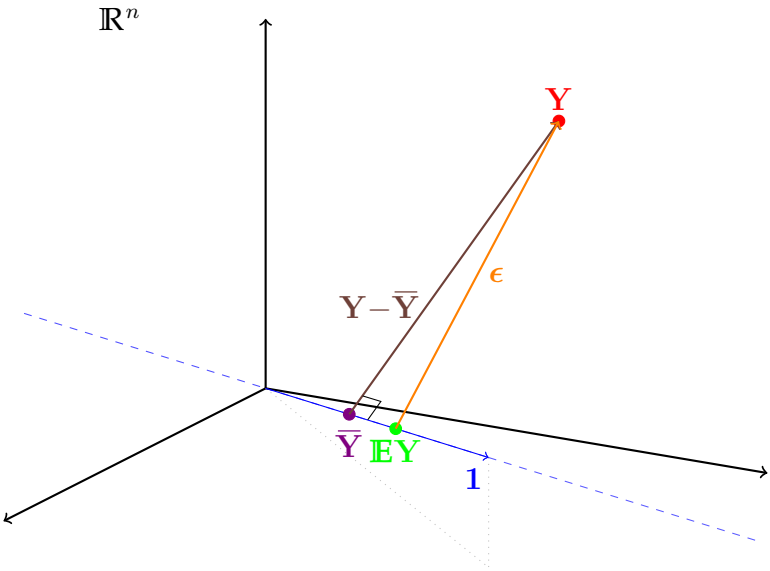


Figure 2.7: A generic picture of the constant vector $\mathbf{1}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, and a realization of the error* $\boldsymbol{\epsilon}$*, assuming all the components of* $\mathbf{Y}$ *do indeed have the same expectation* μ*. The orthogonal projection* $\hat{\mathbf{Y}}$ *of the response* $\mathbf{Y}$ *back onto span*$\{\mathbf{1}\}$ *can be thought of as an estimator for* $\mathbb{E}\mathbf{Y}$*.*

onal 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, then we couldn't possibly hope to tell which one is the *true* parameter value no matter how

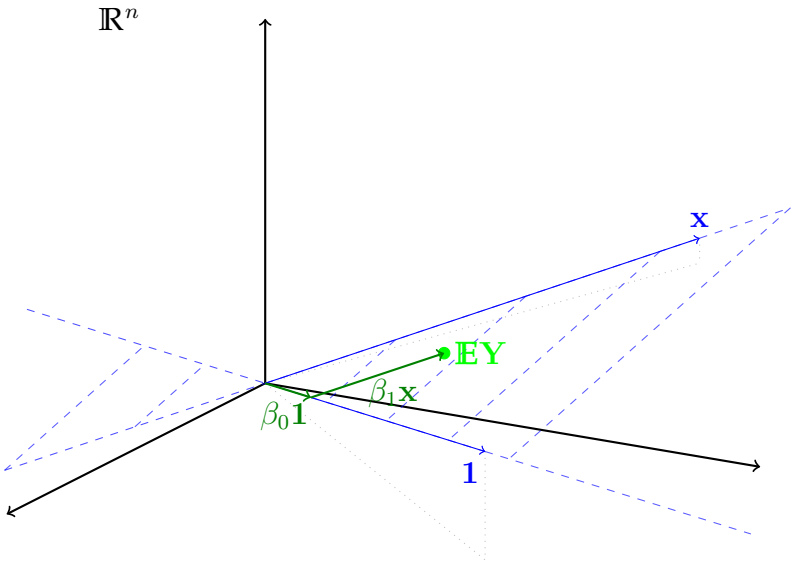


Figure 2.8: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, and the expectation of the response variable vector* $\mathbb{E}\mathbf{Y} = \beta_0\mathbf{1} + \beta_1\mathbf{x}$ *in span*$\{\mathbf{1}, \mathbf{x}\}$*, assuming the simple linear model is true.*

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.

2.7 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 three different pairs of (b_0, b_1) for which $b_0\mathbf{1} + b_1\mathbf{x}$ is equal to the least-squares fit. Argue that the de-

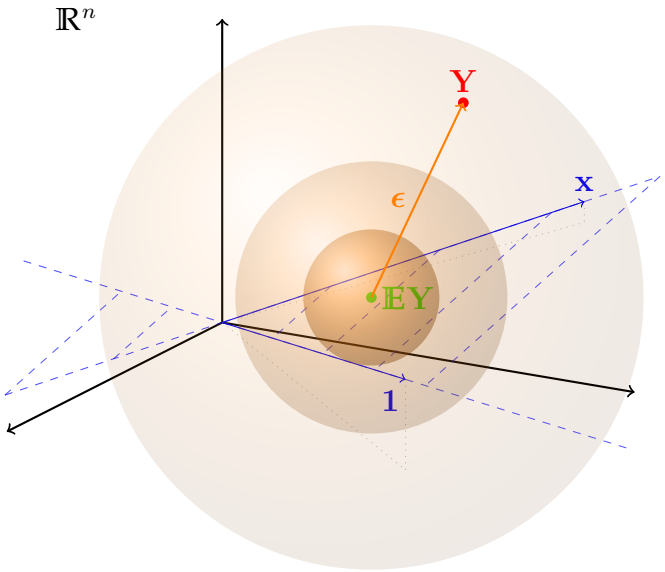


Figure 2.9: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, a density for the error vector, and a realization of that error* $\boldsymbol{\epsilon}$*, assuming the simple linear model is true. (The density depicted is spherically symmetric, bringing to mind the symmetric multivariate Normal density, though we won't specifically assume Normal errors until Chapter 3.) The error vector kicks* $\mathbf{Y}$ *out into space away from its expectation.*

rived parameter $\tilde{b} := b_0 + cb_1$ is identifiable. What is the least-squares estimate for $\tilde{b}$?

2.8 Let X be a random variable with a finite expected value which we will denote μ. Show that for any

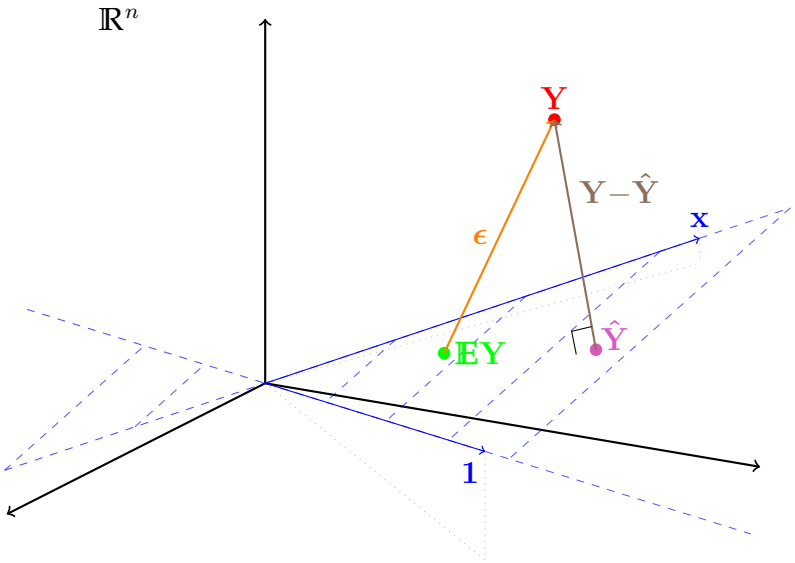


Figure 2.10: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the expected response* $\mathbb{E}\mathbf{Y}$*, and a realization of the error* $\boldsymbol{\epsilon}$*, assuming the simple linear model is true. The orthogonal projection* $\hat{\mathbf{Y}}$ *of the response* $\mathbf{Y}$ *back onto span*$\{\mathbf{1}, \mathbf{x}\}$ *can be thought of as an estimator for* $\mathbb{E}\mathbf{Y}$*.*

$a \in \mathbb{R}$,

$$\mathbb{E}(X-a)^2 = \mathrm{var}(X) + (\mu - a)^2$$

by expressing $(X-a)$ as $(X-\mu)-(a-\mu)$, then multiplying the square. (We will call this identity the *bias-variance decomposition for random variables.*) What value of a minimizes $\mathbb{E}(X-a)^2$? Explain your answer using the bias-variance decomposition.

