3 Multiple linear regression

So far we have dealt with a single explanatory variable. We now generalize the ideas to multiple explanatory variables. Thus, the data is now

$$(x_{i1}, \ldots, x_{ip}, y_i), \qquad i = 1, ..., n.$$

The explanatory variable for the ith observation is a p-dimensional vector, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$, and we assume throughout $p + 1 \leq n$ (we will see soon why the case $p + 1 > n$ is problematic).

In the multiple explanatory variables case it is generally challenging to visualize the data points, because we quickly run out of dimensions if we attempt to generate a scatterplot (the case $p = 2$ is still doable because we can draw a 3-dimensional plot, showing X_1, X_2 on the XY-plane and the response Y on the Z axis). However, the mathematical concepts from the simple regression case can still be extended to the general-p case. Thus, we may still try to predict y by a *linear* function of $x_1, \ldots, x_p$, i.e., the analog of (1) will be

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_p x_p = \sum_{j=0}^{p} \hat{\beta}_j x_j,$$

where for each observation we prepend $x_{i0} \equiv 1$ to the vector of explanatory variables, that is, we define $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ip})$ instead of $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$.

For any method for fitting such a linear function (that is, for calculating $\hat{\beta}_1, \ldots, \hat{\beta}_p$ as a function of the observed data), the *fitted values* are

$$\hat{y}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij},$$

and the *residuals* are defined exactly as before,

$$e_i := y_i - \hat{y}_i$$

We can generalize also the least squares method, by seeking the set of coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$ which, as before, minimize the sum of squared errors. Formally, let $\boldsymbol{b} = (b_0, \ldots, b_p)^\top \in \mathbb{R}^{p+1}$, and define

$$Q(\boldsymbol{b}) := \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2. \tag{8}$$

This is a differentiable function (now of $p + 1$ variables), and we can take partial derivatives and set them to zero. We have

$$\frac{\partial}{\partial b_r} \sum_{j=0}^{p} x_{ij} b_j = x_{ir},$$

and, therefore,

$$\frac{\partial Q}{\partial b_r} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right) \left(\frac{\partial}{\partial b_r} \sum_{j=0}^{p} x_{ij} b_j \right) = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} x_{ij} b_j \right) x_{ir}.$$

Setting this to zero for $r = 0, \ldots, p$ yields the so-called *Normal equations*,

$$\boxed{\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) b_j = \sum_{i=1}^{n} x_{ir} y_i, \qquad r = 0, \ldots, p.} \tag{9}$$

The solution to this set of $p+1$ equations (in $p+1$ variables), assuming it exists and is unique (we will see conditions for this to be the case), gives the least squares coefficients $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

The system of equations (9) can be written equivalently in *matrix* form. Thus, define

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n, \qquad \boldsymbol{X} := \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times (p+1)}.$$

The $n \times (p+1)$ matrix $\boldsymbol{X}$ is called the *design matrix*, or simply the X-matrix. For $i = 1, \ldots, n$, the i th row of $\boldsymbol{X}$ is the $(p+1)$-dimensional row vector $\boldsymbol{x}_i^\top$, the feature vector for the i th sample point (transposed). For $j = 0, \ldots, p$, the j th column of $\boldsymbol{X}$ is the n-dimensional column vector $\boldsymbol{X}_j := (x_{1j}, \ldots, x_{nj})^\top \in \mathbb{R}^n$, the vector of values of the jth feature for each of the n observations. Note that $\boldsymbol{X}_0 = (1, \ldots, 1)^\top =: \mathbf{1}_n$.

With this notation,

$$\sum_{i=1}^{n} x_{ir} x_{ij} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)_{rj}$$

and

$$\sum_{i=1}^{n} x_{ir} y_i = \left(\boldsymbol{X}^\top \boldsymbol{y}\right)_r,$$

so we can write (2.2) as

$$\sum_{j=0}^{p} \left(\boldsymbol{X}^\top \boldsymbol{X}\right)_{rj} b_j = \left(\boldsymbol{X}^\top \boldsymbol{y}\right)_r, \quad r = 0, \ldots, p$$

or, equivalently,

$$\boldsymbol{X}^\top \boldsymbol{X} \boldsymbol{b} = \boldsymbol{X}^\top \boldsymbol{y}$$

