Lecture Notes: Geometric Interpretation of Least Squares and Projection Matrices

Regression Class - March 31st

Administrative Notes

- Apologies for the slight delay at the start - navigating campus can be tricky sometimes!
- Reminder: Keep an eye out for information regarding an upcoming test/assessment.
- We took a short break during the lecture.
- The statistical model underlying regression (introducing error terms, assumptions, etc.) will be discussed in more detail starting next week. For now, our focus is algebraic and geometric.
- Some proof details (like Property 10 and the end of Property 11) are elaborated upon in the supplementary course notes.

1 Introduction: Revisiting Least Squares

We've been discussing the method of least squares for fitting a linear model. Recall our multiple linear regression setup: we have response data $y_1, \ldots, y_n$ and predictor data organized in a matrix X (an $n \times (p+1)$ matrix, including a column of ones for the intercept). We seek coefficients $b_0, b_1, \ldots, b_p$, collected in a vector $\mathbf{b} \in \mathbb{R}^{p+1}$, to minimize the sum of squared residuals. Today, we'll explore a powerful geometric interpretation of this minimization problem, which relies heavily on concepts from linear algebra.

2 Linear Algebra Review: Essential Concepts

Let's refresh some fundamental ideas from linear algebra. Assume we are working in the vector space $\mathbb{R}^k$ for some appropriate dimension k.

Definition 2.1 (Standard Inner Product)**.** For vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$, their standard inner product (or dot product) is:

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = \sum_{i=1}^{k} u_i v_i$$

Definition 2.2 (Norm)**.** The standard Euclidean norm (or length) of a vector $\mathbf{u} \in \mathbb{R}^k$ is induced by the inner product:

$$\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle} = \sqrt{\mathbf{u}^T \mathbf{u}} = \sqrt{\sum_{i=1}^{k} u_i^2}$$

Note that $\|\mathbf{u}\|^2 = \mathbf{u}^T \mathbf{u}$.

Definition 2.3 (Distance)**.** The Euclidean distance between two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$ is the norm of their difference:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\| = \sqrt{\sum_{i=1}^{k} (u_i - v_i)^2}$$

Definition 2.4 (Orthogonality)**.** Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$ are **orthogonal** if their inner product is zero. We denote this by $\mathbf{u} \perp \mathbf{v}$.

$$\mathbf{u} \perp \mathbf{v} \iff \langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^T \mathbf{v} = 0$$

In $\mathbb{R}^2$ and $\mathbb{R}^3$, this corresponds to the usual geometric notion of perpendicularity.

Definition 2.5 (Orthogonal Complement)**.** Let M be a subspace of $\mathbb{R}^k$. The **orthogonal complement** of M, denoted $M^\perp$, is the set of all vectors in $\mathbb{R}^k$ that are orthogonal to *every* vector in M.

$$M^\perp = \{\mathbf{v} \in \mathbb{R}^k \mid \mathbf{v}^T \mathbf{u} = 0 \text{ for all } \mathbf{u} \in M\}$$

It can be shown that $M^\perp$ is also a subspace of $\mathbb{R}^k$. To check if $\mathbf{v} \in M^\perp$, it suffices to check that $\mathbf{v}$ is orthogonal to every vector in a basis for M.

Theorem 2.6 (Pythagorean Theorem)**.** *If $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$ are orthogonal ($\mathbf{u} \perp \mathbf{v}$), then:*

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

Proof. We use the definition of the norm squared and the properties of the inner product (specifically, bilinearity and $\mathbf{u}^T \mathbf{v} = \mathbf{v}^T \mathbf{u} = 0$ due to orthogonality):

$$\begin{aligned}
\|\mathbf{u} + \mathbf{v}\|^2 &= (\mathbf{u} + \mathbf{v})^T (\mathbf{u} + \mathbf{v}) \\
&= (\mathbf{u}^T + \mathbf{v}^T)(\mathbf{u} + \mathbf{v}) \\
&= \mathbf{u}^T \mathbf{u} + \mathbf{u}^T \mathbf{v} + \mathbf{v}^T \mathbf{u} + \mathbf{v}^T \mathbf{v} \\
&= \|\mathbf{u}\|^2 + 0 + 0 + \|\mathbf{v}\|^2 \\
&= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2
\end{aligned}$$

□

3 The Least Squares Objective Function Geometrically

Recall the objective function we want to minimize in ordinary least squares (OLS):

$$Q(\mathbf{b}) = \sum_{i=1}^{n} (y_i - (b_0 x_{i0} + b_1 x_{i1} + \cdots + b_p x_{ip}))^2$$

where we assume $x_{i0} = 1$ for all i.

