Euclidean geometry and projections

4433LALG3: Linear Algebra

Week 2, Lecture 7, Valente Ramírez

Mathematical & Statistical Methods group — Biometris, Wageningen University & Research

Overview

- Euclidean geometry
- Orthogonal projections
- Supplementary material

References:

- Nicholson §4.1 & 4.2 (up to and including *'Projections'*).
- Nicholson §5.3
- 3Blue1Brown Ch.9

Section 1

Euclidean geometry

From vectors to geometry

So far, we've exploited the two main properties of vectors in $\mathbb{R}^n$:

- Two vectors can be added
- A vector can be re-scaled by any real number

However, the space $\mathbb{R}^n$ has an additional *geometric structure*.

The two fundamental operations that provide this geometric structure are:

- We can measure length
- We can measure angles

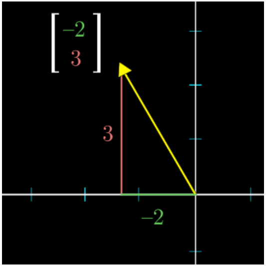
Length of a vector

The length of a vector in $\mathbb{R}^2$ is readily given by *Pythagoras' theorem*:

$$c^2 = a^2 + b^2$$

Thus

$$\ell = \sqrt{(-2)^2 + 3^2} = \sqrt{13}$$



Definition

The **length** or **norm** of a vector $\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \ldots & x_n \end{bmatrix}^\top$ in $\mathbb{R}^n$ is denoted by $\|\mathbf{x}\|$, and given by the formula:

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \ldots + x_n^2}.$$

Properties of the norm

Fact

The norm operation $\|\cdot\|$ satisfies the following properties:

- $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$
- $\|\mathbf{x}\| = 0$ only if $\mathbf{x} = \mathbf{0}$
- $\|\alpha\mathbf{x}\| = |\alpha| \, \|\mathbf{x}\|$ for any scalar α and $\mathbf{x} \in \mathbb{R}^n$

Distance between two vectors

We use the norm to quantify how different two vectors are.

Definition

The **distance** between two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ is the number $\|\mathbf{x} - \mathbf{y}\|$.

The dot product

Definition

Given two vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$, their **dot product**, denoted $\mathbf{v} \cdot \mathbf{w}$, is the real number:

$$\mathbf{v} \cdot \mathbf{w} = \mathbf{v}^\top \mathbf{w} = \begin{bmatrix} v_1 & \dots & v_n \end{bmatrix} \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = v_1 w_1 + \dots + v_n w_n.$$

Remark

This concept is extremely important, more than it seems at first glance.

Like other important concepts, it has multiple names: *dot product, inner product*, or *scalar product* mean all the same thing.

Properties of the dot product

Fundamental properties

Let $\mathbf{u}, \mathbf{v}, \mathbf{w}$ be vectors in $\mathbb{R}^n$ and α a scalar.

The dot product has the following properties:

1. It is a map $\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$
    - This means that given input $(\mathbf{v}, \mathbf{w})$, the output $\mathbf{v} \bullet \mathbf{w}$ is a real number
2. It is symmetric
    - That is, $\mathbf{v} \bullet \mathbf{w} = \mathbf{w} \bullet \mathbf{v}$
3. It is positive-definite
    - $\mathbf{v} \bullet \mathbf{v} \geq 0$ for all $\mathbf{v}$
    - $\mathbf{v} \bullet \mathbf{v} = 0$ only if $\mathbf{v} = \mathbf{0}$
4. It is bilinear
    - $(\mathbf{u} + \mathbf{v}) \bullet \mathbf{w} = \mathbf{u} \bullet \mathbf{w} + \mathbf{v} \bullet \mathbf{w}$
    - $(\alpha \mathbf{v}) \bullet \mathbf{w} = \alpha(\mathbf{v} \bullet \mathbf{w})$

Properties of the dot product

Secondary properties

Moreover, the dot product has the following property:

5. $\mathbf{v} \bullet \mathbf{w} = \|\mathbf{v}\| \, \|\mathbf{w}\| \, \cos\theta$,

   where θ is the angle between the vectors $\mathbf{v}$ and $\mathbf{w}$

   In particular,

   - $\mathbf{v} \bullet \mathbf{v} = \|\mathbf{v}\|^2$ (or equivalently, $\|\mathbf{v}\| = \sqrt{\mathbf{v} \bullet \mathbf{v}}$),
   - $\mathbf{v} \bullet \mathbf{w} = 0$ if and only if $\mathbf{v}$ and $\mathbf{w}$ are *orthogonal*.