Now, $\boldsymbol{X}^\top \boldsymbol{X}$ is a $(p+1) \times (p+1)$ matrix, and it is invertible if and only if the columns of $\boldsymbol{X}$ are linearly independent (prove this!). If $p+1 > n$ the columns of $\boldsymbol{X}$ are necessarily linearly dependent (because column rank $\leq \min(n, p+1) = n < p+1$). If $p+1 \leq n$, the columns of $\boldsymbol{X}$ are linearly independent if no feature vector $\boldsymbol{X}_j$ is a combination of the others, which is a reasonable assumption. If the columns of $\boldsymbol{X}$ are linearly dependent, $\boldsymbol{X}^\top \boldsymbol{X}$ is singular and (2.3) has infinitely many solutions, i.e., the LS coefficients $\hat{\boldsymbol{\beta}}$ are not unique (every solution for (2.3) is then said to be a LS solution). If the columns of $\boldsymbol{X}$ are linearly independent, which we will generally assume from now on, then (2.3) has a unique solution given by

$$\boxed{\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \qquad (10)}$$

Geometric interpretation of the LS solution. There is in fact a more direct (and more intuitive) way to derive the LS solution using *geometric* interpretation. We first recall some basic definitions and facts from linear algebra.

1. For a $n \times m$ matrix $\boldsymbol{A}$, the *image* is the linear subspace of $\mathbb{R}^n$ given by

$$\operatorname{Im}(\boldsymbol{A}) := \{\boldsymbol{A}\boldsymbol{v} : \boldsymbol{v} \in \mathbb{R}^m\}$$

Recall that, if $\boldsymbol{A}_j$ denotes the j-th column of $\boldsymbol{A}$, then $\boldsymbol{A}\boldsymbol{v} = \sum_j \boldsymbol{A}_j v_j$, so we have

$$\mathrm{Im}(\boldsymbol{A}) = \mathrm{sp}\left(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m\right) \subseteq \mathbb{R}^n,$$

i.e., the image is the linear subspace (of $\mathbb{R}^n$) *spanned* by the columns $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$.

2. The (standard) *inner product* of vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is $\boldsymbol{u}^\top \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{u} = \sum_{i=1}^n u_i v_i$.

3. The *norm* of a vector $\boldsymbol{v} \in \mathbb{R}^n$ is $\|\boldsymbol{v}\| = \left(\boldsymbol{v}^\top \boldsymbol{v}\right)^{1/2} = \left(\sum_{i=1}^n v_i^2\right)^{1/2}$. The *Euclidean distance* between two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ is $\|\boldsymbol{u} - \boldsymbol{v}\|$. $\ln \mathbb{R}^2, \mathbb{R}^3$ this coincides with the usual geometric notion of a distance between two points.

4. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$, we say that $\boldsymbol{v}$ is *orthogonal* to $\boldsymbol{u}$, and denote $\boldsymbol{u} \perp \boldsymbol{v}$, if $\boldsymbol{u}^\top \boldsymbol{v} = 0$. In $\mathbb{R}^2, \mathbb{R}^3$ this coincides with the usual geometric notion of perpendicularity.

5. The *orthogonal complement* of a subspace $M \subseteq \mathbb{R}^n$ is the subspace

$$M^\perp := \left\{\boldsymbol{v} : \boldsymbol{v}^\top \boldsymbol{u} = 0 \quad \forall \boldsymbol{u} \in M\right\}$$

(Exercise: verify that $M^\perp$ is indeed a linear space.)

6. *Pythagorean theorem*: If $\boldsymbol{u} \perp \boldsymbol{v}$, then $\|\boldsymbol{v} + \boldsymbol{u}\|^2 = \|\boldsymbol{v}\|^2 + \|\boldsymbol{u}\|^2$.

We now derive the LS solution (10) from an alternative, geometric viewpoint. For any $\boldsymbol{b} = (b_0, \ldots, b_p)^\top \in \mathbb{R}^{p+1}$ we can write the objective function (8) in vector form as

$$Q(\boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}\|^2.$$

Thus, if $\hat{\boldsymbol{\beta}}$ is the minimizer of $Q(\boldsymbol{b})$ over all $\boldsymbol{b} \in \mathbb{R}^{p+1}$, then $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ minimizes the squared norm (equivalently, it minimizes the Euclidean distance) between $\boldsymbol{y}$ and $\boldsymbol{z}$ among all vectors $\boldsymbol{z} \in \mathrm{Im}(\boldsymbol{X})$. Then $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ must be the *projection* of $\boldsymbol{y}$ onto $\mathrm{colsp}(\boldsymbol{X})$ (this claim requires a proof, but it's intuitive geometrically). Now, a necessary and sufficient condition for $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ to be the projection of $\boldsymbol{y}$ onto $\mathrm{colsp}(\boldsymbol{X})$ is

$$\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}, \tag{11}$$

which simply requires that the dot product between the residual $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ and each column of $\boldsymbol{X}$ is zero. Rearranging (11), we get

$$\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

which, assuming again that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, yields (10). The vector