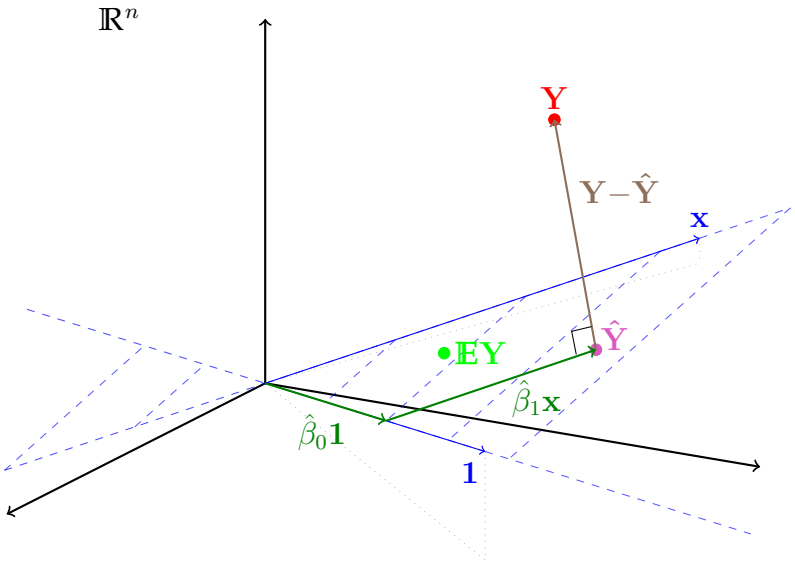


Figure 2.11: A generic picture of the constant vector $\mathbf{1}$*, an explanatory variable vector* $\mathbf{x}$*, the response* $\mathbf{Y}$*, and its orthogonal projection* $\hat{\mathbf{Y}}$ *back into the column space of design matrix. The least-squares coefficients of* $\mathbf{1}$ *and* $\mathbf{x}$ *can be considered estimators for the true coefficients* β_0 *and* β_1*.*

2.2.3. 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. It generalizes the location model and the simple linear model, and yet it is still a special case of the general form (2.1).

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 2.12, 2.13, and 2.14.

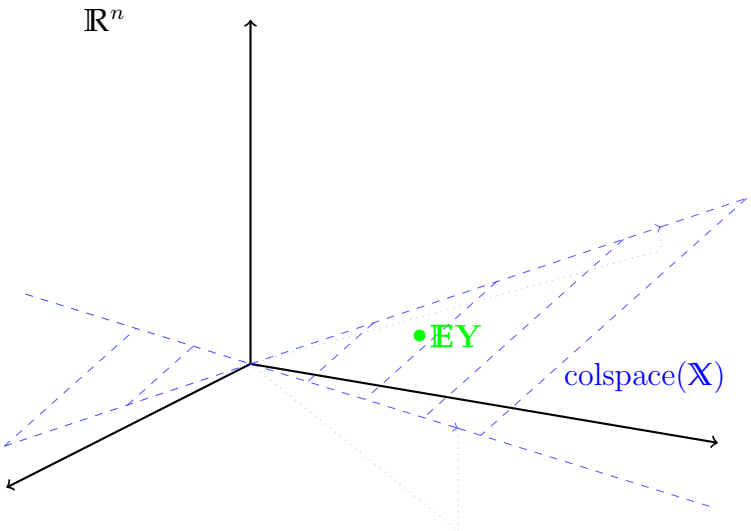


Figure 2.12: Assuming the linear model is true, $\mathbb{E}\mathbf{Y} = \mathbb{X}\boldsymbol{\beta}$ lies in the column space of the $\mathbb{X}$ matrix.

Before analyzing the quantities in the picture any further, it will help if we review some aspects of linear algebra with particular attention to *orthogonal projection* matrices.

Recall that any $n \times n$ real symmetric matrix $\mathbb{M}$ has a representation

$$\mathbb{M} = \mathbb{Q}\mathbb{D}\mathbb{Q}' \tag{2.3}$$

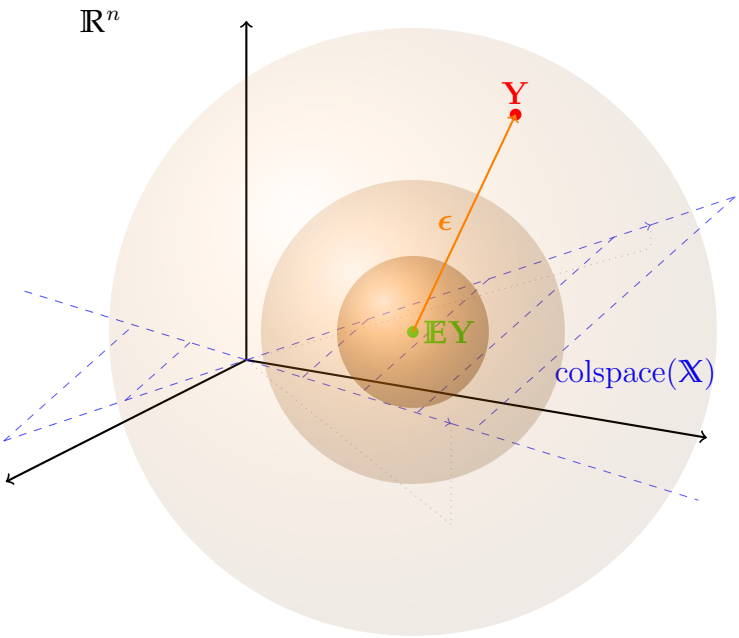


Figure 2.13: The error vector kicks the response $\mathbf{Y}$ *off into space away from its expectation.*

where $\mathbb{Q}$ is a real orthonormal matrix and $\mathbb{D}$ is a diagonal $n \times n$ matrix containing the eigenvalues of $\mathbb{M}$ (which are all real) with their proper multiplicities.