Let's rewrite this using matrix notation. Let $\mathbf{y} \in \mathbb{R}^n$ be the vector of responses, X be the $n \times (p+1)$ design matrix, and $\mathbf{b} \in \mathbb{R}^{p+1}$ be the vector of coefficients. The predicted value for the i-th observation is the i-th element of the vector $X\mathbf{b}$. Let $(X\mathbf{b})_i$ denote this element. Then:

$$(X\mathbf{b})_i = \sum_{j=0}^{p} X_{ij} b_j = b_0 x_{i0} + b_1 x_{i1} + \cdots + b_p x_{ip}$$

So, the objective function is:

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

Now, recognize the structure! This sum is exactly the squared Euclidean distance between the vector $\mathbf{y}$ and the vector $X\mathbf{b}$, both of which live in $\mathbb{R}^n$.

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

The problem of minimizing $Q(\mathbf{b})$ over all possible coefficient vectors $\mathbf{b} \in \mathbb{R}^{p+1}$ is therefore equivalent to finding the vector $X\mathbf{b}$ that is *closest* to the observed vector $\mathbf{y}$ in terms of squared Euclidean distance.

4 Geometric Interpretation: Projection onto a Subspace

Where do the vectors $X\mathbf{b}$ live? As $\mathbf{b}$ varies over all of $\mathbb{R}^{p+1}$, the resulting vector $X\mathbf{b}$ traces out the column space (or image) of the matrix X.

Definition 4.1 (Image or Column Space)**.** The **image** or **column space** of an $n \times m$ matrix A, denoted $\mathrm{Im}(A)$ or $C(A)$, is the set of all possible linear combinations of its columns. Equivalently,

$$\mathrm{Im}(A) = \{A\mathbf{v} \mid \mathbf{v} \in \mathbb{R}^m\}$$

$\mathrm{Im}(A)$ is a subspace of $\mathbb{R}^n$.

In our case, X is $n \times (p+1)$, so $\mathrm{Im}(X)$ is a subspace of $\mathbb{R}^n$. The dimension of this subspace is the rank of X. We typically assume that the columns of X are linearly independent, which means $rank(X) = p+1$. This requires $n \geq p+1$. Under this assumption, the columns of X form a basis for $\mathrm{Im}(X)$.

Our minimization problem can be rephrased: Find the vector $\mathbf{z} \in \mathrm{Im}(X)$ that minimizes $\|\mathbf{y} - \mathbf{z}\|^2$.

Geometrically, we have the vector $\mathbf{y}$ in $\mathbb{R}^n$ and the subspace $\mathrm{Im}(X)$ (which is typically a lower-dimensional subspace, like a plane or hyperplane within $\mathbb{R}^n$ unless $p+1 = n$). If $\mathbf{y}$ happens to already be *in* $\mathrm{Im}(X)$, the minimum distance is 0, achieved when $\mathbf{z} = \mathbf{y}$. This corresponds to a perfect fit with zero residuals.

More commonly, $\mathbf{y}$ does not lie in $\mathrm{Im}(X)$. Our geometric intuition tells us that the point $\mathbf{z}$ in the subspace $\mathrm{Im}(X)$ that is closest to $\mathbf{y}$ is the **orthogonal projection** of $\mathbf{y}$ onto that subspace.

Let $\mathbf{z}^*$ be this minimizing vector (the projection). Let $\hat{\mathbf{b}}$ be the coefficient vector such that $\mathbf{z}^* = X\hat{\mathbf{b}}$. Then $\hat{\mathbf{b}}$ is our least squares estimate. The vector $\hat{\mathbf{y}} = \mathbf{z}^* = X\hat{\mathbf{b}}$ is the vector of fitted values.