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

is called the orthogonal projection of $\boldsymbol{y}$ onto $\mathrm{Im}(\boldsymbol{X})$, and the matrix

$$\boldsymbol{P}_{\boldsymbol{X}} := \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$$

i.e., the matrix such that $\hat{\boldsymbol{y}} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$, is called the projection matrix of $\boldsymbol{X}$ (this is the matrix projecting any vector onto the linear space spanned by the columns of $\boldsymbol{X}$).

Remark. We are assuming, as before, that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, although the projection matrix of $\boldsymbol{X}$ is well-defined also when the columns of $\boldsymbol{X}$ are linearly dependent; for example, this can be achieved by choosing a basis for $\mathrm{Im}(\boldsymbol{X})$ and writing (2.5) using the "thinner" matrix instead of $\boldsymbol{X}$).

Linear algebra interlude: projection matrices and related results.

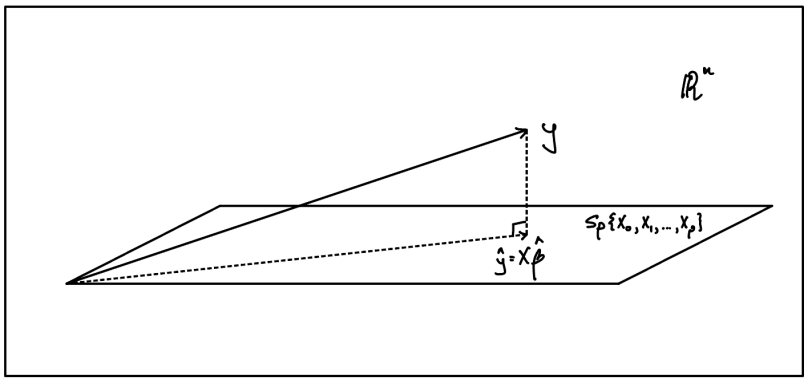

Figure 4: Geometric interpretation of the Least squares estimator

Proposition 4. *Let $\boldsymbol{X}$ be an $n \times m$ matrix and assume that it has linearly independent columns (i.e., full column rank; remember that this implies $m \leq n$). Then the projection matrix $\boldsymbol{P_X}$ has the following properties.*

1. $\boldsymbol{P}_X$ *is symmetric*
2. $\boldsymbol{P}_X$ *is idempotent,* $\boldsymbol{P}_X^2 = \boldsymbol{P}_X$
3. $\boldsymbol{P}_X \boldsymbol{X} = \boldsymbol{X}$
4. $\boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{P_X}) = \boldsymbol{0} \in \mathbb{R}^{m \times n}$
5. $\boldsymbol{P}_X v \in \operatorname{Im}(X)$ *for all* $v \in \mathbb{R}^n$
6. *If* $m = n$ *and* $\boldsymbol{X}$ *is invertible, then* $\boldsymbol{P_X} = \boldsymbol{I}$
7. $(\boldsymbol{I} - \boldsymbol{P_X})\, \boldsymbol{v} \in \operatorname{Im}(\boldsymbol{X})^\perp$ *for all* $\boldsymbol{v} \in \mathbb{R}^n$
8. *If* $w \in \operatorname{lm}(\boldsymbol{X})$*, then* $\boldsymbol{P_X w} = \boldsymbol{w}$
9. *If* $\boldsymbol{w} \in \operatorname{Im}(\boldsymbol{X})^\perp$*, then* $\boldsymbol{P}_X \boldsymbol{w} = \boldsymbol{0}$
10. *If* $\boldsymbol{Z}$ *is another* $n \times m$ *matrix s.t.* $\operatorname{Im}(\boldsymbol{Z}) = \operatorname{Im}(\boldsymbol{X})$*, then* $\boldsymbol{P_Z} = \boldsymbol{P_X}$*. This means that* $\boldsymbol{P_X}$ *depends on* $\boldsymbol{X}$ *only through the span of its columns. Hence, for an arbitrary linear space* M*, we can define the projection matrix* $\boldsymbol{P}_M$ *onto* M *(an explicit form for* $\boldsymbol{P}_M$ *can be obtained by taking any basis of* M *and stacking its elements as columns in a matrix* $\boldsymbol{X}$*, then forming* $\boldsymbol{P_X} := \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\Big)$
11. *If* L *and* M *are two subspaces with* $L \subseteq M$*, then* $\boldsymbol{P}_M \boldsymbol{P}_L = \boldsymbol{P}_L \boldsymbol{P}_M = \boldsymbol{P}_L$*.*

Proof.