Definition

Two vectors $\mathbf{v}$ and $\mathbf{w}$ are **orthogonal** whenever one of the following holds:

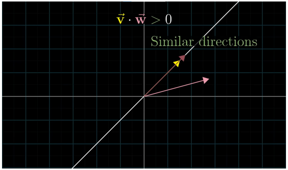
- One of the two vectors is $\mathbf{0}$
- The angle between the two vectors is 90°

The sign of the dot product

It follows from $\mathbf{v} \bullet \mathbf{w} = \|\mathbf{v}\|\ \|\mathbf{w}\| \cos\theta$, that there are three possibilities:

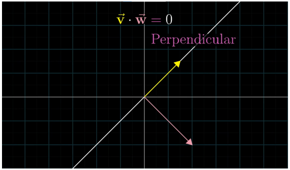
If $\mathbf{v}$ and $\mathbf{w}$ point in similar directions:

$$\mathbf{v} \bullet \mathbf{w} > 0$$



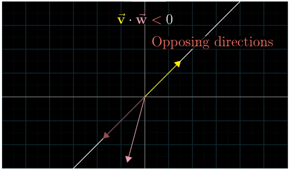
If $\mathbf{v}$ and $\mathbf{w}$ are orthogonal:

$$\mathbf{v} \bullet \mathbf{w} = 0$$



If $\mathbf{v}$ and $\mathbf{w}$ point in opposing directions:

$$\mathbf{v} \bullet \mathbf{w} < 0$$



Application: the dot product

Example

The *theory of multiple intelligences* proposes the differentiation of human intelligence into specific intelligences, rather than defining intelligence as a single, general ability. Two of these differentiated aspects are *interpersonal* intelligence and *logical–mathematical* intelligence.

Consider an intelligence test which aims to measure these two different aspects. Suppose that the scores are reported as (Z_i, Z_ℓ), where the values have been standardized so that each variable follows a standard normal distribution. Assume moreover that the two variables are **independent**.

Two universities use this test as admission criterion. However, they value each aspect of intelligence differently:

- University 1 considers the score: $V = 0.2Z_i + 0.8Z_\ell$
- University 2 considers the score: $W = 0.6Z_i + 0.4Z_\ell$

What is the correlation between the variables V and W?

Application: Analysis

Actually, we're going to focus instead on the covariance.

Covariance formulas

Let X, Y, Z be random variables, and a, b constants.

- $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$
- $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$
- $\operatorname{Cov}(X + Y, Z) = \operatorname{Cov}(X, Z) + \operatorname{Cov}(Y, Z)$
- $\operatorname{Cov}(aX, bY) = ab\operatorname{Cov}(X, Y)$

Using the above formulas, we see that:

$$\begin{aligned}
\operatorname{Cov}(V, W) &= \operatorname{Cov}(0.2Z_i + 0.8Z_\ell,\ 0.6Z_i + 0.4Z_\ell) \\
&= (0.2)(0.6)\operatorname{Var}(Z_i) + [(0.2)(0.4) + (0.8)(0.6)]\operatorname{Cov}(Z_i, Z_\ell) \\
&\quad + (0.8)(0.4)\operatorname{Var}(Z_\ell).
\end{aligned}$$

Becuase Z_i, Z_ℓ are standard normal and independent: $\operatorname{Var}(Z_i) = \operatorname{Var}(Z_\ell) = 1$ and $\operatorname{Cov}(Z_i, Z_\ell) = 0$.

Conclusion: $\operatorname{Cov}(V, W) = (0.2)(0.6) + (0.8)(0.4)$.

The dot product and covariance

If you are very comfortable with (mathematical) statistics, but less so with geometry, the following might be helpful.

You could **imagine** that the dot product is some sort of *covariance operator* between vectors.

In that case,

- $\mathbf{v} \cdot \mathbf{v}$ is analogous to the *variance*.
- Then $\|\mathbf{v}\| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$ plays the role of the *standard deviation*.
- The number $\cos\theta = \dfrac{\mathbf{v} \cdot \mathbf{w}}{\|\mathbf{v}\| \, \|\mathbf{w}\|}$ is a sort of *correlation*.
    - Indeed, $\cos\theta$ is always between -1 and 1,
    - $\cos\theta = 1$ when the vectors are pointing in exactly the same direction,
    - $\cos\theta = -1$ when the vectors are pointing in exactly opposite directions,
    - $\cos\theta = 0$ when the vectors are orthogonal (c.f. uncorrelated).