Recall also the trace operator's *invariance under cyclic permutation* which says that the left-most matrix in a product can be moved to the right-most position without changing the value of the trace of the product. Using this property along with representation (2.3), it is easy to see that

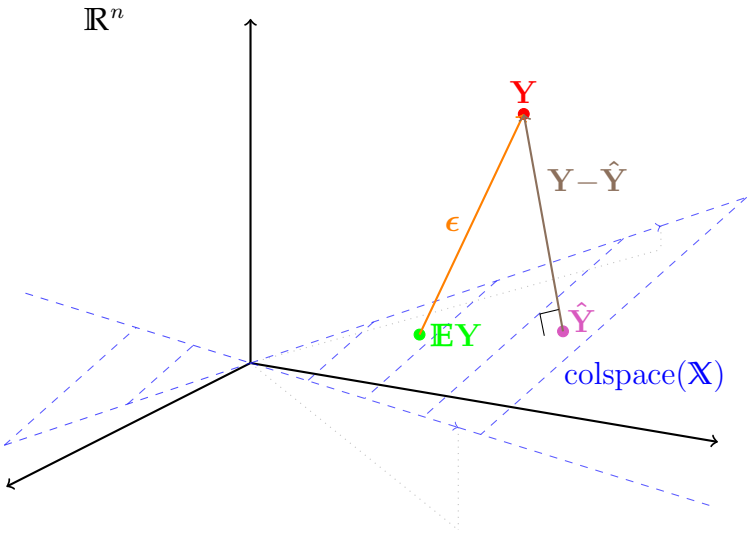


Figure 2.14: The orthogonal projection $\hat{\mathbf{Y}}$ of the response $\mathbf{Y}$ back onto the column space of $\mathbb{X}$ can be considered an estimator for $\mathbb{E}\mathbf{Y}$.

the trace of $\mathbb{M}$ equals the sum of its eigenvalues.

$$\begin{aligned} \operatorname{tr} \mathbb{M} &= \operatorname{tr} [\mathbb{Q}\mathbb{D}\mathbb{Q}'] \\ &= \operatorname{tr} [\mathbb{D}\mathbb{Q}'\mathbb{Q}] \\ &= \operatorname{tr} \mathbb{D} \end{aligned}$$

Now we'll consider the eigenvalues of an orthogonal projection matrix based on its behavior. We know that an orthogonal projection matrix onto the subspace $\mathcal{S} \subseteq \mathbb{R}^n$ maps any given vector in $\mathbb{R}^n$ to the vector in $\mathcal{S}$ that is closest to it. Clearly vectors that are already in $\mathcal{S}$ must be left alone by the orthogonal projection matrix. From this, we can conclude that $\mathcal{S}$ is an eigenspace of the orthogonal

projection vector, and its eigenvalue is 1. Next, consider a vector $\mathbf{v}$ that is orthogonal to $\mathcal{S}$. We know that $\mathbf{v}$ minus its orthogonal projection $\mathbf{v}_{\mathcal{S}}$ is supposed to be orthogonal to $\mathcal{S}$. But if $\mathbf{v}$ is already orthogonal to $\mathcal{S}$, then $\mathbf{v} - \mathbf{v}_{\mathcal{S}} \perp \mathcal{S}$ is satisfied by $\mathbf{v}_{\mathcal{S}} = \mathbf{0}$. So the vectors perpendicular to $\mathcal{S}$ comprise a second eigenspace of the orthogonal projection matrix, and its eigenvalue is 0. These two eigenspaces ($\mathcal{S}$ and its orthogonal complement) together span all of $\mathbb{R}^n$, so there's no room for any other eigenvectors or eigenvalues.

This argument leads us to conclude that the only eigenvalues of an orthogonal projection matrix are 0 and 1. The multiplicity of 1 is the dimension of $\mathcal{S}$, while the multiplicity of 0 is n minus the dimension of $\mathcal{S}$. Therefore the sum of the orthogonal projection matrix's eigenvalues (which also equals its trace) is the dimension of $\mathcal{S}$.

A real matrix $\mathbb{H}$ is an orthogonal projection matrix if and only if it is both *idempotent*[4] ($\mathbb{H}\mathbb{H} = \mathbb{H}$) and *symmetric* ($\mathbb{H}' = \mathbb{H}$). These two properties together let us represent the squared norm of $\mathbb{H}\mathbf{v}$ conveniently as a quadratic form:

$$\begin{aligned} \|\mathbb{H}\mathbf{v}\|^2 &= (\mathbb{H}\mathbf{v})'(\mathbb{H}\mathbf{v}) \\ &= \mathbf{v}'\mathbb{H}'\mathbb{H}\mathbf{v} \\ &= \mathbf{v}'\mathbb{H}\mathbf{v}. \end{aligned} \tag{2.4}$$

Several of the quantities in our picture are orthogonal projections of random vectors. To analyze their expected squared lengths, we'll make use of (2.4) along with a well-

[4]From Latin "idem" (root of "identical") meaning *the same* and "potentia" meaning *power*. It's an appropriate name, as all positive integer powers of the matrix are equal: $\mathbb{H}^k = \mathbb{H}$.

known formula for the expected value of a quadratic form of a random vector. Let's work out that formula first.

A clever trick for manipulating quadratic forms is to realize that the result of a vector-matrix-vector multiplications is just a number, which is a 1×1 matrix and is equal to its trace; this allows us to make use of properties of the trace operator as you will see in the derivation of (2.5).

Let $\mathbf{Y}$ be an $\mathbb{R}^n$-valued random vector with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance matrix $\mathbb{C}$. For any $n \times n$ real matrix $\mathbb{M}$,

$$\begin{aligned}
\mathbb{E}\mathbf{Y}'\mathbb{M}\mathbf{Y} &= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu} + \boldsymbol{\mu})'\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu} + \boldsymbol{\mu}) \\
&= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})'\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu}) + \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})'\mathbb{M}\boldsymbol{\mu} \\
&\quad + \mathbb{E}\boldsymbol{\mu}'\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu}) + \mathbb{E}\boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu} \\
&= \mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})'\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu}) + \boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu} \\
&= \mathbb{E}\,\mathrm{tr}\,[(\mathbf{Y} - \boldsymbol{\mu})'\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu})] + \boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu} \\
&= \mathbb{E}\,\mathrm{tr}\,[\mathbb{M}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] + \boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu} \\
&= \mathrm{tr}\,[\mathbb{M}\,\mathbb{E}(\mathbf{Y} - \boldsymbol{\mu})(\mathbf{Y} - \boldsymbol{\mu})'] + \boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu} \\
&= \mathrm{tr}\,[\mathbb{M}\mathbb{C}] + \boldsymbol{\mu}'\mathbb{M}\boldsymbol{\mu}. \qquad (2.5)
\end{aligned}$$

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.

We discussed earlier in this section that the trace of an orthogonal projection matrix $\mathbb{H}$ is equal to the dimension of the space $\mathcal{S}$ that it projects onto. Now we'll explain that the dimension of $\mathcal{S}$ is exactly equal to the rank of $\mathbb{H}$, because $\mathcal{S}$ is exactly equal to the column space of $\mathbb{H}$.

This isn't hard to see. Let $\mathbb{H}$ be the orthogonal projection matrix onto $\mathcal{S}$. If $\mathbf{v} \in \mathcal{S}$, then we know that it gets projected to itself: $\mathbb{H}\mathbf{v} = \mathbf{v}$. This tells us that $\mathbf{v}$ is in the column space of $\mathbb{H}$; because $\mathbf{v}$ was an arbitrary choice from $\mathcal{S}$, we can tell that $\mathcal{S}$ is a subspace of the column space of $\mathbb{H}$. Next assume that $\mathbf{v}$ is in the column space of $\mathbb{H}$, that is $\mathbf{v} = \mathbb{H}\mathbf{u}$ for some $\mathbf{u} \in \mathbb{R}^n$. It will follow from idempotence that $\mathbb{H}\mathbf{v} = \mathbf{v}$, which tells us that $\mathbf{v}$ is indeed in the space $\mathcal{S}$ that $\mathbb{H}$ projects onto.

$$\begin{aligned}\mathbb{H}\mathbf{v} &= \mathbb{H}\mathbb{H}\mathbf{u} \\ &= \mathbb{H}\mathbf{u} \\ &= \mathbf{v}\end{aligned}$$

We've shown that $\mathcal{S}$ is a subspace of the column space of $\mathbb{H}$ and vice versa, so we conclude that the two spaces are equal.