1. $\boldsymbol{P_X^\top} = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top\right]^\top = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$ where we used the fact $\left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}\right]^\top = \left[\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^\top\right]^{-1} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1}$.

2. $\boldsymbol{P}_{\boldsymbol{X}}^2 = \left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right]\left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right] = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$

3. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}$

4. $\boldsymbol{X}^\top\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}\right) = \left[\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}\right)\boldsymbol{X}\right]^\top = \left[\boldsymbol{X}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\right]^\top = \left[\boldsymbol{X}-\boldsymbol{X}\right]^\top = \boldsymbol{0}^\top \in \mathbb{R}^{n\times m} = \boldsymbol{0} \in \mathbb{R}^{m\times n}$, where we used fact #3

5. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v}$ for all $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{v} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{v} = \boldsymbol{X}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{v}\right] \in \mathrm{Im}(\boldsymbol{X})$

6. If $\boldsymbol{P}_{\boldsymbol{X}}$ is (square and) invertible, $\left[\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\right]^{-1} = \left[\boldsymbol{X}^\top\right]^{-1}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)\boldsymbol{X}^{-1} = \boldsymbol{I}_n$.

7. $(\boldsymbol{X}\boldsymbol{u})^\top\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}\right)\boldsymbol{v} = \boldsymbol{u}^\top\underbrace{\boldsymbol{X}^\top\left(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}\right)}_{\boldsymbol{0}\in\mathbb{R}^{m\times n}}\boldsymbol{v} = 0$

8. $\boldsymbol{P}_{\boldsymbol{X}}\underbrace{(\boldsymbol{X}\boldsymbol{u})}_{w} = \left(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\right)\boldsymbol{u} = \underbrace{X\boldsymbol{u}}_{w}$

9. $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{w} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\underbrace{\boldsymbol{X}^\top\boldsymbol{w}}_{=0} = \boldsymbol{0}$

10. Denote $\boldsymbol{X}_j, \boldsymbol{Z}_j$ for the j th columns of $\boldsymbol{X}, \boldsymbol{Z}$, respectively. Then $\boldsymbol{Z}_j = \boldsymbol{X}\boldsymbol{h}_j$ for some $\boldsymbol{h}_j \in \mathbb{R}^m$ because $\boldsymbol{Z}_j \in \mathrm{Im}(\boldsymbol{Z}) = \mathrm{Im}(\boldsymbol{X})$. Putting $\boldsymbol{H} := [\boldsymbol{h}_1\boldsymbol{h}_2\cdots\boldsymbol{h}_m] \in \mathbb{R}^{m\times m}$, we have $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{H}$. Further, if $\boldsymbol{w} \in \mathbb{R}^m$ is s.t. $\boldsymbol{H}\boldsymbol{w} = \boldsymbol{0}$, then $\mathbf{Z}\boldsymbol{w} = \boldsymbol{X}\boldsymbol{H}\boldsymbol{w} = \boldsymbol{0}$, which implies $\boldsymbol{w} = \boldsymbol{0}$ because $\boldsymbol{Z}$ was assumed to have full column rank. Hence, $\boldsymbol{H}$ is invertible. Then

$$\boldsymbol{P}_{\boldsymbol{Z}} = \boldsymbol{Z}\left(\boldsymbol{Z}^\top\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^\top = \boldsymbol{X}\boldsymbol{H}\left[(\boldsymbol{X}\boldsymbol{H})^\top\boldsymbol{X}\boldsymbol{H}\right]^{-1}(\boldsymbol{X}\boldsymbol{H})^\top =$$
$$= \boldsymbol{X}\boldsymbol{H}\boldsymbol{H}^{-1}\boldsymbol{X}^{-1}\left[(\boldsymbol{X}\boldsymbol{H})^\top\right]^{-1}(\boldsymbol{X}\boldsymbol{H})^\top = \boldsymbol{X}\left[\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\right]\boldsymbol{X}^\top = \boldsymbol{P}_{\boldsymbol{X}}$$

11. Let $\boldsymbol{v}$ be any vector. Then $\boldsymbol{P}_M\underbrace{\boldsymbol{P}_L\boldsymbol{v}}_{\in M} \overset{8}{=} \boldsymbol{P}_L\boldsymbol{v}$, implying $\boldsymbol{P}_M\boldsymbol{P}_L = \boldsymbol{P}_L$. Transposing both sides and using the fact that $\boldsymbol{P}_M, \boldsymbol{P}_L$ are both symmetric, we obtain also $\boldsymbol{P}_L\boldsymbol{P}_M = \boldsymbol{P}_L$.