The dot product and covariance

In the following example, the dot product is literally the covariance.

Given two independent standard normal variables: Z_1, Z_2, the *space* of all possible linear combinations of them is identified with $\mathbb{R}^2$ via:

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \longleftrightarrow \quad V = v_1 Z_1 + v_2 Z_2.$$

In this vector space, the length and the dot product are given by:

$$\|\mathbf{v}\| = \sqrt{v_1^2 + v_2^2} = \sqrt{\mathrm{Var}(v_1 Z_1 + v_2 Z_2)} = \sigma_V,$$
$$\mathbf{v} \bullet \mathbf{w} = v_1 w_1 + v_2 w_2 = \mathrm{Cov}(v_1 Z_1 + v_2 Z_2,\, w_1 Z_1 + w_2 Z_2) = \mathrm{Cov}(V, W).$$

From Slide 10 we can see that:

- $\mathbf{v} \bullet \mathbf{w} = \mathrm{Cov}(V, W) > 0$, when $\mathbf{v}, \mathbf{w}$ point in a similar directions,
- $\mathbf{v} \bullet \mathbf{w} = \mathrm{Cov}(V, W) < 0$, when $\mathbf{v}, \mathbf{w}$ point in opposing directions,
- $\mathbf{v} \bullet \mathbf{w} = \mathrm{Cov}(V, W) = 0$, when $\mathbf{v}, \mathbf{w}$ are orthogonal.

Important remark: In statistics, the notion of orthogonality is usually equivalent to the notion of being uncorrelated.

Euclidean geometry

Definition

We call $\mathbb{R}^n$ the "**Euclidean space** of dimension n" when we regard it as a vector space with the geometry provided by the dot product.

Remark

Notice how the dot product really provides $\mathbb{R}^n$ with a geometric structure:

With the dot product we can:

- Measure length: $\|\mathbf{v}\| = \sqrt{\mathbf{v} \bullet \mathbf{v}}$,
- Measure angles: $\cos\theta = \dfrac{\mathbf{v} \bullet \mathbf{w}}{\|\mathbf{v}\|\,\|\mathbf{w}\|}$.

Remark

The reason why we are giving the good–old $\mathbb{R}^n$ a fancy name is because it is possible to consider aternative geometric structures on it.

Section 2

Orthogonal projections

Notation

Remark

From now on, I will always write $\mathbf{v}^\top \mathbf{w}$ instead of $\mathbf{v} \bullet \mathbf{w}$.

This is to emphasize that projections and transposition are deeply related.

Projections

The concept of an (orthogonal) projection is fundamental in statistics, mathematics and any other science.

Definition

Let $U \subset \mathbb{R}^n$ be a linear subspace of $\mathbb{R}^n$, and $\mathbf{x} \in \mathbb{R}^n$.

The **orthogonal projection** of $\mathbf{x}$ onto U is the (unique) vector $\mathbf{p}$ characterized by the fact that $\mathbf{x}$ can be decomposed as

$$\mathbf{x} = \mathbf{p} + (\mathbf{x} - \mathbf{p}),$$

where:

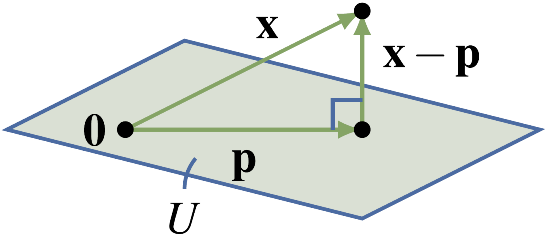
- $\mathbf{p} \in U$,
- $\mathbf{x} - \mathbf{p}$ is orthogonal to *every* vector in U.

The projection of $\mathbf{x}$ onto U is denoted as $\mathbf{p} = \operatorname{proj}_U \mathbf{x}$.

Projections

The projection $\mathbf{p} = \mathrm{proj}_U \mathbf{x}$ answers the question:

Among the vectors in U, which one is the one that resembles $\mathbf{x}$ the most?



Application: Projection

Example

Fertilizers are usually labeled with three numbers indicating the relative content in percentage by weight of the primary macronutrients: nitrogen (N), phosphorus (P), and potassium (K). A fertilizer manufacturer produces three different mixes with different NPK labels, as shown below.