2.9 Let $\mathbb{H}$ be an orthogonal projection matrix. Show that $\mathbb{I} - \mathbb{H}$ is an orthogonal projection matrix onto the orthogonal complement of the column space of $\mathbb{H}$. (Hint: you might want to show that behaves "correctly" for vectors that are either in the column space of $\mathbb{H}$ or orthogonal to it. Then realize that a basis for $\mathbb{R}^n$ can be chosen from among these vectors. Finally, argue that if $\mathbb{I} - \mathbb{H}$ behaves the same as an orthogonal projection matrix on a basis, then it also behaves the same for every vector in $\mathbb{R}^n$.) Use this to observe that the residual vector is an orthogonal projection of $\mathbf{Y}$.

From here on, we'll use $\mathbb{H}$ to represent the orthogonal

projection matrix $\mathbb{X}(\mathbb{X}'\mathbb{X})^{-}\mathbb{X}'$ onto the column space of $\mathbb{X}$.

2.10 Verify that the orthogonal projection matrix $\mathbb{H}$ onto the column space of $\mathbb{X}$ is indeed idempotent and symmetric.

2.11 Suppose the response variable is truly governed by the assumptions of (2.2) and that the errors are uncorrelated and all have the same variance σ^2. Find the covariance matrix of the residual vector. Next, use (2.4) and (2.5) to find the expected values of the sum of squared residuals $\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ and the expected value of the squared loss $\|\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$. (Use the Pythagorean theorem, in conjunction with one of the quantities you calculated in Exercise **2.5**, to verify your answers.) If σ is unknown, devise an unbiased estimator for σ^2 that is proportional to the sum of squared residuals.

2.12 Suppose the response variable is truly governed by the assumptions of (2.2) and that the errors are uncorrelated and all have the same variance σ^2. Consider fits of the form $\alpha\hat{\mathbf{Y}}$. Use (2.4) and steps similar to (2.5) to derive an expression for the expected squared loss $\mathbb{E}\|\alpha\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}\|^2$ that involves two terms: one with the factor α^2 and another with the factor $(1-\alpha)^2$. Find the $\alpha \in \mathbb{R}$ for which $\alpha\hat{\mathbf{Y}}$ has the smallest expected squared loss.

It follows from the *Gauss-Markov Theorem* (which the interested reader can find elsewhere)

that the least-squares fit $\hat{\mathbf{Y}}$ has the smallest possible expected squared loss among all estimators of $\mathbb{E}\mathbf{Y}$ that are both linear (as a function of $\mathbf{Y}$) and unbiased. Explain why your observation about $\alpha\hat{\mathbf{Y}}$ doesn't contradict the Gauss-Markov Theorem.

Up next, we'll add a Normality assumption on the errors which will open the door to a variety of new opportunities for statistical inference.

CHAPTER

3

NORMAL ERRORS

THROUGHOUT THIS CHAPTER, we'll continue analyzing the linear model (2.1), but we will assume in particular that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$ for an unknown σ. The figures in Chapter 2 already showed iid Normal errors, although the results we derived in that chapter didn't make 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 derive much more, including hypothesis tests and confidence intervals.

Let's start with a few warm-up exercises to reinforce some of the things you've learned so far and guide you

toward the ideas that will be developed in this chapter.

3.1 Picking up where Exercise 2.5 left off, assume further that the errors are independent and Normal. What are the distributions of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$? Picking up where Exercise 2.6 left off, assume further that the errors are independent and Normal. What is the distribution of $\hat{\boldsymbol{\beta}}$?

3.2 Find the density of $\boldsymbol{\epsilon} := (\epsilon_1, \ldots, \epsilon_n)$ if $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$. Looking at this density, explain what is meant when the $N(\mathbf{0}, \sigma^2 \mathbb{I})$ distribution is called *spherically symmetric*.

3.3 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.) What is the maximum likelihood procedure for this model? Does it matter whether σ is known or unknown?

3.4 Suppose $Y_1, \ldots, Y_n$ are modeled as iid $N(\mu, \sigma^2)$ with unknown $\mu \in \mathbb{R}$. Use Exercises 3.3 and 1.1 to find the maximum-likelihood estimator $\hat{\mu}_{\text{MLE}}$ for μ. Explain your reasoning. Does it matter whether σ is assumed to be known or unknown?

3.1. Spherical symmetry

Much of the analysis in this chapter will stem from a crucial observation about the spherical symmetry of the $N(\mathbf{0}, \sigma^2 \mathbb{I})$ distribution; spherical level sets for its density are drawn in Figure 3.1 from an arbitrary three-dimensional perspective.

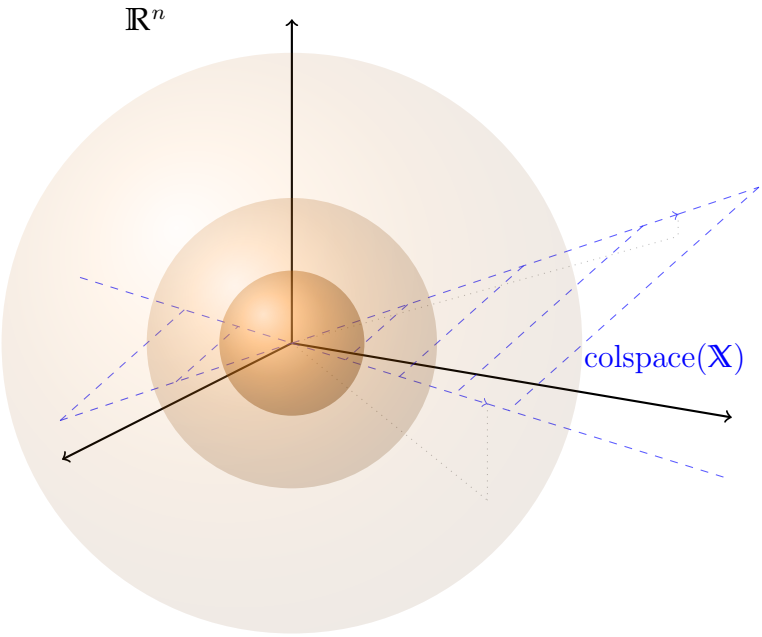


Figure 3.1: If $\epsilon_1, \ldots, \epsilon_n$ are iid $N(0, \sigma^2)$, then $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$. The density of the $N(\mathbf{0}, \sigma^2 \mathbb{I})$ distribution is spherically symmetric about $\mathbf{0}$.

The error vector $\boldsymbol{\epsilon}$ is defined by having the iid Normal draws $\epsilon_1, \ldots, \epsilon_n$ as its components. Alternatively, the components $\epsilon_1, \ldots, \epsilon_n$ can be thought of as the coordinates of a *single draw* of $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbb{I})$. Now suppose that we come up with a new 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, it is clear that 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 $N(0, \sigma^2)$.*

Now suppose in particular that the first $\text{rank}(\mathbb{X})$ new axes are inside the column space of $\mathbb{X}$; the remaining $n - \text{rank}(\mathbb{X})$ new axes must be orthogonal to that column space.[1] See Figure 3.2 for an example. Using the new axes, the orthogonal projection of $\boldsymbol{\epsilon}$ onto the column space of $\mathbb{X}$ must have zeros as its last $n - \text{rank}(\mathbb{X})$ coordinates. Likewise, the orthogonal projection of $\boldsymbol{\epsilon}$ onto the orthogonal complement of the column space of $\mathbb{X}$ must have zeros as its first $\text{rank}(\mathbb{X})$ coordinates. Because the sum of these two vectors equals $\epsilon = (\delta_1, \ldots, \delta_n)_\delta$, the vectors must have the representations

$$\mathbb{H}\boldsymbol{\epsilon} = (\delta_1, \ldots, \delta_{\text{rank}(\mathbb{X})}, 0, \ldots, 0)_\delta$$

and

$$(\mathbb{I} - \mathbb{H})\boldsymbol{\epsilon} = (0, \ldots, 0, \delta_{\text{rank}(\mathbb{X})+1}, \ldots, \delta_n)_\delta$$

with respect to the new axes. Furthermore, these two projected vectors are functions of different independent components and must therefore be independent of each other.