□

Additionally, we have the following results related to projection matrices.

Proposition 5. *Let M be a subspace of $\mathbb{R}^n$ with $\dim(M) = m \leq n$. Then any vector $\boldsymbol{v} \in \mathbb{R}^n$ can be uniquely represented as $\boldsymbol{v} = \boldsymbol{w} + \boldsymbol{z}$ where $\boldsymbol{w} \in M$ and $z \in M^\perp$, the orthogonal complement of M. Moreover, in this representation, $\boldsymbol{w} = \boldsymbol{P}_M\boldsymbol{v}$, the projection of $\boldsymbol{v}$ onto M, and satisfies $\boldsymbol{w} = \underset{\boldsymbol{u}\in M}{\arg\min}\|\boldsymbol{v}-\boldsymbol{u}\|^2$.*

Proof. Taking $\boldsymbol{w} = \boldsymbol{P}_M\boldsymbol{v}, \boldsymbol{z} = \boldsymbol{v} - \boldsymbol{P}_M\boldsymbol{v}$, we have $\boldsymbol{w} \in M, z \in M^\perp$ by the properties above, and $\boldsymbol{v} = \boldsymbol{w} + \boldsymbol{z}$. We show that this representation is unique. Thus, suppose $\boldsymbol{v} = \boldsymbol{w}_1 + \boldsymbol{z}_1, \boldsymbol{v} = \boldsymbol{w}_2 + \boldsymbol{z}_2$ for $\boldsymbol{w}_1, \boldsymbol{w}_2 \in M, \boldsymbol{z}_1, \boldsymbol{z}_2 \in M^\perp$. Then $\boldsymbol{0} = (\boldsymbol{w}_1 + \boldsymbol{z}_1) - (\boldsymbol{w}_2 + \boldsymbol{z}_2) = (\boldsymbol{w}_1 - \boldsymbol{w}_2) + (\boldsymbol{z}_1 - \boldsymbol{z}_2)$, and, furthermore, $(\boldsymbol{w}_1 - \boldsymbol{w}_2) \perp (\boldsymbol{z}_1 - \boldsymbol{z}_2)$ because $(\boldsymbol{w}_1 - \boldsymbol{w}_2) \in M, (\boldsymbol{z}_1 - \boldsymbol{z}_2) \in M^\perp$ (remember that each of $M, M^\perp$ is a subspace so closed under addition/subtraction). Therefore,

$$0 = \|\boldsymbol{0}\|^2 = \|\,(\boldsymbol{w}_1 - \boldsymbol{w}_2) + (\boldsymbol{z}_1 - \boldsymbol{z}_2)\,\|^2 = \|\boldsymbol{w}_1 - \boldsymbol{w}_2\|^2 + \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2 \Rightarrow (\boldsymbol{w}_1 - \boldsymbol{w}_2) = \boldsymbol{0}, (\boldsymbol{z}_1 - \boldsymbol{z}_2) = \boldsymbol{0}.$$

It remains to show that $\boldsymbol{w}$ minimizes the Euclidean distance from $\boldsymbol{v}$ to M. Take any $\boldsymbol{u} \in M$. Then

$$\|\boldsymbol{v}-\boldsymbol{u}\|^2 = \|(\boldsymbol{w}+\boldsymbol{z})-\boldsymbol{u}\|^2 = \|\boldsymbol{w}-\boldsymbol{u}+\boldsymbol{z}\|^2 = \|\boldsymbol{w}-\boldsymbol{u}\|^2 + \|\boldsymbol{z}\|^2 \geq \|\boldsymbol{z}\|^2,$$

and on the other hand, for $\boldsymbol{u} = \boldsymbol{w} = \boldsymbol{P}_M \boldsymbol{v}$ we have $\|\boldsymbol{v}-\boldsymbol{u}\|^2 = \|\boldsymbol{z}\|^2$. □

