

Orthogonal Projections and Their Applications

Jesse Perla, Thomas J. Sargent and John Stachurski

December 4, 2020

1 Contents

- Overview [2](#)
- Key Definitions [3](#)
- The Orthogonal Projection Theorem [4](#)
- Orthonormal Basis [5](#)
- Projection Using Matrix Algebra [6](#)
- Least Squares Regression [7](#)
- Orthogonalization and Decomposition [8](#)
- Exercises [9](#)
- Solutions [10](#)

2 Overview

Orthogonal projection is a cornerstone of vector space methods, with many diverse applications.

These include, but are not limited to,

- Least squares projection, also known as linear regression
- Conditional expectations for multivariate normal (Gaussian) distributions
- Gram–Schmidt orthogonalization
- QR decomposition
- Orthogonal polynomials
- etc

In this lecture we focus on

- key ideas
- least squares regression

2.1 Further Reading

For background and foundational concepts, see our lecture [on linear algebra](#).

For more proofs and greater theoretical detail, see [A Primer in Econometric Theory](#).

For a complete set of proofs in a general setting, see, for example, [\[1\]](#).

For an advanced treatment of projection in the context of least squares prediction, see [this book chapter](#).

3 Key Definitions

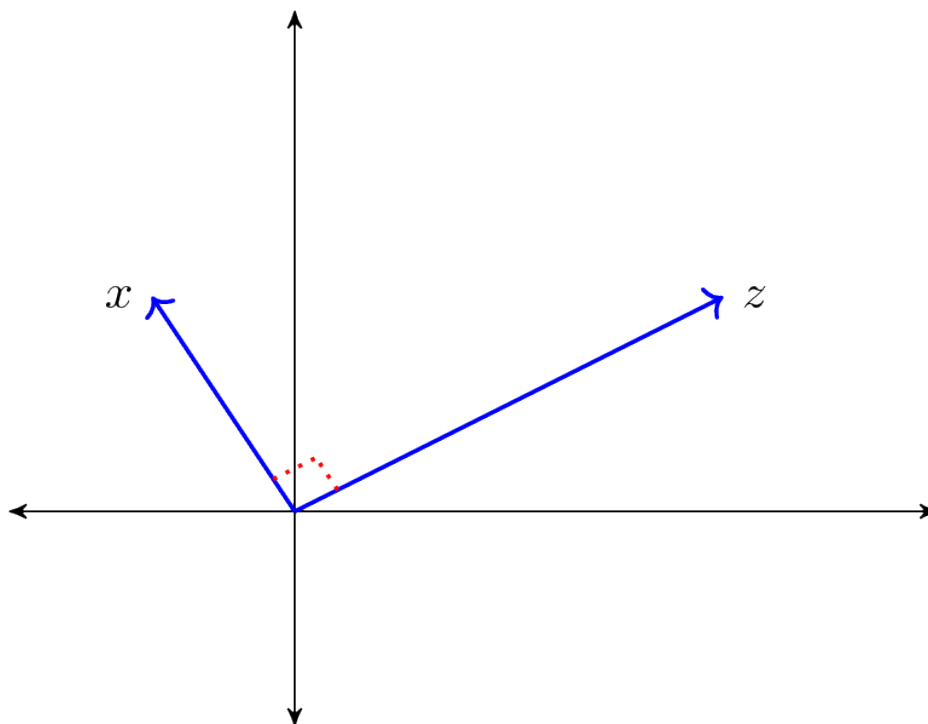
Assume $x, z \in \mathbb{R}^n$.

Define $\langle x, z \rangle = \sum_i x_i z_i$.