3.5 Show that the residual vector $\mathbf{R} := \mathbf{Y} - \hat{\mathbf{Y}}$ is exactly equal to the vector

[1] We will make reference to this basis to prove facts about various statistics. Note that we don't actually need to *construct* a satisfactory new coordinate basis; we only care that such a basis *exists*.

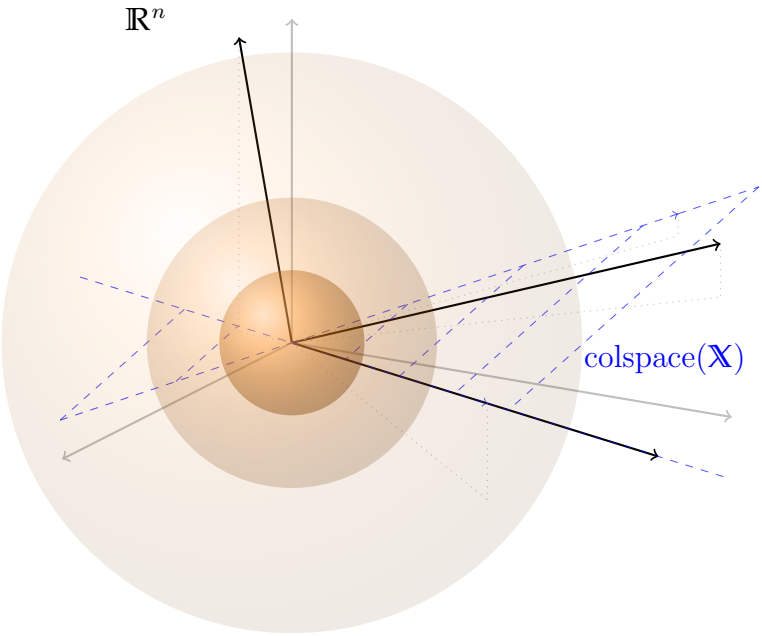


Figure 3.2: What if $\boldsymbol{\epsilon}$ is represented by another choice of orthogonal coordinate axes? By spherical symmetry, we can see that the joint distribution of the coordinate random variables must be the same regardless of the choice of orthogonal coordinate axes. In particular, we consider a choice in which the first 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.

$(0, \ldots, 0, \delta_{\text{rank}(\mathbb{X})+1}, \ldots, \delta_n)_\delta$ as defined above. In other words, show that $(\mathbb{I} - \mathbb{H})\mathbf{Y} = (\mathbb{I} - \mathbb{H})\boldsymbol{\epsilon}$.

3.6 Show that $\hat{\boldsymbol{\beta}}$ is independent of $\mathbf{R}$. (Hint: Express

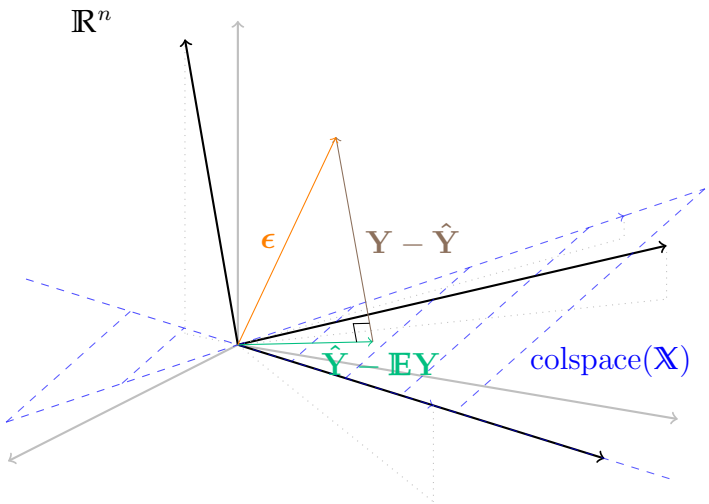


Figure 3.3: The random vectors $\hat{\mathbf{Y}} - \mathbb{E}\mathbf{Y}$ *and* $\mathbf{Y} - \hat{\mathbf{Y}}$ *are functions of distinct subsets of the* δ*-coordinates. This implies that they are independent.*

> $\hat{\boldsymbol{\beta}}$ as a function of $\hat{\mathbf{Y}}$, then express $\hat{\mathbf{Y}}$ as a function of $\mathbb{H}\boldsymbol{\epsilon}$. Connect this to the discussion above to argue for independence.)
>
> **3.7** Use the rotated axes described above to figure out the distributions of $\|\mathbf{R}\|^2/\sigma^2$ and $\|\mathbb{H}\epsilon\|^2/\sigma^2$?

3.2. Inference

Let's use these results to devise a hypothesis test for a coefficient estimate. Assume the linear model is true, the errors are iid Normal, and that $\mathbb{X} \in \mathbb{R}^{n\times(m+1)}$ is full-

rank. Each component of $\hat{\boldsymbol{\beta}}$ is Normal with expectation equal to the true coefficient and variance equal to σ^2 times the corresponding diagonal entry of the covariance matrix $(\mathbf{X}'\mathbf{X})^{-1}$. If σ were known, we could divide $\hat{\beta}_j$ by σs_j (defining $s_j := \sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}$) to get a z-score for the null hypothesis that $\hat{\beta}_j$ is zero. But σ isn't known. This should seem familiar if you've studied t-tests. You may remember substituting an estimate of σ to create a statistic with a known distribution.

Recall how t-distributions arise: if $X \sim N(0,1)$, $V \sim \chi^2_k$, and X and V are independent, then $\frac{X}{\sqrt{V/k}} \sim t_k$. In Exercise **2.11**, we realized that $\frac{\|\mathbf{R}\|^2}{n-m-1}$ is an unbiased estimator for σ^2. In light of this, we define

$$\hat{\sigma} := \frac{1}{\sqrt{n-m-1}}\|\mathbf{R}\|.$$

Now consider the statistic

$$\begin{aligned} T_j &:= \frac{\hat{\beta}_j/s_j}{\hat{\sigma}} \\ &= \frac{\hat{\beta}_j/s_j}{\|\mathbf{R}\|/\sqrt{n-m-1}} \\ &= \frac{\hat{\beta}_j/\sigma s_j}{\sqrt{(\frac{1}{\sigma^2}\|\mathbf{R}\|^2)/(n-m-1)}} \end{aligned} \tag{3.1}$$

The numerator is standard Normal. According to Exercise **3.6**, the numerator and denominator of (3.1) are independent of each other. And according to Exercise **3.7**, the square of the denominator is a χ^2_{n-m-1} random variable divided by $n-m-1$. We conclude that $T_j \sim t_{n-m-1}$.

3.8 Let $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$ with σ unknown. Define $\hat{\sigma} := \sqrt{\frac{1}{n-1}\sum(Y_i - \bar{Y})^2}$. If the null hypothesis is that μ is zero, then the t-statistic of the data is $\frac{\bar{Y}}{\hat{\sigma}/\sqrt{n}}$. Explain how $\hat{\sigma}$ and the t-statistic fit into the above discussion of linear models as a special case.

