Lecture Notes: Multiple Linear Regression

Your Friendly Math Teacher

April 2, 2025

1 Introduction: Beyond a Single Predictor

In our previous discussions, we explored how to model a response variable y using a single explanatory variable x. While useful, real-world phenomena are often influenced by multiple factors. Multiple linear regression extends the concepts we've learned to handle situations where we have several explanatory variables influencing our response.

1.1 The Data

Our dataset now consists of n observations, where each observation i includes p explanatory variables and one response variable:

$$(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i), \quad i = 1, \ldots, n$$

. We collect the explanatory variables for the i-th observation into a p-dimensional vector $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$. A crucial assumption we'll make throughout is that $n \geq p + 1$. We'll see later why this is important for obtaining a unique solution.

1.2 Visualization Challenges

Visualizing this data becomes tricky as p increases. For simple linear regression ($p = 1$), a 2D scatter plot suffices. For $p = 2$, we can manage a 3D plot, showing x_1 and x_2 on the base plane and y on the vertical axis. Beyond that, direct scatter plots are impractical. However, the underlying mathematical framework elegantly extends.

2 The Multiple Linear Regression Model

Our goal is to predict y using a linear combination of *all* the explanatory variables $x_1, \ldots, x_p$. The model takes the form:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

. Here, $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are the coefficients we need to estimate from the data. $\hat{\beta}_0$ is the intercept, and $\hat{\beta}_j$ ($j = 1, \ldots, p$) represents the estimated change in y for a one-unit change in x_j, holding all other predictors constant.

2.1 Introducing the Intercept Term

To simplify notation, we often incorporate the intercept term into the summation. We achieve this by defining a new "zeroth" explanatory variable x_{i0} that is always equal to 1 for every observation

i. Our feature vector for observation i then becomes $(p+1)$-dimensional: $\mathbf{x}_i = (x_{i0}, x_{i1}, \ldots, x_{ip})^\top = (1, x_{i1}, \ldots, x_{ip})^\top$. With this convention, the predicted value for the i-th observation is:

$$\hat{y}_i = \hat{\beta}_0 x_{i0} + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} = \sum_{j=0}^{p} \hat{\beta}_j x_{ij}$$

.

2.2 Fitted Values and Residuals

Just like in simple regression, for any estimated set of coefficients $\hat{\beta} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top$, we can calculate:

- **Fitted values** ($\hat{y}_i$): The model's prediction for observation i, given by $\hat{y}_i = \sum_{j=0}^{p} \hat{\beta}_j x_{ij}$.
- **Residuals** (e_i): The difference between the observed value and the fitted value, $e_i = y_i - \hat{y}_i$. These represent the prediction errors.

3 The Method of Least Squares

How do we find the "best" coefficients $\hat{\beta}_0, \ldots, \hat{\beta}_p$? We generalize the method of least squares: we seek the coefficients that minimize the sum of the squared residuals (SSR).

Definition 3.1 (Least Squares Criterion)**.** Let $\mathbf{b} = (b_0, b_1, \ldots, b_p)^\top \in \mathbb{R}^{p+1}$ be a vector of candidate coefficients. The sum of squared residuals function is:

$$Q(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right)^2$$

. The least squares estimate $\hat{\beta}$ is the vector $\mathbf{b}$ that minimizes $Q(\mathbf{b})$.

3.1 Finding the Minimum: The Normal Equations

$Q(\mathbf{b})$ is a differentiable function of the $p+1$ variables $b_0, \ldots, b_p$. To find the minimum, we take the partial derivative with respect to each coefficient b_r and set it to zero. First, note that for a fixed i and r:

$$\frac{\partial}{\partial b_r} \left(\sum_{j=0}^{p} b_j x_{ij} \right) = x_{ir}$$

. Using the chain rule for differentiation:

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

. Setting $\frac{\partial Q}{\partial b_r} = 0$ for each $r = 0, 1, \ldots, p$, we get:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} b_j x_{ij} \right) x_{ir} = 0$$

Rearranging this gives:

$$\sum_{i=1}^{n} x_{ir} y_i = \sum_{i=1}^{n} x_{ir} \sum_{j=0}^{p} b_j x_{ij} = \sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) b_j$$

This yields a system of $p+1$ linear equations for the $p+1$ unknown coefficients $b_0, \ldots, b_p$.