	Mix 1	Mix 2	Mix 3
N	8%	10%	5%
P	4%	4%	5%
K	6%	4%	10%

For years, a farmer has been adding the following combination:

$$200 \text{ kg of Mix 1}, \quad 200 \text{ kg of Mix 2}, \quad 100 \text{ kg of Mix 3}.$$

Unfortunately, the company will stop producing Mix 3.

Is it possible to find a combination of mixes 1 & 2 that provides exactly the same amount of nutrients?

Application: Analysis

Let us first figure out how many kilos of each N, P, K, the farmer is used to.

```
> M <- matrix(c(0.08,0.04,0.06, 0.10,0.04,0.04, 0.05,0.05,0.10),
    nrow=3)  # matrix with all 3 mixes
> input <- c(200,200,100)  # quantities usually added by farmer
> x <- M %*% input; x  # weight per macronutrient
     [,1]
[1,]   41
[2,]   21
[3,]   30
```

Can we combine Mix 1 & 2 to get $41\ \mathrm{kg}$ N, $21\ \mathrm{kg}$ P, and $30\ \mathrm{kg}$ K?

Let's set a system of equations and solve it.

```
> A <- M[,1:2]  # matrix with only first 2 columns

# We want to solve for a in the system Aa=x
> gaussianElimination(A, x)
     [,1] [,2]        [,3]
[1,]    1    0 242.8571429
[2,]    0    1  10.7142857
[3,]    0    0   0.3571429
```

The last row means that the system is inconsistent!

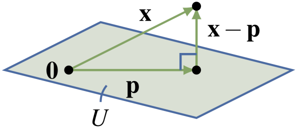
Application: Analysis

It is **impossible** to combine $\begin{bmatrix} 0.08 \\ 0.04 \\ 0.06 \end{bmatrix}$ and $\begin{bmatrix} 0.10 \\ 0.04 \\ 0.04 \end{bmatrix}$ in order to get $\begin{bmatrix} 41 \\ 21 \\ 30 \end{bmatrix}$.

The farmer still wants to fertilize the land using mixes 1 & 2. The NPK combination 41-21-30 is impossible.

But what is the closest we can get to 41-21-30?

- The best choice is the vector we get by projecting $\begin{bmatrix} 41 \\ 21 \\ 30 \end{bmatrix}$ onto the plane spanned by $\begin{bmatrix} 0.08 \\ 0.04 \\ 0.06 \end{bmatrix}$ and $\begin{bmatrix} 0.10 \\ 0.04 \\ 0.04 \end{bmatrix}$.



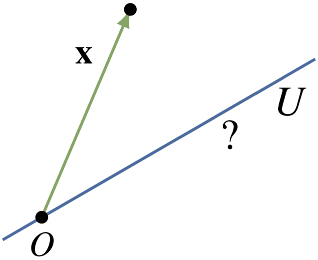
Projections minimize distance

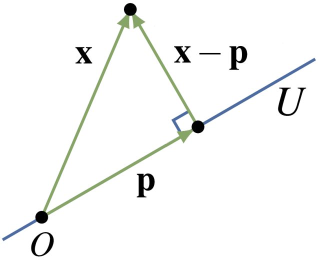
For each vector in $\mathbf{u} \in U$, consider the distance between $\mathbf{x}$ and $\mathbf{u}$: $\|\mathbf{x} - \mathbf{u}\|$.

The projection has the following fundamental property.

Property

The vector $\mathbf{p} = \operatorname{proj}_U \mathbf{x}$ is the vector in U whose distance to $\mathbf{x}$ is **minimal**.





Projection to a line

Consider the case of a line U (i.e. a one–dimensional subspace).

Suppose the line is given as: $U = \operatorname{span}(\mathbf{u})$, where $\mathbf{u}$ is a *unit vector*.

Definition

A vector $\mathbf{u} \in \mathbb{R}^n$ is called a **unit vector** if $\|\mathbf{u}\| = 1$.

Then $\operatorname{proj}_U \mathbf{x}$ must be a multiple of $\mathbf{u}$: $\operatorname{proj}_U \mathbf{x} = \alpha \mathbf{u}$.

The question is: what is the right α?