In the context of hypothesis testing, the hypothesized value of a coefficient is typically zero. But more generally,

$$\frac{(\hat{\beta}_j - \beta_j)/s_j}{\hat{\sigma}}$$

has a t_{n-m-1} distribution. We can use this observation to devise confidence intervals. With t^* representing the $(1-\alpha/2)$ quantile of the t_{n-m-1} distribution,

$$\mathbb{P}\left\{-t^* \leq \frac{(\hat{\beta}_j - \beta_j)/s_j}{\hat{\sigma}} \leq t^*\right\} = 1-\alpha$$
$$\Rightarrow \mathbb{P}\left\{\hat{\beta}_j - t^*\hat{\sigma}s_j \leq \beta_j \leq \hat{\beta}_j + t^*\hat{\sigma}s_j\right\} = 1-\alpha$$

Thus $\hat{\beta}_j \pm t^*\hat{\sigma}s_j$ is a $(1-\alpha)$-level confidence interval for β_j.

Now, consider instead estimation of a linear combination of the coefficients: $\mathbf{v}'\boldsymbol{\beta}$ for some $\mathbf{v} \in \mathbb{R}^{m+1}$.[2] An unbiased estimator is $\mathbf{v}'\hat{\boldsymbol{\beta}}$, which has variance

$$\begin{aligned}\operatorname{var}\mathbf{v}'\hat{\boldsymbol{\beta}} &= \mathbf{v}'(\operatorname{cov}\hat{\boldsymbol{\beta}})\mathbf{v} \\ &= \sigma^2\mathbf{v}'(\mathbb{X}'\mathbb{X})^{-1}\mathbf{v}.\end{aligned}$$

[2]Note that this generalizes the problem of estimating β_j because $\beta_j = \mathbf{e}_j'\boldsymbol{\beta}$ where $\mathbf{e}_j$ is the unit vector in the jth coordinate direction.

By the same reasoning as before, $\mathbf{v}'\hat{\boldsymbol{\beta}} \pm t^* \hat{\sigma} \sqrt{\mathbf{v}'(\mathbb{X}'\mathbb{X})^{-1}\mathbf{v}}$ is a $(1-\alpha)$-level confidence interval for $\mathbf{v}'\boldsymbol{\beta}$.

Let's use this result in the context of simple linear modeling to draw confidence intervals for the true line in the observations picture. At each $x \in \mathbb{R}$, the least-squares line $(1, x)'\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x$ is the estimate of $(1, x)'\boldsymbol{\beta} = \beta_0 + \beta_1 x$, and the confidence interval should reach upward and downward from the estimate by $t^*\hat{\sigma}$ times the square root of

$$\begin{bmatrix} 1 & x \end{bmatrix} (\mathbb{X}'\mathbb{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}.$$

In this case, t^* is the $(1-\alpha/2)$ quantile of t_{n-2}.

3.9 Use the Pythagorean theorem to show that

$$\|\mathbf{x} - \bar{\mathbf{x}}\|^2 = \|\mathbf{x}\|^2 - n\bar{x}^2. \tag{3.2}$$

Explain how this equation can also be derived using the familiar variance formula $\mathrm{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$. (Hint: what if the distribution of X is the empirical distribution of $\mathbf{x}$?)

3.10 In simple linear regression, what is the dimension of the matrix $\mathbb{X}'\mathbb{X}$? Work out $\mathbb{X}'\mathbb{X}$ and its inverse explicitly; simplify the inverse's denominator using (3.2) to end up with

$$(\mathbb{X}'\mathbb{X})^{-1} = \frac{1}{\|\mathbf{x} - \bar{\mathbf{x}}\|^2} \begin{bmatrix} \|\mathbf{x}\|^2/n & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

If the simple linear model is true, and the errors are uncorrelated with variance σ^2, what is the covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)$? Under what condition

are the two coefficient estimates uncorrelated? Do your observations in this exercise require Normal errors?

3.11 Use your derivation in Exercise **3.10** to work out

$$\begin{bmatrix} 1 & x \end{bmatrix} (\mathbb{X}'\mathbb{X})^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

which is one of the quantities involved in calculating confidence intervals for the true line. Complete the square, then use (3.2) to simplify your result to $\frac{1}{\|\mathbf{x}-\bar{\mathbf{x}}\|^2}(x-\bar{x})^2 + \frac{1}{n}$.

3.12 Assume the simple linear model is true and that the errors are iid Normal. Let the data $x_1, \ldots, x_4$ and $y_1, \ldots, y_4$ be the same as in Exercise **1.7**. Sketch the scatterplot of the data along with the least-squares line showing the domain from -1 to 10. Based on the above discussion and the result in Exercise **3.11**, we can calculate a confidence interval at every $x \in \mathbb{R}$. On your scatterplot, sketch dashed lines above and below the least-squares line to indicate the upper and lower endpoints of 90% confidence intervals for the true line. Likewise, sketch the 95% confidence intervals as a pair of dotted curves above and below the least-squares line. Suppose a new observation has $x_5 = 5.5$; give an estimate and a 90% confidence interval for $\mathbb{E}Y_5$. If x_5 had instead been 10 what would the estimate and the 90% confidence interval for $\mathbb{E}Y_5$ have been? Comment on

how the confidence intervals' widths vary along the domain.

3.13 Assume the simple linear model is true and that the errors are iid Normal. Again let the data $x_1, \ldots, x_4$ and $y_1, \ldots, y_4$ be the same as in Exercise 1.7. If the true slope were $\beta_1 = 0$, what would the distribution of $\hat{\beta}_1$ be? Calculate the t-statistic T_1. What would its distribution be if $\beta_1 = 0$? What's the probability that T_1 would be at least as far from zero as we observe in this data set? (i.e. Calculate the significance probability of the test that $\beta_1 = 0$.)

The significance probabilities and the confidence intervals resulting from the above approach are legitimate if only one coefficient (or linear combination of coefficients) is under consideration. But if the data analyst is interested in multiple quantities at once, then alternative methods are more appropriate.

Recall how f-distributions arise: if $V \sim \chi^2_k$, $W \sim \chi^2_l$, and V and W are independent, then $\frac{V/k}{W/l} \sim f_{k,l}$. Consider the hypothesis that a certain subset of the coefficients are all zero. Let $\mathcal{S}_1$ be the span of the variables whose coefficients are hypothesized to be zero, and let $\mathcal{S}_0$ be the span of the remaining variables. Let $\widetilde{\mathbf{Y}}$ be the orthogonal projection of $\mathbf{Y}$ onto $\mathcal{S}_0$ (i.e. the least-squares fit if the variables in question are ignored). Let $\hat{\mathbf{Y}}$ be the least-squares fit for

$\mathbf{Y}$ when all the variables are used. We define the statistic

$$F := \frac{\|\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}}\|^2/k}{\hat{\sigma}^2} = \frac{\|\hat{\mathbf{Y}} - \widetilde{\mathbf{Y}}\|^2/k}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n - \text{rank}(\mathbb{X}))} \tag{3.3}$$

where k is the dimension of the orthogonal complement of $\mathcal{S}_0$ within $\text{span}(\mathcal{S}_0 \cup \mathcal{S}_1)$.