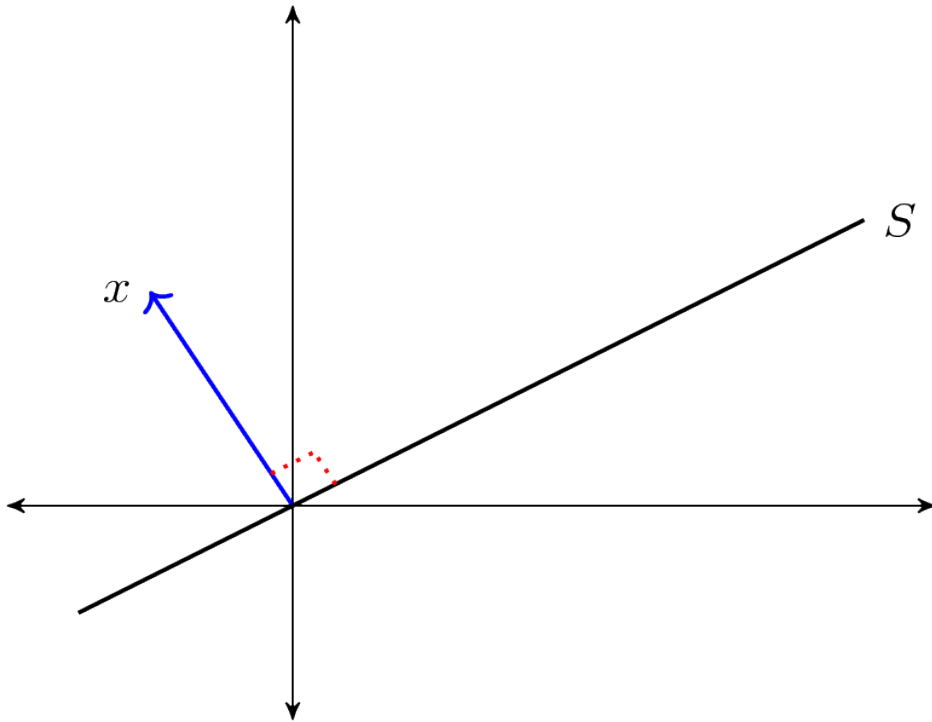
Recall $\|x\|^2 = \langle x, x \rangle$.

The **law of cosines** states that $\langle x, z \rangle = \|x\| \|z\| \cos(\theta)$ where θ is the angle between the vectors x and z .

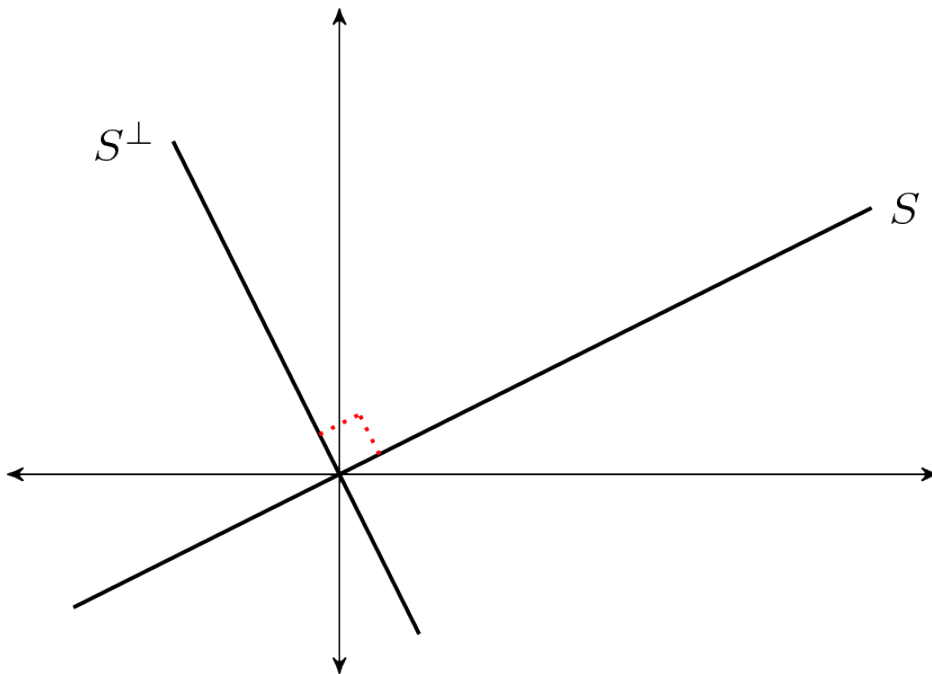
When $\langle x, z \rangle = 0$, then $\cos(\theta) = 0$ and x and z are said to be **orthogonal** and we write $x \perp z$



For a linear subspace $S \subset \mathbb{R}^n$, we call $x \in \mathbb{R}^n$ **orthogonal to S** if $x \perp z$ for all $z \in S$, and write $x \perp S$



The **orthogonal complement** of linear subspace $S \subset \mathbb{R}^n$ is the set $S^\perp := \{x \in \mathbb{R}^n : x \perp S\}$



S^\perp is a linear subspace of \mathbb{R}^n

- To see this, fix $x, y \in S^\perp$ and $\alpha, \beta \in \mathbb{R}$.
- Observe that if $z \in S$, then

$$\langle \alpha x + \beta y, z \rangle = \alpha \langle x, z \rangle + \beta \langle y, z \rangle = \alpha \times 0 + \beta \times 0 = 0$$

- Hence $\alpha x + \beta y \in S^\perp$, as was to be shown

A set of vectors $\{x_1, \dots, x_k\} \subset \mathbb{R}^n$ is called an **orthogonal set** if $x_i \perp x_j$ whenever $i \neq j$.