What characterizes this projection $\mathbf{z}^*$? The key geometric insight is that the **residual vector**, $\mathbf{e}^* = \mathbf{y} - \mathbf{z}^* = \mathbf{y} - X\hat{\mathbf{b}}$, must be **orthogonal** to the subspace $\text{Im}(X)$.

$$(\mathbf{y} - X\hat{\mathbf{b}}) \perp \text{Im}(X)$$

This means the residual vector must be orthogonal to *every* vector in $\text{Im}(X)$.

4.1 Deriving the Normal Equations Geometrically

If the residual vector $\mathbf{e}^* = \mathbf{y} - X\hat{\mathbf{b}}$ is orthogonal to the entire subspace $\text{Im}(X)$, it must be orthogonal to every vector that spans the subspace. In particular, it must be orthogonal to each column of X. Let X_j denote the j-th column of X (for $j = 0, \ldots, p$). Then:

$$\langle X_j, \mathbf{y} - X\hat{\mathbf{b}} \rangle = X_j^T(\mathbf{y} - X\hat{\mathbf{b}}) = 0 \quad \text{for all } j = 0, \ldots, p$$

We can write these $p + 1$ orthogonality conditions compactly using the full matrix X. The condition that all columns X_j are orthogonal to $\mathbf{e}^*$ is equivalent to:

$$X^T(\mathbf{y} - X\hat{\mathbf{b}}) = \mathbf{0}$$

where $\mathbf{0}$ is the zero vector in $\mathbb{R}^{p+1}$.

Rearranging this equation gives:

$$X^T\mathbf{y} - X^TX\hat{\mathbf{b}} = \mathbf{0}$$

$$\boxed{X^TX\hat{\mathbf{b}} = X^T\mathbf{y}}$$

These are precisely the **Normal Equations** we derived earlier using calculus!

Assuming the columns of X are linearly independent, the $(p + 1) \times (p + 1)$ matrix X^TX is invertible. We can then solve for $\hat{\mathbf{b}}$:

$$\hat{\mathbf{b}} = (X^TX)^{-1}X^T\mathbf{y}$$

This geometric derivation provides a beautiful alternative perspective on why the least squares solution takes this form.

5 The Projection Matrix

Let's look again at the vector of fitted values:

$$\hat{\mathbf{y}} = X\hat{\mathbf{b}} = X(X^TX)^{-1}X^T\mathbf{y}$$

Notice that we can obtain the fitted vector $\hat{\mathbf{y}}$ by multiplying the original data vector $\mathbf{y}$ by a specific matrix.

Definition 5.1 (Projection Matrix)**.** Let X be an $n \times m$ matrix with linearly independent columns (so $n \geq m$). The **projection matrix** (or hat matrix) that projects vectors orthogonally onto the column space $\text{Im}(X)$ is:

$$P_X = X(X^TX)^{-1}X^T$$

In our regression context, $m = p + 1$. P_X is an $n \times n$ matrix. Using this matrix, the fitted values are simply:

$$\hat{\mathbf{y}} = P_X \mathbf{y}$$

The matrix P_X takes any vector $\mathbf{y} \in \mathbb{R}^n$ and maps it to its orthogonal projection onto the subspace spanned by the columns of X.

6 Properties of Projection Matrices

Let X be an $n \times m$ matrix with linearly independent columns, and let $P_X = X(X^T X)^{-1} X^T$ be the projection matrix onto $\mathrm{Im}(X)$. P_X has several important properties:

Property 6.1 (Symmetry)**.** *P_X is symmetric, i.e., $P_X^T = P_X$.*

Proof. Recall that $(ABC)^T = C^T B^T A^T$ and $(A^{-1})^T = (A^T)^{-1}$. Also, $(X^T X)^T = X^T (X^T)^T = X^T X$, so $X^T X$ is symmetric.

$$\begin{aligned} P_X^T &= (X(X^T X)^{-1} X^T)^T \\ &= (X^T)^T ((X^T X)^{-1})^T X^T \\ &= X((X^T X)^T)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &= P_X \end{aligned}$$

□

Property 6.2 (Idempotence)**.** *P_X is idempotent, i.e., $P_X^2 = P_X P_X = P_X$. (Projecting a second time doesn't change anything, since the vector is already in the subspace).*

Proof.

$$\begin{aligned} P_X^2 &= (X(X^T X)^{-1} X^T)(X(X^T X)^{-1} X^T) \\ &= X(X^T X)^{-1} (X^T X)(X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} I (X^T X)^{-1} X^T \\ &= X(X^T X)^{-1} X^T \\ &= P_X \end{aligned}$$