Definition 3.2 (Normal Equations (Scalar Form))**.** The set of equations that determine the least squares coefficients $\mathbf{b}$ is:

$$\sum_{j=0}^{p} \left(\sum_{i=1}^{n} x_{ir} x_{ij} \right) b_j = \sum_{i=1}^{n} x_{ir} y_i, \quad \text{for } r = 0, 1, \ldots, p$$

. The solution to this system, if unique, gives the least squares estimate $\hat{\beta}$.

4 Matrix Formulation

Writing these equations individually is cumbersome. Linear algebra provides a much more compact and insightful representation.

4.1 Key Matrices and Vectors

Let's define the following:

- **Response vector $\mathbf{y} \in \mathbb{R}^n$**: The vector of observed responses.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

  .

- **Design Matrix $\mathbf{X} \in \mathbb{R}^{n \times (p+1)}$**: The matrix whose rows are the feature vectors (including the leading 1 for the intercept) for each observation.

$$\mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

  . The i-th row is $\mathbf{x}_i^\top$. The j-th column, denoted $\mathbf{X}_j$, contains all n observations for the j-th predictor ($j = 1, \ldots, p$). The 0-th column $\mathbf{X}_0$ is a vector of all ones, $\mathbf{1}_n$.

- **Coefficient vector $\mathbf{b} \in \mathbb{R}^{p+1}$**: The vector of coefficients we want to estimate.

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

  .

4.2 Normal Equations in Matrix Form

Now, let's look at the terms in the scalar normal equations:

- $\sum_{i=1}^{n} x_{ir}x_{ij}$: This is the dot product of the r-th column of $\mathbf{X}$ and the j-th column of $\mathbf{X}$. This is precisely the element at row r, column j of the matrix product $\mathbf{X}^\top\mathbf{X}$. That is, $(\mathbf{X}^\top\mathbf{X})_{rj} = \sum_{i=1}^{n} x_{ir}x_{ij}$.
- $\sum_{i=1}^{n} x_{ir}y_i$: This is the dot product of the r-th column of $\mathbf{X}$ and the vector $\mathbf{y}$. This is the r-th element of the vector $\mathbf{X}^\top\mathbf{y}$. That is, $(\mathbf{X}^\top\mathbf{y})_r = \sum_{i=1}^{n} x_{ir}y_i$.

Substituting these into the scalar normal equations $\sum_{j=0}^{p}(\mathbf{X}^\top\mathbf{X})_{rj}b_j = (\mathbf{X}^\top\mathbf{y})_r$ for $r = 0, \ldots, p$, we recognize this as the definition of matrix-vector multiplication.

Theorem 4.1 (Normal Equations (Matrix Form))**.** The system of normal equations in matrix form is:

$$(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{y}$$

.

This is a compact equation for the vector $\mathbf{b}$!

4.3 Solving for the Coefficients

The equation $(\mathbf{X}^\top\mathbf{X})\mathbf{b} = \mathbf{X}^\top\mathbf{y}$ is a standard linear system $A\mathbf{x} = \mathbf{c}$, where $A = \mathbf{X}^\top\mathbf{X}$ is a $(p+1) \times (p+1)$ square matrix, $\mathbf{x} = \mathbf{b}$, and $\mathbf{c} = \mathbf{X}^\top\mathbf{y}$.

A unique solution exists if and only if the matrix $\mathbf{X}^\top\mathbf{X}$ is invertible. When is this true?

- $\mathbf{X}^\top\mathbf{X}$ is invertible if and only if the columns of the design matrix $\mathbf{X}$ are linearly independent. This means no predictor column (including the column of ones) can be written as a linear combination of the others.
- A necessary condition for the columns of $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$ to be linearly independent is that the number of columns does not exceed the number of rows, i.e., $p + 1 \leq n$. This is the assumption we made earlier! If $p + 1 > n$, the columns *must* be linearly dependent, $\mathbf{X}^\top\mathbf{X}$ is singular (not invertible), and the system has infinitely many solutions. The least squares coefficients are not unique in this case.
- Assuming $p+1 \leq n$, linear independence usually holds unless there's perfect multicollinearity (one predictor is an exact linear function of others) or a predictor is constant (and thus a multiple of the intercept column).