If $\{x_1, \dots, x_k\}$ is an orthogonal set, then the **Pythagorean Law** states that

$$\|x_1 + \dots + x_k\|^2 = \|x_1\|^2 + \dots + \|x_k\|^2$$

For example, when $k = 2$, $x_1 \perp x_2$ implies

$$\|x_1 + x_2\|^2 = \langle x_1 + x_2, x_1 + x_2 \rangle = \langle x_1, x_1 \rangle + 2\langle x_2, x_1 \rangle + \langle x_2, x_2 \rangle = \|x_1\|^2 + \|x_2\|^2$$

3.1 Linear Independence vs Orthogonality

If $X \subset \mathbb{R}^n$ is an orthogonal set and $0 \notin X$, then X is linearly independent.

Proving this is a nice exercise.

While the converse is not true, a kind of partial converse holds, as we'll [see below](#).

4 The Orthogonal Projection Theorem

What vector within a linear subspace of \mathbb{R}^n best approximates a given vector in \mathbb{R}^n ?

The next theorem provides answers this question.

Theorem (OPT) Given $y \in \mathbb{R}^n$ and linear subspace $S \subset \mathbb{R}^n$, there exists a unique solution to the minimization problem

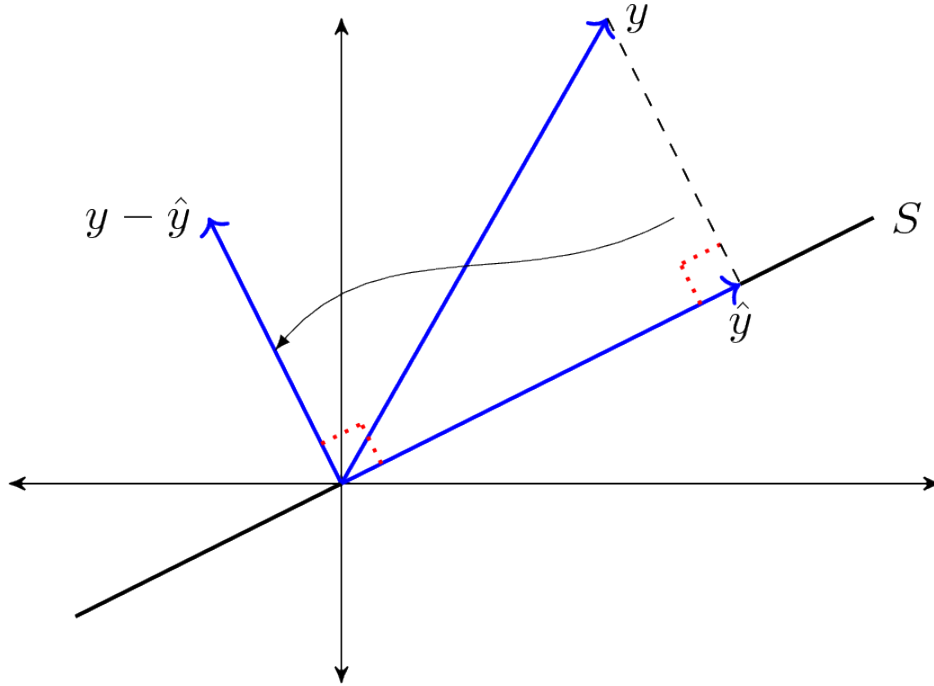
$$\hat{y} := \arg \min_{z \in S} \|y - z\|$$

The minimizer \hat{y} is the unique vector in \mathbb{R}^n that satisfies

- $\hat{y} \in S$
- $y - \hat{y} \perp S$

The vector \hat{y} is called the **orthogonal projection** of y onto S .

The next figure provides some intuition



4.1 Proof of sufficiency

We'll omit the full proof.

But we will prove sufficiency of the asserted conditions.

To this end, let $y \in \mathbb{R}^n$ and let S be a linear subspace of \mathbb{R}^n .

Let \hat{y} be a vector in \mathbb{R}^n such that $\hat{y} \in S$ and $y - \hat{y} \perp S$.