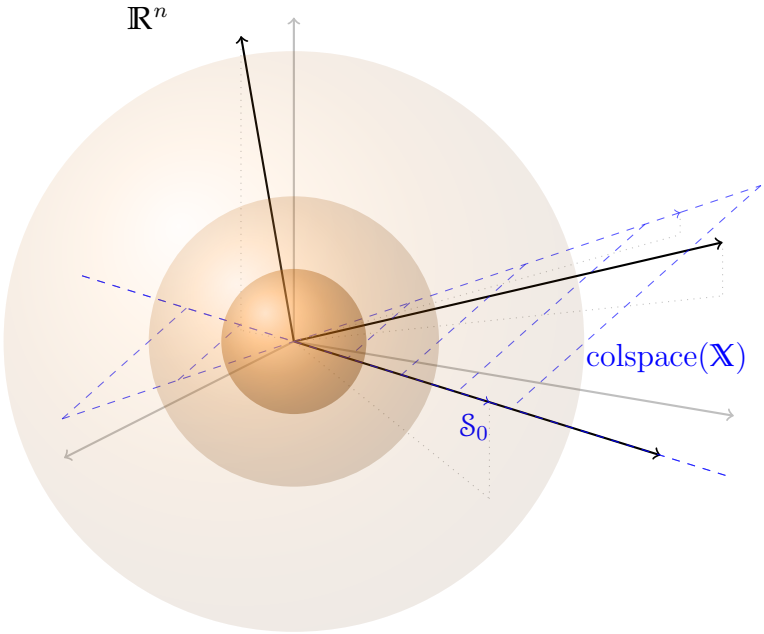


Figure 3.4: Let $\mathcal{S}_0$ be the span of a particular subset of the columns of $\mathbb{X}$. The first $dim(\mathcal{S}_0)$ coordinate axes can be chosen from $\mathcal{S}_0$, then next $rank(\mathbb{X}) - dim(\mathcal{S}_0)$ axes can be chosen from the column space of $\mathbb{X}$ (and necessarily orthogonal to $\mathcal{S}_0$), while the remaining $n - rank(\mathbb{X})$ axes can be chosen orthogonally to the column space of $\mathbb{X}$.

3.14 Let $\mathbb{H}_1$ be the orthogonal projection matrix onto $\mathcal{S}_1 \subseteq \mathbb{R}^n$, and let $\mathbb{H}_0$ be the orthogonal projection matrix onto $\mathcal{S}_0 \subseteq \mathcal{S}_1$. Show that $\mathbb{H}_0 \circ \mathbb{H}_1 = \mathbb{H}_1 \circ \mathbb{H}_0 = \mathbb{H}_0$. (Hint: to analyze $\mathbb{H}_0 \circ \mathbb{H}_1$, you might want to use the fact than any $\mathbf{v} \in \mathbb{R}^n$ has a unique representation as the sum of a vector $\mathbf{v}_0 \in \mathcal{S}_0$ and $\mathbf{v} - \mathbf{v}_0$ orthogonal to $\mathcal{S}_0$.) Explain why $\mathbb{H}_1\mathbf{v} - \mathbb{H}_0\mathbf{v}$ is orthogonal to $\mathcal{S}_0$.

3.15 Assuming the linear model is true and that the errors are iid Normal, what is the distribution of F defined in (3.3)? Justify your answer using an argument analogous to our earlier discussion of the statistic T_j. (Hint: think about Figure 3.4 and Exercise **3.14**.)

The larger the F-statistic is, the more it indicates that the null hypothesis is suspect. Therefore the significance probability of the test is the right-hand tail of the distribution from Exercise **3.15**.

3.3. Categorical explanatory variables

Linear modeling can also be used when some or all of the explanatory variables are categorical. Consider the simplest such case in which there is only one explanatory variable and that it splits the observations into k distinct groups, i.e. each $x_i \in \{1, \ldots, k\}$.

Suppose each of the k groups has its own mean; a natural way to represent this model is

$$\begin{aligned} Y_i &= \mu_{x_j} + \epsilon_i \\ &= \mu_1 \mathbb{I}(x_i = 1) + \ldots + \mu_k \mathbb{I}(x_i = k) + \epsilon_i. \end{aligned} \tag{3.4}$$

Column j of the design matrix has ones at the rows in which the observations belong to group j, and it has zeros elsewhere. The least-squares fit $\hat{\mathbf{Y}}$ estimates each group's expected value by the average of the response values belonging to that group:

$$\begin{aligned} \hat{\mu}_j &:= \bar{Y}_j \\ &:= \frac{1}{n_j} \sum_{i=1}^{n_j} Y_i^{(j)} \end{aligned}$$

where n_j denotes the number of observations from group j and $Y_i^{(j)}$ denotes the ith value in the jth group.

3.16 Explain why it's impossible for the design matrix to have linearly dependent columns in model (3.4). Work out $(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbf{y}$ to verify that the least-squares estimates are indeed the group averages.

An alternative parametrization of the model (3.4) is

$$Y_i = \mu_1 + (\mu_2 - \mu_1)\mathbb{I}(x_i = 2) + \ldots + (\mu_k - \mu_1)\mathbb{I}(x_i = k) + \epsilon_i. \tag{3.5}$$

In this formulation, the design matrix has a **1** column, and it has an indicator variable column for each group except the first (which can be considered the *baseline* group). The

intercept coefficient is the baseline group's mean, and each of the other coefficients is the difference between the corresponding group's mean and the baseline group's mean. The column space of the design matrix is the same whether you use formulation (3.4) or (3.5).

If you want to conduct a hypothesis test or construct a confidence interval for the difference in means between a particular pair of groups, then you can designate one of them to be the baseline group and analyze the other group's least-squares coefficient in (3.5) using the standard inference theory developed above (assuming iid Normal errors).[3]

On the other hand, if you want to test the hypothesis that *all the groups have the same expected value*, you can perform the F-test that assumes all coefficients in (3.5) are zero except the intercept. In this context, that F-test is called **analysis of variance** (or "ANOVA" for short), and the Pythagorean theorem relating $\mathbf{Y}$, $\hat{\mathbf{Y}}$, and $\overline{\mathbf{Y}}$ is called the *ANOVA decomposition*. Because $\hat{\mathbf{Y}}$ fits each observation with its group's sample mean, the *residual sum of squares* term is

$$\begin{aligned} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 &= \sum_i (Y_i - \hat{Y}_i)^2 \\ &= \sum_i (Y_i - \overline{Y}_{x_i})^2 \end{aligned}$$

where $\bar{Y}_{x_i}$ denotes the group sample mean for group x_i.

[3]This generalizes the familiar two-sample t-test. The difference is that if there are more than two groups, the observations from the other groups will also contribute to the $\hat{\sigma}$ estimate.

The *regression sum of squares* term simplifies to

$$\begin{aligned}\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 &= \sum_i (\hat{Y}_i - \overline{Y})^2 \\ &= \sum_i (\overline{Y}_{x_i} - \overline{Y})^2 \\ &= \sum_{j=1}^{k} n_j (\overline{Y}_j - \overline{Y})^2\end{aligned}$$

where n_j is the number of observations in group j. Thus the ANOVA decomposition is

$$\begin{aligned}\sum_i (Y_i - \bar{Y})^2 &= \|\mathbf{Y} - \overline{\mathbf{Y}}\|^2 \\ &= \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 + \|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 \\ &= \sum_i (Y_i - \overline{Y}_{x_i})^2 + \sum_{j=1}^{k} n_j (\overline{Y}_j - \overline{Y})^2. \quad (3.6)\end{aligned}$$

The F-statistic

$$\begin{aligned}F &:= \frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2/(k-1)}{\hat{\sigma}^2} \\ &= \frac{\|\hat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2/(k-1)}{\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2/(n-k)} \\ &= \frac{(\sum_{j=1}^{k} n_j (\overline{Y}_j - \bar{Y})^2)/(k-1)}{(\sum_i (Y_i - \overline{Y}_{x_i})^2)/(n-k)}\end{aligned}$$

has an $f_{k-1,n-k}$ distribution if the errors are iid Normal and all groups share the same expectation.