Fact

The projection of $\mathbf{x}$ onto (the span of) a unit vector $\mathbf{u}$ is given by:

$$\operatorname{proj}_{\mathbf{u}} \mathbf{x} = (\mathbf{u}^\top \mathbf{x})\, \mathbf{u}.$$

Projection to a line

Suppose now that $U = \operatorname{span}(\mathbf{v})$, where $\mathbf{v}$ is a vector of arbitrary length.

We are looking for α such that $\operatorname{proj}_{\mathbf{v}} \mathbf{x} = \alpha\mathbf{v}$.

The fundamental fact is: $\mathbf{v}^\top(\mathbf{x} - \alpha\mathbf{v}) = 0$.

The above equation can be rewritten as: $\mathbf{v}^\top\mathbf{x} = \alpha\mathbf{v}^\top\mathbf{v}$.

Solving for α gives: $\alpha = \dfrac{\mathbf{v}^\top\mathbf{x}}{\mathbf{v}^\top\mathbf{v}}$.

Theorem

The projection of $\mathbf{x}$ onto (the span of) a vector $\mathbf{v} \neq \mathbf{0}$ is given by:

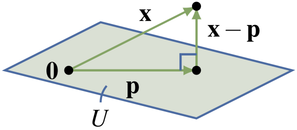
$$\operatorname{proj}_{\mathbf{v}} \mathbf{x} = \frac{\mathbf{v}^\top\mathbf{x}}{\mathbf{v}^\top\mathbf{v}}\,\mathbf{v}.$$

Notice that the denominator can be rewritten using the identity $\mathbf{v}^\top\mathbf{v} = \|\mathbf{v}\|^2$.

Projection to a subspace

We now consider $U \subset \mathbb{R}^n$ a linear subspace of dimension k.

Suppose U has basis $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\}$, and take a vector $\mathbf{x} \notin U$.



We wish to decompose: $\mathbf{x} = \mathbf{p} + (\mathbf{x} - \mathbf{p})$, such that

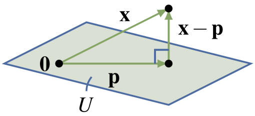
- $\mathbf{p} \in U$,
- $\mathbf{x} - \mathbf{p} \in U^{\perp}$ (i.e. is orthogonal to all vectors in U).

In particular, $\mathbf{x} - \mathbf{p}$ is orthogonal to each of the basis vectors $\mathbf{v}_i$.

Projection to a subspace

So, we have $\mathbf{v}_i^\top(\mathbf{x}-\mathbf{p}) = 0$, for $i = 1, 2, \ldots, k$. Or, in matrix notation:

$$\begin{bmatrix} \mathbf{v}_1^\top \\ \mathbf{v}_2^\top \\ \vdots \\ \mathbf{v}_k^\top \end{bmatrix} (\mathbf{x}-\mathbf{p}) = \begin{bmatrix} \mathbf{v}_1^\top(\mathbf{x}-\mathbf{p}) \\ \mathbf{v}_2^\top(\mathbf{x}-\mathbf{p}) \\ \vdots \\ \mathbf{v}_k^\top(\mathbf{x}-\mathbf{p}) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$



If we let $A = \begin{bmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_k \end{bmatrix}$, we get:

$$\begin{bmatrix} \mathbf{v}_1^\top \\ \vdots \\ \mathbf{v}_k^\top \end{bmatrix} (\mathbf{x}-\mathbf{p}) = \mathbf{0} \quad \Rightarrow \quad A^\top(\mathbf{x}-\mathbf{p}) = \mathbf{0}.$$

Because $\mathbf{p} \in U$, we can write $\mathbf{p} = \alpha_1\mathbf{v}_1 + \ldots + \alpha_k\mathbf{v}_k = A\boldsymbol{\alpha}$.

From: $A^\top(\mathbf{x} - A\boldsymbol{\alpha}) = \mathbf{0}$, we arrive to:

The **normal equation**: $\boxed{A^\top A\boldsymbol{\alpha} = A^\top\mathbf{x}}$

Projection to a subspace

Fact

If the columns of A are linearly independent, then $A^\top A$ is **invertible**.

Using the above fact, we can conclude:

$$A^\top A\boldsymbol{\alpha} = A^\top \mathbf{x},$$

$$\boldsymbol{\alpha} = (A^\top A)^{-1}A^\top \mathbf{x},$$

$$\mathbf{p} = A\boldsymbol{\alpha} = A(A^\top A)^{-1}A^\top \mathbf{x}.$$

By the way, we also get: $\mathbf{x} - \mathbf{p} = \left(I_n - A(A^\top A)^{-1}A^\top\right)\mathbf{x}$.