Theorem 4.2 (Least Squares Solution)**.** If the columns of the design matrix $\mathbf{X}$ are linearly independent (which implies $p + 1 \leq n$ and requires $\mathbf{X}^\top\mathbf{X}$ to be invertible), the unique least squares estimate $\hat{\beta}$ is given by:

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

.

Remark 4.1. From now on, we will generally assume the columns of $\mathbf{X}$ are linearly independent, ensuring a unique solution exists.

5 Geometric Interpretation

There's a beautiful geometric way to understand the least squares solution, connecting it to concepts from linear algebra like vector spaces and projections.

5.1 Linear Algebra Preliminaries

Let's refresh some key ideas:

- **Image (Column Space)**: For an $n \times m$ matrix $\mathbf{A}$, its image, $\text{Im}(\mathbf{A})$, is the subspace of $\mathbb{R}^n$ spanned by its columns $\mathbf{A}_1, \ldots, \mathbf{A}_m$. It's the set of all possible vectors you can form by linear combinations $A\mathbf{v}$ for $\mathbf{v} \in \mathbb{R}^m$. We often call this the column space, denoted $colsp(\mathbf{A})$.
- **Inner Product**: For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$, the standard inner product (dot product) is $\mathbf{u}^\top \mathbf{v} = \sum_{i=1}^n u_i v_i$.
- **Norm**: The Euclidean norm (length) of $\mathbf{v}$ is $\|\mathbf{v}\| = \sqrt{\mathbf{v}^\top \mathbf{v}} = \sqrt{\sum v_i^2}$. The distance between $\mathbf{u}$ and $\mathbf{v}$ is $\|\mathbf{u} - \mathbf{v}\|$.
- **Orthogonality**: $\mathbf{u}$ is orthogonal to $\mathbf{v}$ ($\mathbf{u} \perp \mathbf{v}$) if $\mathbf{u}^\top \mathbf{v} = 0$.
- **Orthogonal Complement**: For a subspace $M \subseteq \mathbb{R}^n$, its orthogonal complement $M^\perp$ is the subspace of all vectors orthogonal to *every* vector in M: $M^\perp = \{\mathbf{v} \in \mathbb{R}^n \mid \mathbf{v}^\top \mathbf{u} = 0 \text{ for all } \mathbf{u} \in M\}$.
- **Pythagorean Theorem**: If $\mathbf{u} \perp \mathbf{v}$, then $\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$.

5.2 Least Squares as Projection

Recall the least squares objective function $Q(\mathbf{b}) = \sum_{i=1}^n (y_i - \sum_{j=0}^p b_j x_{ij})^2$. Let $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ be the vector of predicted values corresponding to the coefficient vector $\mathbf{b}$. Then we can write the objective function using vector norms:

$$Q(\mathbf{b}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

. Minimizing $Q(\mathbf{b})$ means finding the vector $\mathbf{b}$ such that the vector $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ is *closest* to the observed vector $\mathbf{y}$ in terms of Euclidean distance.

Where do the possible predicted vectors $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ live? Since $\hat{\mathbf{y}}$ is a linear combination of the columns of $\mathbf{X}$ (with coefficients given by $\mathbf{b}$), all possible predicted vectors $\hat{\mathbf{y}}$ form the column space of $\mathbf{X}$, $colsp(\mathbf{X}) = \text{Im}(\mathbf{X})$.

So, the least squares problem is equivalent to finding the vector $\hat{\mathbf{y}}$ within the subspace $colsp(\mathbf{X})$ that is closest to the vector $\mathbf{y}$. Geometrically, this closest vector is the **orthogonal projection** of $\mathbf{y}$ onto the subspace $colsp(\mathbf{X})$. Let $\hat{\beta}$ be the coefficient vector that produces this projection, so $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$.

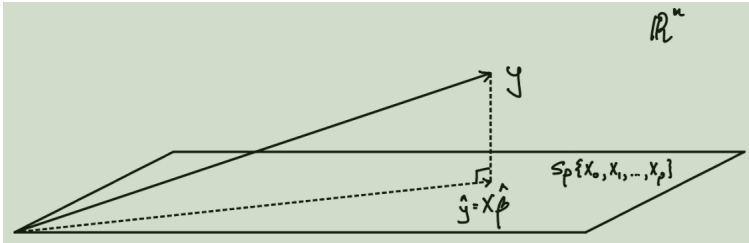

Figure 1: Geometric interpretation: $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ (denoted $Sp\{x_0, ..., x_p\}$ in the diagram).