Let z be any other point in S and use the fact that S is a linear subspace to deduce

$$\|y - z\|^2 = \|(y - \hat{y}) + (\hat{y} - z)\|^2 = \|y - \hat{y}\|^2 + \|\hat{y} - z\|^2$$

Hence $\|y - z\| \geq \|y - \hat{y}\|$, which completes the proof.

4.2 Orthogonal Projection as a Mapping

For a linear space Y and a fixed linear subspace S , we have a functional relationship

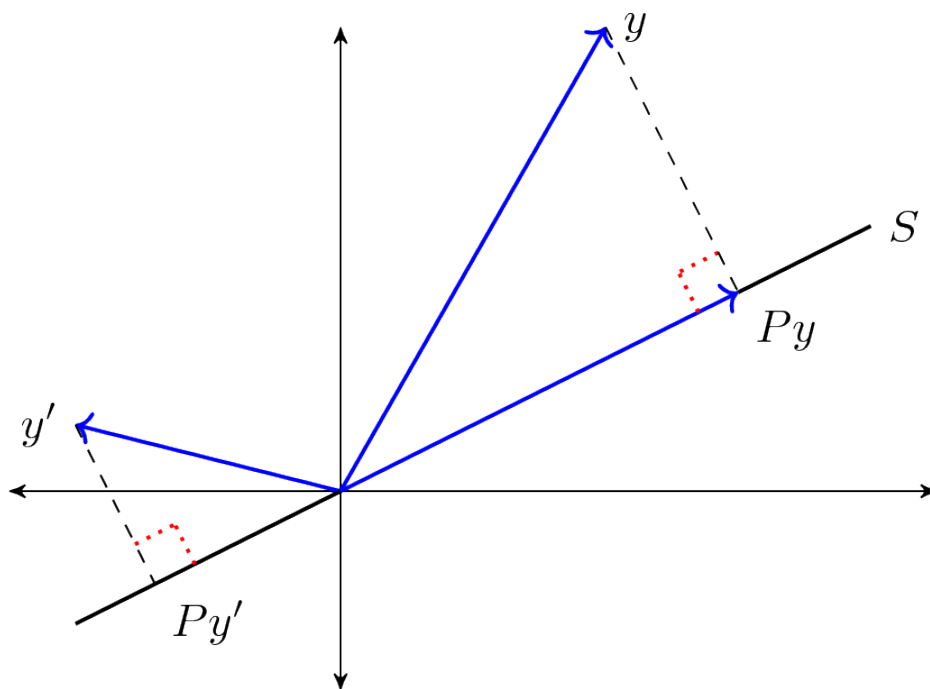
$$y \in Y \mapsto \text{its orthogonal projection } \hat{y} \in S$$

By the OPT, this is a well-defined mapping or *operator* from \mathbb{R}^n to \mathbb{R}^n .

In what follows we denote this operator by a matrix P

- Py represents the projection \hat{y} .
- This is sometimes expressed as $\hat{E}_S y = Py$, where \hat{E} denotes a **wide-sense expectations operator** and the subscript S indicates that we are projecting y onto the linear subspace S .

The operator P is called the **orthogonal projection mapping onto S**



It is immediate from the OPT that for any $y \in \mathbb{R}^n$

1. $Py \in S$ and
2. $y - Py \perp S$

From this we can deduce additional useful properties, such as

1. $\|y\|^2 = \|Py\|^2 + \|y - Py\|^2$ and
2. $\|Py\| \leq \|y\|$

For example, to prove 1, observe that $y = Py + y - Py$ and apply the Pythagorean law.

4.2.1 Orthogonal Complement

Let $S \subset \mathbb{R}^n$.

The **orthogonal complement** of S is the linear subspace S^\perp that satisfies $x_1 \perp x_2$ for every $x_1 \in S$ and $x_2 \in S^\perp$.

Let Y be a linear space with linear subspace S and its orthogonal complement S^\perp .

We write

$$Y = S \oplus S^\perp$$

to indicate that for every $y \in Y$ there is unique $x_1 \in S$ and a unique $x_2 \in S^\perp$ such that $y = x_1 + x_2$.