Proposition 6. *We have*

1. $\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}} = P_{Im(\boldsymbol{X})^\perp}$
2. *if L and M are two subspaces of $\mathbb{R}^n$ with $L \subseteq M$, then $\boldsymbol{P}_M - \boldsymbol{P}_L = \boldsymbol{P}_{M \cap L^\perp}$*

Proof. (a) this part follows from the uniqueness of the representation in Proposition 1. (b) this also follows from the uniqueness of the representation in Proposition 1, taking the original space to be M (I.e., the space to which v belongs) and the subspace (what's denoted M in Proposition 1) to be L. □

Proposition 7. *Let $\boldsymbol{Q}$ be an $n \times n$ matrix of rank $m \leq n$ which is symmetric and idempotent, $\boldsymbol{Q}^\top = \boldsymbol{Q}, \boldsymbol{Q}^2 = \boldsymbol{Q}$. Then $\boldsymbol{Q} = \boldsymbol{P}_M$ where $M := \mathrm{Im}(\boldsymbol{Q})$.*

Proof. Exercise. □

Diagonalizability and positive-semidefiniteness of a projection matrix. First, we recall some facts:

1. A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is diagonalizable (over $\mathbb{R}$) if there is an invertible matrix $\boldsymbol{P} \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$.

   Remark: $\boldsymbol{P}^{-1}$ is the transition matrix from the standard basis of $\mathbb{R}^n$ into the basis in the columns of $\boldsymbol{P}$).

2. If $\boldsymbol{A} \in \mathbb{R}$ is *symmetric*, then there is an *orthogonal* matrix $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and a diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$.

   Remark: (i) a square matrix $\boldsymbol{U}$ is *orthogonal* if $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$ (this implies $\boldsymbol{U}^\top = \boldsymbol{U}^{-1}$, because for any two square matrices $\boldsymbol{A}, \boldsymbol{B}$, we have $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I} \Rightarrow \boldsymbol{B}\boldsymbol{A} = \boldsymbol{I}$; prove this!). (ii) consistent with the above, you should have seen a proof in your linear algebra course that if $\boldsymbol{A}$ is a (real) symmetric matrix, then it is diagonalizable, and all of its eigenvalues are real.

3. A symmetric matrix $\boldsymbol{A}$ is *positive semidefinite* if all of its eigenvalues are nonnegative. It is called *positive definite* if all of its eigenvalues are positive.

   Remark: this is equivalent to saying that the *quadratic form* given by $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is *nonnegative* for all vectors $\boldsymbol{x}$ (positive semidefinite) or that $\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}$ is *positive* (positive definite).

4. Let $\boldsymbol{A}$ be a positive semidefinite matrix. Then:

   (i) there is $\boldsymbol{B}$ square such that $\boldsymbol{B}^2 = \boldsymbol{A}$, and in fact this representation is unique: take $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top$, where $\boldsymbol{D}^{1/2} := \mathrm{diag}(d_1, ..., d_n)$, so $\boldsymbol{B}^2 = \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^\top = \boldsymbol{A}$ (this $\boldsymbol{B}$ is often called the square root of $\boldsymbol{A}$).

   (ii) there is $\boldsymbol{B}$ square such that $\boldsymbol{B}\boldsymbol{B}^\top = \boldsymbol{A}$: take $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{/12}$, then $\boldsymbol{B}^2 = \boldsymbol{U}\boldsymbol{D}^{/12}(\boldsymbol{U}\boldsymbol{D}^{/12})^\top = \boldsymbol{A}$

Now suppose that $\boldsymbol{Q} \in \mathbb{R}^{n \times n}$ is a projection matrix onto a subspace $M \subseteq \mathbb{R}^n$ of dimension $\dim(M) = m$, then $\boldsymbol{Q}$ is positive semidefinite, and there is a representation

$$\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top, \qquad \boldsymbol{D} = \mathrm{diag}(\underbrace{1, ..., 1}_{m}, \underbrace{0, ..., 0}_{n-m}),$$

where $\boldsymbol{U}$ is orthogonal, such that the first m columns of $\boldsymbol{U}$ are an orthonormal basis of M, and the last $n-m$ columns of $\boldsymbol{U}$ are an orthonormal basis of $M^{\perp}$.
Remark: consistent with the previous item, note that $\boldsymbol{Q} = \boldsymbol{B}\boldsymbol{B}^{\top}$ for $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{D}^{1/2} = \tilde{\boldsymbol{U}} := [\boldsymbol{U}_1, ..., \boldsymbol{U}_m]$.