□

Property 6.3 (Action on Columns of X)**.** *$P_X X = X$. (Projecting the columns of X onto their own span leaves them unchanged).*

Proof.

$$\begin{aligned} P_X X &= (X(X^T X)^{-1} X^T) X \\ &= X(X^T X)^{-1} (X^T X) \\ &= XI \\ &= X \end{aligned}$$

□

Property 6.4 (Orthogonality of Residual Projection)**.** *Let I be the $n \times n$ identity matrix. The matrix $I - P_X$ projects vectors onto the orthogonal complement subspace,* $(\mathrm{Im}(X))^{\perp}$*. We have $X^T(I - P_X) = 0$. (The columns of X are orthogonal to the space onto which $I - P_X$ projects).*

Proof.

$$\begin{aligned} X^T(I - P_X) &= X^T I - X^T P_X \\ &= X^T - X^T(X(X^T X)^{-1} X^T) \\ &= X^T - (X^T X)(X^T X)^{-1} X^T \\ &= X^T - I X^T \\ &= X^T - X^T = 0 \quad \text{(the } m \times n \text{ zero matrix)} \end{aligned}$$

□

Property 6.5 (Image of P_X)**.** *For any vector $\mathbf{v} \in \mathbb{R}^n$, the projected vector $P_X\mathbf{v}$ lies in the column space of X, i.e., $P_X\mathbf{v} \in \mathrm{Im}(X)$.*

Proof. Let $P_X\mathbf{v} = X(X^T X)^{-1} X^T \mathbf{v}$. Let $\mathbf{a} = (X^T X)^{-1} X^T \mathbf{v}$. This vector $\mathbf{a}$ is in $\mathbb{R}^m$. Then $P_X\mathbf{v} = X\mathbf{a}$. By definition, any vector of the form $X\mathbf{a}$ is a linear combination of the columns of X and is therefore in $\mathrm{Im}(X)$. □

Property 6.6 (Projection in Full Space)**.** *If X is an invertible $n \times n$ matrix (so $m = n$), then* $\mathrm{Im}(X) = \mathbb{R}^n$*. In this case, $P_X = I_n$. (Projecting onto the entire space is the identity operation).*

Proof. Since X is square and invertible, X^T is also invertible. $(X^T X)^{-1} = X^{-1}(X^T)^{-1}$.

$$\begin{aligned} P_X &= X(X^T X)^{-1} X^T \\ &= X(X^{-1}(X^T)^{-1}) X^T \\ &= (X X^{-1})((X^T)^{-1} X^T) \\ &= I_n I_n \\ &= I_n \end{aligned}$$

□

Property 6.7 (Projection onto Orthogonal Complement)**.** *For any $\mathbf{v} \in \mathbb{R}^n$, the vector $(I - P_X)\mathbf{v}$ lies in the orthogonal complement of the column space of X, i.e., $(I - P_X)\mathbf{v} \in (\mathrm{Im}(X))^{\perp}$. (This vector is the residual part of the projection).*

Proof. We need to show that $(I - P_X)\mathbf{v}$ is orthogonal to any vector $\mathbf{u}$ in $\mathrm{Im}(X)$. Any such $\mathbf{u}$ can be written as $\mathbf{u} = X\mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^m$. We compute the inner product:

$$\begin{aligned} \langle \mathbf{u}, (I - P_X)\mathbf{v} \rangle &= \mathbf{u}^T (I - P_X)\mathbf{v} \\ &= (X\mathbf{a})^T (I - P_X)\mathbf{v} \\ &= \mathbf{a}^T X^T (I - P_X)\mathbf{v} \\ &= \mathbf{a}^T (X^T (I - P_X))\mathbf{v} \\ &= \mathbf{a}^T (0)\mathbf{v} \quad \text{(using Property 4)} \\ &= 0 \end{aligned}$$

Since the inner product is zero for any $\mathbf{u} \in \mathrm{Im}(X)$, the vector $(I - P_X)\mathbf{v}$ must be in $(\mathrm{Im}(X))^{\perp}$. □

Remark 6.8. The vector $(I - P_X)\mathbf{y}$ is exactly the residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ from the least squares fit. This property confirms geometrically that the residual vector is orthogonal to the space spanned by the predictors (including the intercept).