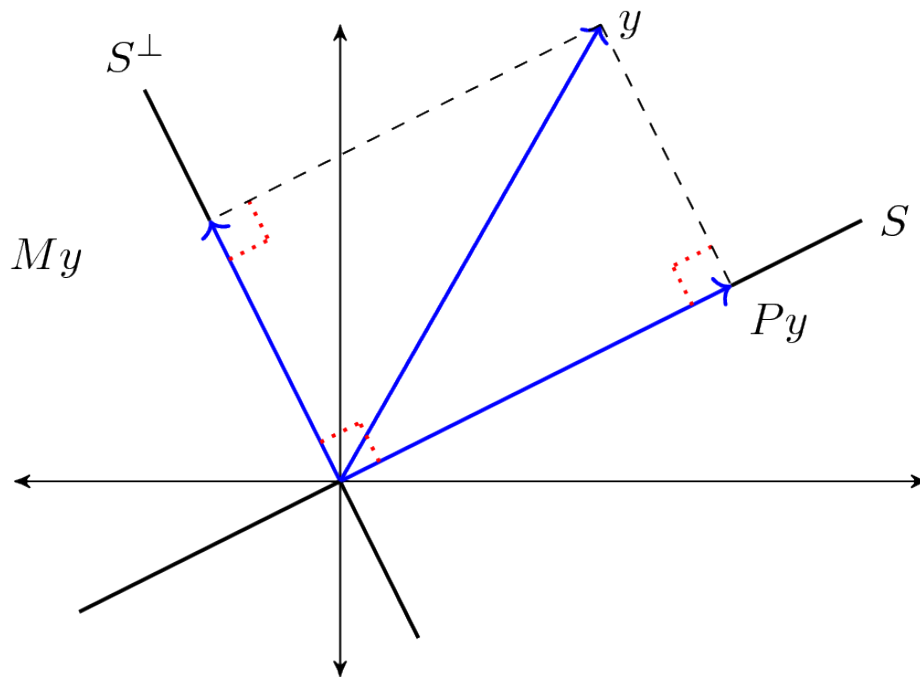
Moreover, $x_1 = \hat{E}_S y$ and $x_2 = y - \hat{E}_S y$.

This amounts to another version of the OPT:

Theorem. If S is a linear subspace of \mathbb{R}^n , $\hat{E}_S y = Py$ and $\hat{E}_{S^\perp} y = My$, then

$$Py \perp My \quad \text{and} \quad y = Py + My \quad \text{for all } y \in \mathbb{R}^n$$

The next figure illustrates



5 Orthonormal Basis

An orthogonal set of vectors $O \subset \mathbb{R}^n$ is called an **orthonormal set** if $\|u\| = 1$ for all $u \in O$.

Let S be a linear subspace of \mathbb{R}^n and let $O \subset S$.

If O is orthonormal and $\text{span } O = S$, then O is called an **orthonormal basis** of S .

O is necessarily a basis of S (being independent by orthogonality and the fact that no element is the zero vector).

One example of an orthonormal set is the canonical basis $\{e_1, \dots, e_n\}$ that forms an orthonormal basis of \mathbb{R}^n , where e_i is the i th unit vector.

If $\{u_1, \dots, u_k\}$ is an orthonormal basis of linear subspace S , then

$$x = \sum_{i=1}^k \langle x, u_i \rangle u_i \quad \text{for all } x \in S$$

To see this, observe that since $x \in \text{span}\{u_1, \dots, u_k\}$, we can find scalars $\alpha_1, \dots, \alpha_k$ that verify

$$x = \sum_{j=1}^k \alpha_j u_j \tag{1}$$

Taking the inner product with respect to u_i gives

$$\langle x, u_i \rangle = \sum_{j=1}^k \alpha_j \langle u_j, u_i \rangle = \alpha_i$$

Combining this result with (1) verifies the claim.

5.1 Projection onto an Orthonormal Basis

When the subspace onto which are projecting is orthonormal, computing the projection simplifies:

Theorem If $\{u_1, \dots, u_k\}$ is an orthonormal basis for S , then

$$Py = \sum_{i=1}^k \langle y, u_i \rangle u_i, \quad \forall y \in \mathbb{R}^n \tag{2}$$

Proof: Fix $y \in \mathbb{R}^n$ and let Py be defined as in (2).

Clearly, $Py \in S$.

We claim that $y - Py \perp S$ also holds.

It suffices to show that $y - Py \perp$ any basis vector u_i (why?).

This is true because

$$\left\langle y - \sum_{i=1}^k \langle y, u_i \rangle u_i, u_j \right\rangle = \langle y, u_j \rangle - \sum_{i=1}^k \langle y, u_i \rangle \langle u_i, u_j \rangle = 0$$

6 Projection Using Matrix Algebra

Let S be a linear subspace of \mathbb{R}^n and let $y \in \mathbb{R}^n$.