What defines the orthogonal projection? The vector connecting $\hat{\mathbf{y}}$ to $\mathbf{y}$, which is the residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, must be orthogonal to the subspace $colsp(\mathbf{X})$. This means $\mathbf{e}$ must be orthogonal to every vector in $colsp(\mathbf{X})$, and in particular, it must be orthogonal to each basis vector of the subspace, i.e., the columns of $\mathbf{X}$.

Mathematically, the orthogonality condition is $\mathbf{X}_j^\top \mathbf{e} = 0$ for all $j = 0, \ldots, p$. We can write this compactly in matrix form:

$$\mathbf{X}^\top \mathbf{e} = \mathbf{0}$$

Substituting $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, we get:

$$\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$$

. Rearranging this gives:

$$\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{0} \implies \mathbf{X}^\top \mathbf{X}\hat{\beta} = \mathbf{X}^\top \mathbf{y}$$

. Look familiar? These are exactly the Normal Equations we derived earlier using calculus!

Assuming again that $\mathbf{X}^\top \mathbf{X}$ is invertible, we solve for $\hat{\beta}$:

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

. This provides a satisfying geometric intuition for the algebraic solution.

6 The Projection Matrix (Hat Matrix)

The vector of fitted values $\hat{\mathbf{y}}$ is obtained by projecting $\mathbf{y}$ onto $colsp(\mathbf{X})$. We can express this projection using a matrix operation. Substituting the formula for $\hat{\beta}$ into $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$, we get:

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

.

Definition 6.1 (Projection Matrix / Hat Matrix)**.** The matrix $\mathbf{P_X} \in \mathbb{R}^{n \times n}$ defined as

$$\mathbf{P_X} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

is the orthogonal projection matrix onto the column space of $\mathbf{X}$, $colsp(\mathbf{X})$. It projects any vector $\mathbf{v} \in \mathbb{R}^n$ onto $colsp(\mathbf{X})$. The fitted values are given by $\hat{\mathbf{y}} = \mathbf{P_X}\mathbf{y}$. (It's sometimes called the "hat matrix" because it puts the hat on $\mathbf{y}$).

Remark 6.1. This definition assumes $\mathbf{X}^\top \mathbf{X}$ is invertible (i.e., columns of $\mathbf{X}$ are linearly independent). The concept of projection still exists even if columns are dependent, but the formula needs adjustment, often by first finding a basis for $colsp(\mathbf{X})$.

6.1 Properties of Projection Matrices

Let $\mathbf{P_X}$ be the projection matrix onto $colsp(\mathbf{X})$, assuming $\mathbf{X}$ has full column rank $(p+1)$. It has several important properties:

Proposition 6.1. Let $\mathbf{X}$ be $n \times (p+1)$ with linearly independent columns. Then $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ satisfies:

1. **Symmetry**: $\mathbf{P_X}^\top = \mathbf{P_X}$.
2. **Idempotence**: $\mathbf{P_X}^2 = \mathbf{P_X}$. (Projecting something already in the subspace doesn't change it).
3. $\mathbf{P_X}\mathbf{X} = \mathbf{X}$. (Projecting the columns of $\mathbf{X}$ onto their own span leaves them unchanged).
4. $\mathbf{X}^\top(\mathbf{I} - \mathbf{P_X}) = \mathbf{0}$. (The residual space, associated with $\mathbf{I} - \mathbf{P_X}$, is orthogonal to the column space).
5. For any $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{P_X}\mathbf{v} \in \mathrm{Im}(\mathbf{X})$.
6. If $p + 1 = n$ (so $\mathbf{X}$ is square and invertible), then $\mathbf{P_X} = \mathbf{I}_n$.
7. For any $\mathbf{v} \in \mathbb{R}^n$, $(\mathbf{I} - \mathbf{P_X})\mathbf{v} \in \mathrm{Im}(\mathbf{X})^\perp$. ($\mathbf{I} - \mathbf{P_X}$ projects onto the orthogonal complement).
8. If $\mathbf{w} \in \mathrm{Im}(\mathbf{X})$, then $\mathbf{P_X}\mathbf{w} = \mathbf{w}$.
9. If $\mathbf{w} \in \mathrm{Im}(\mathbf{X})^\perp$, then $\mathbf{P_X}\mathbf{w} = \mathbf{0}$.
10. $\mathbf{P_X}$ depends only on the subspace $\mathrm{Im}(\mathbf{X})$, not the specific basis chosen for it. If $\mathrm{Im}(\mathbf{Z}) = \mathrm{Im}(\mathbf{X})$, then $\mathbf{P_Z} = \mathbf{P_X}$.
11. If L, M are subspaces with $L \subseteq M$, then $\mathbf{P}_M\mathbf{P}_L = \mathbf{P}_L\mathbf{P}_M = \mathbf{P}_L$. (Projecting onto L then M is the same as just projecting onto L).