Property 6.9 (Action on Vectors in Image)**.** *If* $\mathbf{w} \in \mathrm{Im}(X)$*, then* $P_X\mathbf{w} = \mathbf{w}$*. (Projecting a vector already in the subspace leaves it unchanged).*

Proof. If $\mathbf{w} \in \mathrm{Im}(X)$, then $\mathbf{w} = X\mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^m$.

$$\begin{aligned} P_X\mathbf{w} &= P_X(X\mathbf{a}) \\ &= (P_X X)\mathbf{a} \\ &= X\mathbf{a} \quad \text{(using Property 3)} \\ &= \mathbf{w} \end{aligned}$$

□

Property 6.10 (Action on Vectors in Orthogonal Complement)**.** *If* $\mathbf{w} \in (\mathrm{Im}(X))^\perp$*, then* $P_X\mathbf{w} = \mathbf{0}$*. (Vectors orthogonal to the subspace project to the zero vector).*

Proof. If $\mathbf{w} \in (\mathrm{Im}(X))^\perp$, it means $\mathbf{w}$ is orthogonal to every column of X. This is equivalent to $X^T\mathbf{w} = \mathbf{0}$.

$$\begin{aligned} P_X\mathbf{w} &= X(X^TX)^{-1}X^T\mathbf{w} \\ &= X(X^TX)^{-1}\mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

□

Property 6.11 (Independence of Basis)**.** *If* Z *is another* $n \times m$ *matrix such that* $\mathrm{Im}(Z) = \mathrm{Im}(X)$ *(i.e., the columns of* Z *span the same subspace as the columns of* X*, meaning* Z*'s columns are another basis for the same subspace), then* $P_Z = P_X$*. The projection matrix depends only on the subspace, not on the specific basis chosen to represent it.*

Proof. (Sketch, details in course notes) Requires showing that if $\mathrm{Im}(Z) = \mathrm{Im}(X)$, then $Z = XC$ for some invertible $m \times m$ matrix C, and $X = ZD$ for some invertible $m \times m$ matrix D (where $D = C^{-1}$). Substitute these into the formula for P_Z and simplify using properties of inverses and transposes to show it equals P_X. □

Remark 6.12. This is crucial: the geometric operation of projection onto a subspace is unique; the matrix representing it is unique, even though the subspace itself can be described by different sets of basis vectors (different matrices X or Z).

Property 6.13 (Nested Subspaces)**.** *Let* M *and* N *be two subspaces of* $\mathbb{R}^n$ *such that* $M \subseteq N$*. Let* P_M *and* P_N *be the projection matrices onto* M *and* N*, respectively. Then:*

$$P_M P_N = P_N P_M = P_M$$

Proof. (Intuition provided in lecture, formal proof involves basis matrices or properties like Property 8). Consider $P_N P_M \mathbf{v}$. First, $\mathbf{v}_M = P_M\mathbf{v}$ projects $\mathbf{v}$ onto M. Since $M \subseteq N$, the vector $\mathbf{v}_M$ is already in N. Projecting it again onto N leaves it unchanged (by Property 8 applied to P_N), so $P_N(P_M\mathbf{v}) = P_N\mathbf{v}_M = \mathbf{v}_M = P_M\mathbf{v}$. Thus $P_N P_M = P_M$.

Consider $P_M P_N \mathbf{v}$. First, $\mathbf{v}_N = P_N \mathbf{v}$ projects $\mathbf{v}$ onto N. Then we project $\mathbf{v}_N$ onto the subspace M. Since $M \subseteq N$, projecting onto N first and then the smaller subspace M is equivalent to just projecting onto M directly. (A more formal argument is needed here, often involving showing $(P_M P_N v - P_M v)$ is zero by showing it's in M and $M^\perp$). This leads to $P_M P_N = P_M$. Combining these gives the result. $\square$

7 Conclusion

Understanding least squares through the lens of orthogonal projection onto the column space of the design matrix provides deep insights. It explains the origin of the normal equations and introduces the projection matrix P_X (or hat matrix), a fundamental tool in regression diagnostics and theory. The properties of P_X directly reflect the geometric nature of the least squares fitting procedure.