Exercise:

Check that $\mathbf{x} - \mathbf{p}$ (given by the above formula) is orthogonal to every column of A by verifying that $A^\top(\mathbf{x} - \mathbf{p}) = \mathbf{0}$.

Example: projection onto a plane

Example

The plane U in $\mathbb{R}^3$ is spanned by $\mathbf{v}_1 = \begin{bmatrix}1\\1\\0\end{bmatrix}$ and $\mathbf{v}_2 = \begin{bmatrix}1\\0\\-1\end{bmatrix}$.

What is the projection of $\mathbf{x} = \begin{bmatrix}6\\2\\-1\end{bmatrix}$ onto U?

Solution:

Set $A = \begin{bmatrix}\mathbf{v}_1 & \mathbf{v}_2\end{bmatrix}$. The projection is $\mathbf{p} = A\boldsymbol{\alpha} = \begin{bmatrix}1 & 1\\1 & 0\\0 & -1\end{bmatrix}\begin{bmatrix}\alpha_1\\\alpha_2\end{bmatrix}$, for some vector of coefficients $\boldsymbol{\alpha}$ that we need to find.

The *normal equation*, $A^\top A\boldsymbol{\alpha} = A^\top\mathbf{x}$, becomes:

$$\begin{bmatrix}1 & 1 & 0\\1 & 0 & -1\end{bmatrix}\begin{bmatrix}1 & 1\\1 & 0\\0 & -1\end{bmatrix}\begin{bmatrix}\alpha_1\\\alpha_2\end{bmatrix} = \begin{bmatrix}1 & 1 & 0\\1 & 0 & -1\end{bmatrix}\begin{bmatrix}6\\2\\-1\end{bmatrix}.$$

Example: projection onto a plane (continued)

After multiplying the matrices we get:

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 7 \end{bmatrix}$$

We need to solve this equation for α_1, α_2.

The augmented matrix $\left[\begin{array}{cc|c} 2 & 1 & 8 \\ 1 & 2 & 7 \end{array}\right]$ reduces to $\left[\begin{array}{cc|c} 1 & 0 & 3 \\ 0 & 1 & 2 \end{array}\right]$, therefore $\alpha_1 = 3$ and $\alpha_2 = 2$.

Thus, $\mathbf{p} = 3 \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix}$, and $\mathbf{x} - \mathbf{p} = \begin{bmatrix} 6 \\ 2 \\ -1 \end{bmatrix} - \begin{bmatrix} 5 \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$.

We check our answer by verifying that:

$$\begin{bmatrix} 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 0 \quad \text{and} \quad \begin{bmatrix} 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix} = 0.$$

The orthogonal complement

Definition

Let $U \subset \mathbb{R}^n$ be a linear subspace.

The **orthogonal complement** of U is the linear subspace:

$$U^\perp = \{\mathbf{v} \in \mathbb{R}^n \,|\, \mathbf{v}^\top \mathbf{u} = 0 \text{ for all } \mathbf{u} \in U\}.$$

Let $\mathbf{x} \in \mathbb{R}^n$, and let $\mathbf{p} = \mathrm{proj}_U \mathbf{x}$. Then:

$$\mathbf{x} = \mathbf{p} + (\mathbf{x} - \mathbf{p}), \quad \text{where} \quad \mathbf{p} \in U \text{ and } (\mathbf{x} - \mathbf{p}) \in U^\perp.$$

Fact

Every vector in $\mathbf{x} \in \mathbb{R}^n$ can be expressed as the sum of two vectors: $\mathbf{x} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in U$ and $\mathbf{v} \in U^\perp$.

Moreover, there is a *unique* way to decompose $\mathbf{x}$ like this.

In symbols, we can write: $\mathbb{R}^n = U \oplus U^\perp$ to denote this fact.

Mean vector and centering matrix

Example

Consider the one-vector: $\mathbf{1}_n \in \mathbb{R}^n$. Let $M = \operatorname{span}\{\mathbf{1}_n\}$.

What is the orthogonal complment of M?

What is the interpretation of decomposing a vector $\mathbf{x}$ into its M and $M^\perp$ parts?

Mean vector and centering matrix

We are looking for the subspace consisting of vectors $\mathbf{x}$ such that $\mathbf{1}_n^\top \mathbf{x} = 0$.