Proof. (Selected proofs, see original notes for others)

1. $\mathbf{P_X}^\top = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top]^\top = (\mathbf{X}^\top)^\top((\mathbf{X}^\top\mathbf{X})^{-1})^\top\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{P_X}$. (Used $(\mathbf{A}^{-1})^\top = (\mathbf{A}^\top)^{-1}$ and $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$, and $(\mathbf{X}^\top\mathbf{X})$ is symmetric).
2. $\mathbf{P_X}^2 = [\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top][\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top] = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{I}\mathbf{X}^\top = \mathbf{P_X}$.
3. $\mathbf{P_X}\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{X}) = \mathbf{X}\mathbf{I} = \mathbf{X}$.

□

6.2 Orthogonal Decomposition and Projections

Projection matrices are fundamental to decomposing vectors relative to a subspace and its orthogonal complement.

Proposition 6.2 (Orthogonal Decomposition)**.** Let M be a subspace of $\mathbb{R}^n$. Any vector $\mathbf{v} \in \mathbb{R}^n$ can be uniquely represented as $\mathbf{v} = \mathbf{w} + \mathbf{z}$, where $\mathbf{w} \in M$ and $\mathbf{z} \in M^\perp$. In this decomposition, $\mathbf{w} = \mathbf{P}_M\mathbf{v}$ and $\mathbf{z} = (\mathbf{I} - \mathbf{P}_M)\mathbf{v} = \mathbf{P}_{M^\perp}\mathbf{v}$. Furthermore, $\mathbf{w} = \mathbf{P}_M\mathbf{v}$ is the unique vector in M that minimizes the squared distance to $\mathbf{v}$, i.e., $\mathbf{w} = \arg\min_{\mathbf{u}\in M} \|\mathbf{v} - \mathbf{u}\|^2$.

Proof. Let $\mathbf{w} = \mathbf{P}_M\mathbf{v}$ and $\mathbf{z} = \mathbf{v} - \mathbf{P}_M\mathbf{v}$. By properties of $\mathbf{P}_M$, $\mathbf{w} \in M$ and $\mathbf{z} \in M^\perp$. Clearly $\mathbf{v} = \mathbf{w} + \mathbf{z}$. For uniqueness, suppose $\mathbf{v} = \mathbf{w}_1 + \mathbf{z}_1 = \mathbf{w}_2 + \mathbf{z}_2$ with $\mathbf{w}_1, \mathbf{w}_2 \in M$ and $\mathbf{z}_1, \mathbf{z}_2 \in M^\perp$. Then $\mathbf{w}_1 - \mathbf{w}_2 = \mathbf{z}_2 - \mathbf{z}_1$. The LHS is in M and the RHS is in $M^\perp$. Since the only vector in both M and $M^\perp$ is $\mathbf{0}$, we must have $\mathbf{w}_1 - \mathbf{w}_2 = \mathbf{0}$ and $\mathbf{z}_2 - \mathbf{z}_1 = \mathbf{0}$, so $\mathbf{w}_1 = \mathbf{w}_2$ and $\mathbf{z}_1 = \mathbf{z}_2$. For the minimization property: take any $\mathbf{u} \in M$. Then $\mathbf{w} - \mathbf{u} \in M$, and $\mathbf{z} \in M^\perp$, so $(\mathbf{w} - \mathbf{u}) \perp \mathbf{z}$. By Pythagoras:

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

. This is minimized when $\|\mathbf{w} - \mathbf{u}\|^2 = 0$, i.e., when $\mathbf{u} = \mathbf{w}$. The minimum value is $\|\mathbf{z}\|^2 = \|(\mathbf{I} - \mathbf{P}_M)\mathbf{v}\|^2$. □

Proposition 6.3. We have the following identities:

1. $\mathbf{I} - \mathbf{P}_\mathbf{X}$ is the projection matrix onto $\mathrm{Im}(\mathbf{X})^\perp$. We denote it $\mathbf{P}_{\mathrm{Im}(\mathbf{X})^\perp}$.
2. If $L \subseteq M$ are subspaces, then $\mathbf{P}_M - \mathbf{P}_L$ is the projection matrix onto the subspace $M \cap L^\perp$ (the part of M that is orthogonal to L).