We want to compute the matrix P that verifies

$$\hat{E}_S y = Py$$

Evidently Py is a linear function from $y \in \mathbb{R}^n$ to $Py \in \mathbb{R}^n$.

This reference is useful https://en.wikipedia.org/wiki/Linear_map#Matrices.

Theorem. Let the columns of $n \times k$ matrix X form a basis of S . Then

$$P = X(X'X)^{-1}X'$$

Proof: Given arbitrary $y \in \mathbb{R}^n$ and $P = X(X'X)^{-1}X'$, our claim is that

1. $Py \in S$, and
2. $y - Py \perp S$

Claim 1 is true because

$$Py = X(X'X)^{-1}X'y = Xa \quad \text{when} \quad a := (X'X)^{-1}X'y$$

An expression of the form Xa is precisely a linear combination of the columns of X , and hence an element of S .

Claim 2 is equivalent to the statement

$$y - X(X'X)^{-1}X'y \perp Xb \quad \text{for all} \quad b \in \mathbb{R}^K$$

This is true: If $b \in \mathbb{R}^K$, then

$$(Xb)'[y - X(X'X)^{-1}X'y] = b'[X'y - X'y] = 0$$

The proof is now complete.

6.1 Starting with X

It is common in applications to start with $n \times k$ matrix X with linearly independent columns and let

$$S := \text{span } X := \text{span}\{\text{col } X_1, \dots, \text{col } X_k\}$$

Then the columns of X form a basis of S .

From the preceding theorem, $P = X(X'X)^{-1}X'y$ projects y onto S .

In this context, P is often called the **projection matrix**.

- The matrix $M = I - P$ satisfies $My = \hat{E}_{S^\perp}y$ and is sometimes called the **annihilator matrix**.

6.2 The Orthonormal Case

Suppose that U is $n \times k$ with orthonormal columns.

Let $u_i := \text{col } U_i$ for each i , let $S := \text{span } U$ and let $y \in \mathbb{R}^n$.

We know that the projection of y onto S is

$$Py = U(U'U)^{-1}U'y$$

Since U has orthonormal columns, we have $U'U = I$.

Hence

$$Py = UU'y = \sum_{i=1}^k \langle u_i, y \rangle u_i$$

We have recovered our earlier result about projecting onto the span of an orthonormal basis.

6.3 Application: Overdetermined Systems of Equations

Let $y \in \mathbb{R}^n$ and let X is $n \times k$ with linearly independent columns.

Given X and y , we seek $b \in \mathbb{R}^k$ satisfying the system of linear equations $Xb = y$.

If $n > k$ (more equations than unknowns), then b is said to be **overdetermined**.

Intuitively, we may not be able find a b that satisfies all n equations.

The best approach here is to

- Accept that an exact solution may not exist
- Look instead for an approximate solution

By approximate solution, we mean a $b \in \mathbb{R}^k$ such that Xb is as close to y as possible.

The next theorem shows that the solution is well defined and unique.

The proof uses the OPT.

Theorem The unique minimizer of $\|y - Xb\|$ over $b \in \mathbb{R}^K$ is

$$\hat{\beta} := (X'X)^{-1}X'y$$

Proof: Note that

$$X\hat{\beta} = X(X'X)^{-1}X'y = Py$$

Since Py is the orthogonal projection onto $\text{span}(X)$ we have

$$\|y - Py\| \leq \|y - z\| \text{ for any } z \in \text{span}(X)$$

Because $Xb \in \text{span}(X)$

$$\|y - X\hat{\beta}\| \leq \|y - Xb\| \text{ for any } b \in \mathbb{R}^K$$

This is what we aimed to show.

7 Least Squares Regression

Let's apply the theory of orthogonal projection to least squares regression.

This approach provides insights about many geometric properties of linear regression.

We treat only some examples.

7.1 Squared risk measures

Given pairs $(x, y) \in \mathbb{R}^K \times \mathbb{R}$, consider choosing $f: \mathbb{R}^K \rightarrow \mathbb{R}$ to minimize the **risk**

$$R(f) := \mathbb{E}[(y - f(x))^2]$$

If probabilities and hence \mathbb{E} are unknown, we cannot solve this problem directly.

However, if a sample is available, we can estimate the risk with the **empirical risk**:

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^N (y_n - f(x_n))^2$$

Minimizing this expression is called **empirical risk minimization**.