This means:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_{i=1}^{n} x_i = 0.$$

Notice how $\sum x_i = 0$ if and only if $\bar{x} = \frac{1}{n} \sum x_i = 0$.

Fact

The orthogonal complement of $\operatorname{span}\{\mathbf{1}_n\}$ is the space of zero-mean vectors:

$$M^\perp = \{\mathbf{x} \in \mathbb{R}^n \mid \bar{x} = 0\}.$$

Mean vector and centering matrix

We now project an arbitrary vector onto M:

$$\begin{aligned} \operatorname{proj}_{\mathbf{1}_n} \mathbf{x} &= \frac{\mathbf{1}_n^\top \mathbf{x}}{\|\mathbf{1}_n\|^2}\, \mathbf{1}_n \\ &= \frac{\sum x_i}{n}\, \mathbf{1}_n \\ &= \bar{x}\, \mathbf{1}_n \end{aligned}$$

Conclusion:

Projection of $\mathbf{x}$ onto $\mathbf{1}_n$ computes the mean of $\mathbf{x}$, and returns $\bar{x}\mathbf{1}_n$.

Mean vector and centering matrix

We now project an arbitrary vector onto $M^\perp$.

Fact

$\mathbf{x} = \operatorname{proj}_M \mathbf{x} + \operatorname{proj}_{M^\perp} \mathbf{x}$.

$$\begin{aligned}
\operatorname{proj}_{M^\perp} \mathbf{x} &= \mathbf{x} - \operatorname{proj}_M \mathbf{x} \\
&= \mathbf{x} - \frac{\mathbf{1}_n^\top \mathbf{x}}{\|\mathbf{1}_n\|^2} \mathbf{1}_n \\
&= \mathbf{x} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \mathbf{x} \\
&= \left(I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top \right) \mathbf{x} \\
&= C_n \mathbf{x}
\end{aligned}$$

Mean vector and centering matrix

Definition

The matrix on the previous slide,

$$C_n = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top,$$

is called the (n–dimensional) **centering matrix**.

It represents the *centering* transformation:

$$C_n \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}.$$

Section 3

Supplementary material

Application: Simple linear regression

Consider n observations on two variables, x and y. We would like to fit a simple linear regression to explain y in terms of x. As usual, we arrange these data into an $n \times 2$ matrix M.

We look for the parameters β_0, β_1 that minimize the sum of squares:

$$S = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

We are used to visualizing the data in a scatter plot; namely, by regarding M as a collection of n vectors in $\mathbb{R}^2$. However, it is useful to also think of M as a collection of two vectors in $\mathbb{R}^n$.

When thinking about n–dimensional vectors, we can recognize S as $S = \|\mathbf{e}\|^2$,

where $\mathbf{e}$ is the residual vector: $\mathbf{e} = \begin{bmatrix} y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \beta_0 - \beta_1 x_2 \\ \vdots \\ y_n - \beta_0 - \beta_1 x_n \end{bmatrix}.$

Application: Simple linear regression

When fitting parameters β_0, β_1, we are trying to achieve:

$$y_1 \approx \beta_0 + \beta_1 x_1, \qquad y_2 \approx \beta_0 + \beta_1 x_2, \qquad \dots \qquad y_n \approx \beta_0 + \beta_1 x_n.$$

Let's now not think about individual observations, but about n–vectors. We are looking for β_0, β_1 such that:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \approx \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

Summarizing: We have a vector $\mathbf{y}$, and we want to find which vector in $\mathrm{span}\{\mathbf{1}_n, \mathbf{x}\}$ resembles $\mathbf{y}$ the most.

The next slide shows that indeed the least-squares fit corresponds to an *orthogonal* projection.

Application: Simple linear regression

Set $X = \begin{bmatrix} \mathbf{1}_n & \mathbf{x} \end{bmatrix}$. Then the residual vector can be re-written as:

$$\mathbf{e} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} - \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \mathbf{y} - X\boldsymbol{\beta}.$$

Conclusion 1

The problem of fitting a linear regression boils down to finding a linear combination $X\boldsymbol{\beta}$ of the vectors $\mathbf{1}_n$ and $\mathbf{x}$ which minimizes $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$.

Geometric insight: $\|\mathbf{y} - X\boldsymbol{\beta}\|^2$ is minimal when we have an orthogonal projection.