3.17 The bias-variance decomposition can also be used to derive the ANOVA decomposition. Let $y_1, \ldots, y_n$ be real numbers that are partitioned into k subsets with sizes $n_1, \ldots, n_k$. For $j \in \{1, \ldots, k\}$ and $i \in \{1, \ldots, n_j\}$, let $y_i^{(j)}$ be the ith value in the jth group. For any $a \in \mathbb{R}$,

$$\begin{aligned}\sum_i (y_i - a)^2 &= \sum_j \sum_{i=1}^{n_j} (y_i^{(j)} - a)^2 \\ &= n \sum_j \frac{n_j}{n} \sum_i \frac{1}{n_j} (y_i^{(j)} - a)^2 \\ &= n \mathbb{E}\, \mathbb{E}_{Y \sim \mathbb{P}_J} (Y - a)^2\end{aligned}$$

where the random variable J takes the value j with probability n_j/n and $\mathbb{P}_J$ is the uniform distribution on the response values in group J. Apply the bias-variance decomposition from Exercise **2.8** to $\mathbb{E}_{Y \sim \mathbb{P}_J} (Y - a)^2$, then proceed to derive the ANOVA decomposition (3.6). (Don't worry about the fact that we're writing constants rather than random variables in this exercise; the decomposition for constants implies the same decomposition for random variables.)

3.18 Let X be a random variable with a finite expected value. True or False: $\mathbb{E}X^2$ is finite iff $\mathrm{var} X$ is finite. Explain. (Hint: use the bias-variance decomposition.)

3.19 Explain how the Pythagorean fact observed in Exercise **1.11** can also be proven using the bias-

variance decomposition (Exercise 2.8). (Hint: Divide both sides by n, then think about Exercise 1.2.)

3.20 Let $Y_1, \ldots, Y_n$ be random variables. Let N be uniformly distributed on $\{1, \ldots, n\}$ and independent of $Y_1, \ldots, Y_n$. What is the conditional expected value of Y_N, conditioning on $Y_1, \ldots, Y_n$? (The conditional distribution of Y_N is called the *empirical distribution* of $Y_1, \ldots, Y_n$.) Explain why Exercise 1.2 is a special case of this. (In other words, explain why a constant can be treated as a special case of a random variable. Recall the definition of random variables to answer this question.)

3.4. Prediction

Often, the point of modeling the relationship between a response variable as a set of explanatory variables is to use future explanatory variable values to *predict* the corresponding response variable values. We'll consider the task of prediction in the context of the multiple linear model.

Suppose n observations are available for use in estimation and that the design matrix is full rank. Let $\mathbf{x}_{n+1} := (1, x_{n+1}^{(1)}, \ldots, x_{n+1}^{(m)})$ be the explanatory vector of a new observation, and let Y_{n+1} represent the corresponding response value. If your goal is to minimize the expected squared error of your prediction $\mathbb{E}(Y_{n+1} - \hat{Y}_{n+1})^2$, then the best choice of $\hat{Y}_{n+1}$ would be the expectation of Y_{n+1} given

the explanatory variable values. That expectation is unknown, but if the multiple linear model is true then $\mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}}$ is an unbiased estimator:

$$\begin{aligned}\mathbb{E}\mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}} &= \mathbb{E}(\hat{\beta}_0 + \hat{\beta}_1 x_{n+1}^{(1)} + \ldots + \hat{\beta}_m x_{n+1}^{(m)}) \\ &= \beta_0 + \beta_1 x_{n+1}^{(1)} + \ldots + \beta_m x_{n+1}^{(m)} \\ &= \mathbb{E}Y_{n+1}.\end{aligned}$$

This was observed in Section 3.2 in the context of confidence intervals; from that discussion, we can see that if $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$ then the variance of the estimator is $\sigma^2 v_{n+1}$, defining

$$v_{n+1} := \mathbf{x}'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}\mathbf{x}_{n+1}.$$

If we further assume that the errors are iid Normal, then we can derive a *prediction interval* for Y_{n+1}, an interval that has a specified probability of containing the new response. First, consider the distribution of

$$Y_{n+1} - \hat{Y}_{n+1} = \mathbf{x}'_{n+1}\boldsymbol{\beta} + \epsilon_{n+1} - \mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}}.$$

It is a constant plus a sum of $\mathbf{x}'_{n+1}\hat{\boldsymbol{\beta}}$ and ϵ_{n+1} which are both Normal and are independent of each other, so $Y_{n+1} - \hat{Y}_{n+1}$ is itself Normal. Because $\hat{\boldsymbol{\beta}}$ is unbiased for β and the error has mean zero, the expected value of $Y_{n+1} - \hat{Y}_{n+1}$ is zero. The variance is $\sigma^2 + \sigma^2 v_{n+1}$, so

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

We know that $\hat{\sigma}$ is independent of ϵ_{n+1} and $\hat{\boldsymbol{\beta}}$, so we can conclude that it is independent of $Y_{n+1} - \hat{Y}_{n+1}$. We also

know that $\|\mathbf{R}\|^2/\sigma^2 \sim \chi^2_{n-m-1}$. Putting these observations together,

$$\begin{aligned}\frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{v_{n+1}+1}} &= \frac{(Y_{n+1} - \hat{Y}_{n+1})/(\sigma\sqrt{v_{n+1}+1})}{\sqrt{\frac{1}{\sigma^2}\|\mathbf{R}\|^2/(n-m-1)}} \\ &\sim t_{n-m-1}.\end{aligned}$$

Following the pattern we used earlier to derive confidence intervals,

$$\mathbb{P}\left\{-t^* \leq \frac{Y_{n+1} - \hat{Y}_{n+1}}{\hat{\sigma}\sqrt{v_{n+1}+1}} \leq t^*\right\} = 1 - \alpha.$$

By rearranging, we find that $\hat{Y}_{n+1} \pm t^*\hat{\sigma}\sqrt{v_{n+1}+1}$ is a $(1-\alpha)$-level prediction interval for Y_{n+1}.[4]

3.21 Assume the simple linear model is true and that the errors are iid Normal. Let the data $x_1, \ldots, x_4$ and $y_1, \ldots, y_4$ be the same as in Exercise **1.7**. Sketch the scatterplot of the data along with the least-squares line showing the domain from -1 to 10. Based on the above discussion and the result in Exercise **3.11**, we can calculate a prediction interval at every $x \in \mathbb{R}$. On your scatterplot, sketch the 90% prediction intervals for a new observation

[4] As the sample size increases, the width of the confidence interval for a new observation's expectation tends toward zero. However, the width of the prediction interval will tend toward $2z^*\sigma$ where z^* is the $(1-\alpha/2)$ quantile of the standard Normal distribution. This reflects the fact that with enough data, a parameter can be estimated arbitrarily well, but a new observation comes with its own inherent randomness that limits how well we can predict its value.

using a pair of dashed curves above and below the least-squares line. Sketch the 95% prediction intervals using a pair of dotted curves above and below the least-squares line. Suppose a new observation has $x_5 = 5.5$; give a prediction and a 90% prediction interval for Y_5. If x_5 had instead been 10 what would the prediction and the 90% prediction interval for Y_5 have been? Comment on how the prediction intervals' widths vary along the domain.

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.

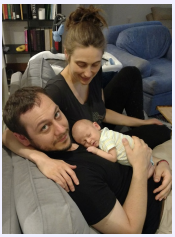
Remarkably, random variables can also be profitably understood with two pictures that are perfectly analogous to those we've been studying. A random variable defined on a probability space has a distribution on $\mathbb{R}$ (the observation picture), but it can also be thought of as a vector in the space 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*.

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.