Proposition 6.4. Any symmetric ($\mathbf{Q}^\top = \mathbf{Q}$) and idempotent ($\mathbf{Q}^2 = \mathbf{Q}$) matrix $\mathbf{Q}$ is the projection matrix onto its own image, $M = \mathrm{Im}(\mathbf{Q})$.

6.3 Spectral Properties of Projection Matrices

Projection matrices have a particularly simple structure when viewed through their eigenvalues and eigenvectors. Recall:

- A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is diagonalizable by an orthogonal matrix: $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^\top$, where $\mathbf{U}$ is orthogonal ($\mathbf{U}^\top\mathbf{U} = \mathbf{I}$, so $\mathbf{U}^{-1} = \mathbf{U}^\top$) and $\mathbf{D}$ is diagonal. The diagonal entries of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{U}$ are the corresponding orthonormal eigenvectors. All eigenvalues of a real symmetric matrix are real.
- A symmetric matrix $\mathbf{A}$ is positive semidefinite (PSD) if all its eigenvalues are ≥ 0. This is equivalent to $\mathbf{x}^\top\mathbf{A}\mathbf{x} \geq 0$ for all $\mathbf{x}$. It is positive definite (PD) if eigenvalues are > 0 (equiv. $\mathbf{x}^\top\mathbf{A}\mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$).
- Any PSD matrix $\mathbf{A}$ has a unique PSD square root $\mathbf{B} = \mathbf{A}^{1/2} = \mathbf{U}\mathbf{D}^{1/2}\mathbf{U}^\top$ such that $\mathbf{B}^2 = \mathbf{A}$. It can also be written as $\mathbf{A} = \mathbf{B}\mathbf{B}^\top$ for some $\mathbf{B}$ (e.g., $\mathbf{B} = \mathbf{U}\mathbf{D}^{1/2}$).

Now, consider a projection matrix $\mathbf{P}_M$ onto a subspace $M \subseteq \mathbb{R}^n$ with $\dim(M) = m$.

- Since $\mathbf{P}_M$ is symmetric, it's orthogonally diagonalizable: $\mathbf{P}_M = \mathbf{U}\mathbf{D}\mathbf{U}^\top$.
- Since $\mathbf{P}_M$ is idempotent ($\mathbf{P}_M^2 = \mathbf{P}_M$), its eigenvalues λ must satisfy $\lambda^2 = \lambda$. This means the only possible eigenvalues are 0 or 1.
- The dimension of the image (column space) is the rank of the matrix, which equals the number of non-zero eigenvalues. Since $\dim(\mathrm{Im}(\mathbf{P}_M)) = \dim(M) = m$, there must be exactly m eigenvalues equal to 1 and $n - m$ eigenvalues equal to 0.
- Therefore, the diagonal matrix $\mathbf{D}$ in the spectral decomposition can be arranged as $\mathbf{D} = \mathrm{diag}(\underbrace{1, \ldots, 1}_{m \text{ times}}, \underbrace{0, \ldots, 0}_{n-m \text{ times}})$.

- The columns of $\mathbf{U}$ corresponding to the eigenvalue 1 form an orthonormal basis for the subspace $M = \mathrm{Im}(\mathbf{P}_M)$.
- The columns of $\mathbf{U}$ corresponding to the eigenvalue 0 form an orthonormal basis for the orthogonal complement $M^\perp = \ker(\mathbf{P}_M)$.
- Since all eigenvalues are ≥ 0, $\mathbf{P}_M$ is positive semidefinite. (It's positive definite only if $m = n$, in which case $\mathbf{P}_M = \mathbf{I}_n$).

This eigenvalue structure cleanly reflects the geometric action of projection.