Conclusion 2

The problem of fitting a linear regression is equivalent to finding the **orthogonal projection** of the vector $\mathbf{y}$ onto the linear subspace $\operatorname{span}\{\mathbf{1}_n, \mathbf{x}\}$. Therefore,

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{y}.$$

Application: Simple linear regression

The geometric interpretation of linear models is also discussed in John Fox's book, in Chapter 10.

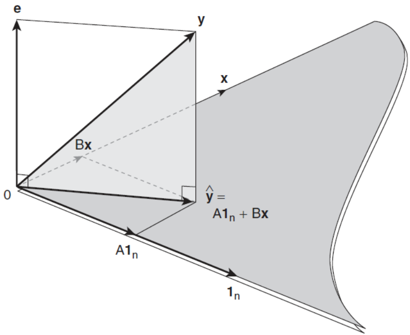


Figure 10.2 The vector geometry of least-squares fit in simple regression. Minimizing the residual sum of squares is equivalent to making the **e** vector as short as possible. The $\widehat{\mathbf{y}}$ vector is, therefore, the orthogonal projection of **y** onto the $\{\mathbf{1_n}, \mathbf{x}\}$ plane.

Source: *John Fox. Applied Regression Analysis and Generalized Linear Models*, 3^{rd} *ed.*

Application: Multiple linear regression

Consider a ***multiple*** linear regression model with k explanatory variables:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \varepsilon.$$

Suppose we have n observations of the variables: $\mathbf{x}_1, \ldots, \mathbf{x}_k, \mathbf{y} \in \mathbb{R}^n$.

Let $X = \begin{bmatrix} \mathbf{1}_n & \mathbf{x}_1 & \ldots & \mathbf{x}_k \end{bmatrix}$, and $\mathbf{e} = \mathbf{y} - X\boldsymbol{\beta}$.

Ordinary least squares

The method of *ordinary least squares* for fitting a linear regression provides parameter estimates $\hat{\beta}$ for which the squared-length of the residual vector $\mathbf{e}$ is minimal. It corresponds to projecting $\mathbf{y}$ onto $\operatorname{span}\{\mathbf{1}_n, \mathbf{x}_1, \ldots, \mathbf{x}_k\}$.

The estimates are given by:

$$\hat{\boldsymbol{\beta}} = \left(X^\top X\right)^{-1} X^\top \mathbf{y}.$$

Note

In matrix notation, the formula for the estimates is always the same (as long as the columns of X remain independent!). The only thing that changes is the **design matrix** X.

Idempotent matrices

Definition

A square matrix P is called **idempotent** if it satisfies: $P^2 = P$.

These matrices appear often in statistics. Below some important properties.

Properties

Let P be an idempotent matrix.

- If λ is an eigenvalue of P, then either $\lambda = 0$ or $\lambda = 1$.
- P is a diagonalizable matrix.
- If P is symmetric, then P represents an *orthogonal* projection.

Idempotent matrices

The last point on the previous slides deserves extra attention.

Suppose P is an idempotent $n \times n$ matrix. Let $U = \mathrm{im}(T_P)$. Recall that

$$\mathrm{im}(T_P) = \{\mathbf{y} \mid \mathbf{y} = P\mathbf{x} \text{ for some } \mathbf{x}\}.$$

If $\mathbf{y} \in U$, then $P\mathbf{y} = P(P\mathbf{x}) = P^2\mathbf{x} = P\mathbf{x} = \mathbf{y}$.

Therefore, the linear transformation T_P has two important properties:

- It squishes all of $\mathbb{R}^n$ onto the subspace U,
- If $\mathbf{y}$ is already in U then $T_P(\mathbf{y}) = \mathbf{y}$.

If you think about it, it sounds a lot like the transformation "projection onto U". This is somewhat true, but if you want an orthogonal projection, then P must be symmetric.

Idempotent matrices

Fact

If P is symmetric and idempotent, then T_P is the linear transformation $T_P(\mathbf{x}) = \operatorname{proj}_U \mathbf{x}$, where $U = \operatorname{im}(T_P)$.

If you look at Slide 28, you'll see that we computed that

$$\operatorname{proj}_U \mathbf{x} = A(A^\top A)^{-1}A^\top \mathbf{x}.$$

You can see that the projection is thus obtained by multiplying $\mathbf{x}$ with the matrix $P = A(A^\top A)^{-1}A^\top$.

Exercise:

Verify for yourself that the matrix $P = A(A^\top A)^{-1}A^\top$ satisfies:

- $P = P^2$,
- $P = P^\top$.