The set \mathcal{F} is sometimes called the hypothesis space.

The theory of statistical learning tells us that to prevent overfitting we should take the set \mathcal{F} to be relatively simple.

If we let \mathcal{F} be the class of linear functions $1/N$, the problem is

$$\min_{b \in \mathbb{R}^K} \sum_{n=1}^N (y_n - b'x_n)^2$$

This is the sample **linear least squares problem**.

7.2 Solution

Define the matrices

$$y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad x_n := \begin{pmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{nK} \end{pmatrix} = \text{n-th obs on all regressors}$$

and

$$X := \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_N \end{pmatrix} ::= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}$$

We assume throughout that $N > K$ and X is full column rank.

If you work through the algebra, you will be able to verify that $\|y - Xb\|^2 = \sum_{n=1}^N (y_n - b'x_n)^2$.

Since monotone transforms don't affect minimizers, we have

$$\arg \min_{b \in \mathbb{R}^K} \sum_{n=1}^N (y_n - b'x_n)^2 = \arg \min_{b \in \mathbb{R}^K} \|y - Xb\|$$

By our results about overdetermined linear systems of equations, the solution is

$$\hat{\beta} := (X'X)^{-1}X'y$$

Let P and M be the projection and annihilator associated with X :

$$P := X(X'X)^{-1}X' \quad \text{and} \quad M := I - P$$

The **vector of fitted values** is

$$\hat{y} := X\hat{\beta} = Py$$

The **vector of residuals** is

$$\hat{u} := y - \hat{y} = y - Py = My$$

Here are some more standard definitions:

- The **total sum of squares** is $:= \|y\|^2$.
- The **sum of squared residuals** is $:= \|\hat{u}\|^2$.
- The **explained sum of squares** is $:= \|\hat{y}\|^2$.

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

We can prove this easily using the OPT.

From the OPT we have $y = \hat{y} + \hat{u}$ and $\hat{u} \perp \hat{y}$.

Applying the Pythagorean law completes the proof.

8 Orthogonalization and Decomposition

Let's return to the connection between linear independence and orthogonality touched on above.

A result of much interest is a famous algorithm for constructing orthonormal sets from linearly independent sets.

The next section gives details.

8.1 Gram-Schmidt Orthogonalization

Theorem For each linearly independent set $\{x_1, \dots, x_k\} \subset \mathbb{R}^n$, there exists an orthonormal set $\{u_1, \dots, u_k\}$ with

$$\text{span}\{x_1, \dots, x_i\} = \text{span}\{u_1, \dots, u_i\} \quad \text{for } i = 1, \dots, k$$

The **Gram-Schmidt orthogonalization** procedure constructs an orthogonal set $\{u_1, u_2, \dots, u_n\}$.

One description of this procedure is as follows:

- For $i = 1, \dots, k$, form $S_i := \text{span}\{x_1, \dots, x_i\}$ and S_i^\perp
- Set $v_1 = x_1$
- For $i \geq 2$ set $v_i := \hat{E}_{S_{i-1}^\perp} x_i$ and $u_i := v_i / \|v_i\|$

The sequence u_1, \dots, u_k has the stated properties.

A Gram-Schmidt orthogonalization construction is a key idea behind the Kalman filter described in [A First Look at the Kalman filter](#).

In some exercises below you are asked to implement this algorithm and test it using projection.

8.2 QR Decomposition

The following result uses the preceding algorithm to produce a useful decomposition.

Theorem If X is $n \times k$ with linearly independent columns, then there exists a factorization $X = QR$ where

- R is $k \times k$, upper triangular, and nonsingular
- Q is $n \times k$ with orthonormal columns

Proof sketch: Let

- $x_j := \text{col}_j(X)$
- $\{u_1, \dots, u_k\}$ be orthonormal with same span as $\{x_1, \dots, x_k\}$ (to be constructed using Gram-Schmidt)
- Q be formed from cols u_i

Since $x_j \in \text{span}\{u_1, \dots, u_j\}$, we have

$$x_j = \sum_{i=1}^j \langle u_i, x_j \rangle u_i \quad \text{for } j = 1, \dots, k$$

Some rearranging gives $X = QR$.

8.3 Linear Regression via QR Decomposition

For matrices X and y that overdetermine β in the linear equation system $y = X\beta$, we found the least squares approximator $\hat{\beta} = (X'X)^{-1}X'y$.

Using the QR decomposition $X = QR$ gives

$$\begin{aligned} \hat{\beta} &= (R'Q'QR)^{-1}R'Q'y \\ &= (R'R)^{-1}R'Q'y \\ &= R^{-1}(R')^{-1}R'Q'y = R^{-1}Q'y \end{aligned}$$

Numerical routines would in this case use the alternative form $R\hat{\beta} = Q'y$ and back substitution.

9 Exercises

9.1 Exercise 1

Show that, for any linear subspace $S \subset \mathbb{R}^n$, $S \cap S^\perp = \{0\}$.

9.2 Exercise 2

Let $P = X(X'X)^{-1}X'$ and let $M = I - P$. Show that P and M are both idempotent and symmetric. Can you give any intuition as to why they should be idempotent?

9.3 Exercise 3

Using Gram-Schmidt orthogonalization, produce a linear projection of y onto the column space of X and verify this using the projection matrix $P := X(X'X)^{-1}X'$ and also using QR decomposition, where:

$$y := \begin{pmatrix} 1 \\ 3 \\ -3 \end{pmatrix},$$

and

$$X := \begin{pmatrix} 1 & 0 \\ 0 & -6 \\ 2 & 2 \end{pmatrix}$$

10 Solutions

10.1 Exercise 1

If $x \in S$ and $x \in S^\perp$, then we have in particular that $\langle x, x \rangle = 0$. But then $x = 0$.

10.2 Exercise 2

Symmetry and idempotence of M and P can be established using standard rules for matrix algebra. The intuition behind idempotence of M and P is that both are orthogonal projections. After a point is projected into a given subspace, applying the projection again makes no difference. (A point inside the subspace is not shifted by orthogonal projection onto that space because it is already the closest point in the subspace to itself).

10.3 Exercise 3

Here's a function that computes the orthonormal vectors using the GS algorithm given in the lecture.

10.4 Setup

```
In [1]: using InstantiateFromURL
        # optionally add arguments to force installation: instantiate = true,
        ↪precompile = true
        github_project("QuantEcon/quantecon-notebooks-julia", version = "0.8.0")
```

```
In [2]: using LinearAlgebra, Statistics
```

```
In [3]: function gram_schmidt(X)

        U = similar(X, Float64) # for robustness

        function normalized_orthogonal_projection(b, Z)
            # project onto the orthogonal complement of the col span of Z
            orthogonal = I - Z * inv(Z'Z) * Z'
            projection = orthogonal * b
            # normalize
            return projection / norm(projection)
        end

        for col in 1:size(U, 2)
            # set up
            b = X[:,col] # vector we're going to project
            Z = X[:,1:col - 1] # first i-1 columns of X
            U[:,col] = normalized_orthogonal_projection(b, Z)
        end

        return U
    end
```

```
Out[3]: gram_schmidt (generic function with 1 method)
```

Here are the arrays we'll work with

```
In [4]: y = [1, 3, -3]
        X = [1 0; 0 -6; 2 2];
```

First let's do ordinary projection of y onto the basis spanned by the columns of X .

```
In [5]: Py1 = X * inv(X'X) * X' * y
```

```
Out[5]: 3-element Array{Float64,1}:
        -0.5652173913043479
         3.260869565217391
        -2.217391304347826
```

Now let's orthogonalize first, using Gram-Schmidt:

```
In [6]: U = gram_schmidt(X)
```

```
Out[6]: 3x2 Array{Float64,2}:
 0.447214 -0.131876
 0.0      -0.989071
 0.894427  0.065938
```

Now we can project using the orthonormal basis and see if we get the same thing:

```
In [7]: Py2 = U * U' * y
```

```
Out[7]: 3-element Array{Float64,1}:
 -0.5652173913043477
  3.260869565217391
 -2.2173913043478257
```

The result is the same. To complete the exercise, we get an orthonormal basis by QR decomposition and project once more.

```
In [8]: Q, R = qr(X)
        Q = Matrix(Q)
```

```
Out[8]: 3x2 Array{Float64,2}:
 -0.447214 -0.131876
  0.0      -0.989071
 -0.894427  0.065938
```

```
In [9]: Py3 = Q * Q' * y
```

```
Out[9]: 3-element Array{Float64,1}:
 -0.5652173913043474
  3.2608695652173907
 -2.2173913043478253
```

Again, the result is the same.

References

- [1] Steven Roman. *Advanced linear algebra*, volume 3. Springer, 2005.