Lecture 8: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia*

February 25, 2025 (last updated March 8, 2025)

Recap

Last time, we began discussing multivariate linear "regression", with $p > 1$ features, $\mathcal{X} = \mathbb{R}^p$, and $\mathcal{Y} = \mathbb{R}$. Given a data set $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, we were looking to predict using candidate functions that define a hyperplane. Recall this means we have an equation of the form

$$\begin{aligned} 0 &= w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_p x_p \\ &= \begin{bmatrix} 1 & x_1 & \cdots & x_p \end{bmatrix}^\top \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} \\ &= \mathbf{x}^\top \mathbf{w} \end{aligned}$$

Because of this, we overloaded our notation for the observations $\mathbf{x}_i$, so that instead of denoting a p-dimensional vector, it denotes a $(p+1)$-dimensional vector whose first entry is 1. Hence, for $\mathbf{x} \in \{1\} \times \mathbb{R}^p \subset \mathbb{R}^{p+1}$, the set of candidate functions is given by

$$\mathcal{H} = \{\mathbf{x}^\top \mathbf{w} = 0 : \mathbf{w} \in \mathbb{R}^{p+1}\}$$

We defined X to be a matrix of size $n \times (p+1)$, where the ith row consists of a 1 entry followed by the entries of $\mathbf{x}_i$ (the p features of the ith observation).

$$X := \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{bmatrix}$$

In other words, we make a matrix whose first column is $\vec{\mathbf{1}}_n$, and where every other observation in the feature space becomes a row. Note the jth column consists of the values of the jth feature across all n observations.

If $g \in \mathcal{H}$, then there is a fixed $\mathbf{w} \in \mathbb{R}^{p+1}$ such that $\hat{y}_i = g(\mathbf{x}_i) = \mathbf{w}^T \mathbf{x}_i$. Equivalently, since $\hat{y}_i$ is a scalar, and we think of $\hat{y}_i$ as the 1×1 matrix $[\hat{y}_i]$, then $\hat{y}_i = \hat{y}_i^\top$ (equals its transpose), so

$$\hat{y}_i = \hat{y}_i^\top = (\mathbf{w}^\top \mathbf{x}_i)^\top = \mathbf{x}_i^\top \mathbf{w}, \quad \forall i$$

*Based on lectures of Dr. Adam Kapelner at Queens College. See also the course GitHub page.

Recall that we can think of the ith row of the matrix product $A\mathbf{u}$ as the dot product of the ith row of A and $\mathbf{u}$. If we let $\hat{\mathbf{y}}$ be the prediction vector obtained by applying g to the $\mathbf{x}_i$ in $\mathbb{D}$, then

$$\hat{\mathbf{y}} := \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^\top \mathbf{w} \\ \vdots \\ \mathbf{x}_n^\top \mathbf{w} \end{bmatrix} = X\mathbf{w}$$

Define the response vector

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

Then the error (residual) is given by $\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}}$, and in order to obtain the best g (the least squares approximation), we need to minimize the SSE, which is given by

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \mathbf{e}^\top \mathbf{e} \\ &= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^T X\mathbf{w} + +\mathbf{w}^\top X^\top X\mathbf{w}. \end{aligned} \tag{1}$$

We will present some results from calculus and linear algebra that will help us tackle this problem.

Interlude of Linear Algebra and Calculus

Proposition 1. *Let $\mathbf{x} \in \mathbb{R}^n$, and let $a \in \mathbb{R}$ be a constant with respect to all entries of $\mathbf{x}$. Then*

$$\frac{\partial}{\partial \mathbf{x}}[a] := \begin{bmatrix} \frac{\partial}{\partial x_1}[a] \\ \frac{\partial}{\partial x_2}[a] \\ \vdots \\ \frac{\partial}{\partial x_n}[a] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}_n \tag{2}$$

Proposition 2. *Let $\mathbf{a} \in \mathbb{R}^n$, where all entries are constant with respect to all entries of $\mathbf{x}$. Then*

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{a} \cdot \mathbf{x}] = \frac{\partial}{\partial \mathbf{x}}[\mathbf{a}^\top \mathbf{x}] = \mathbf{a} \tag{3}$$

Proposition 3. *Let $\mathbf{x} \in \mathbb{R}^n$, $a, b \in \mathbb{R}$, and $f, g : \mathbb{R}^n \to \mathbb{R}$ be differentiable functions. Then*

$$\frac{\partial}{\partial \mathbf{x}}[af(\mathbf{x}) + bg(\mathbf{x})] = a \cdot \frac{\partial}{\partial \mathbf{x}}[f(\mathbf{x})] + b \cdot \frac{\partial}{\partial \mathbf{x}}[g(\mathbf{x})] \tag{4}$$

In other words, the derivative is linear for differentiable functions of several variables.

Proposition 4. *Let A be an $n \times n$ matrix ($A \in \mathbb{R}^{n \times n}$) that is symmetric (recall this means $a_{i,j} = a_{j,i}$ for all i, j), and whose entries are all constant with respect to $\mathbf{x} \in \mathbb{R}^n$. Then*

$$\frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^\top A\mathbf{x}] = 2 \cdot A\mathbf{x} \tag{5}$$

***Remark**: The expression $\mathbf{x}^\top A\mathbf{x}$ is a scalar, and the form of this product is known as a **quadratic form**. Perhaps the name is from the fact that it resembles $\frac{d}{dx}(ax^2) = 2ax$, where $a, x \in \mathbb{R}$.*

Minimizing the SSE in Multivariate Least Squares

We can now return to our main goal, which is a minimization problem

$$\mathcal{A} : \mathbf{b} = \underset{\mathbf{w} \in \mathbb{R}^{p+1}}{\operatorname{argmin}}\{SSE(\mathbf{w})\}$$

Consider Equation 1, and note the following observations:

- $\mathbf{y}^T\mathbf{y}$ is constant with respect to $\mathbf{w}$.
- Using the property that $(AB)^\top = B^\top A^\top$, we can write $\mathbf{y}^\top X = (X^\top \mathbf{y})^\top$.
- The matrix $X^\top X$ is symmetric: $(X^\top X)^\top = X^\top (X^\top)^\top = X^\top X$.

We proceed to taking the derivative with respect to $\mathbf{w} \in \mathbb{R}^{p+1}$ and using the equations from our interlude along with our observations above:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}[SSE] &= \frac{\partial}{\partial \mathbf{w}}[\mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top X\mathbf{w} + \mathbf{w}^\top X^\top X\mathbf{w}] \\
&\underbrace{=}_{\text{Eqn 4}} \underbrace{\frac{\partial}{\partial \mathbf{w}}[\mathbf{y}^\top \mathbf{w}]}_{\text{Eqn 2}} - 2\underbrace{\frac{\partial}{\partial \mathbf{w}}[(X^\top \mathbf{y})^\top \mathbf{w}]}_{\text{Eqn 3}} + \underbrace{\frac{\partial}{\partial \mathbf{w}}[\mathbf{w}^\top X^\top X\mathbf{w}]}_{\text{Eqn 5}} \\
&= \mathbf{0}_{p+1} - 2 \cdot X^\top \mathbf{y} + 2 \cdot X^\top X\mathbf{w}
\end{aligned}$$

If we set it to $\mathbf{0}_{p+1}$, then the solution is the vector $\mathbf{w} = \mathbf{b}$ that minimizes the SSE.

$$\begin{aligned}
\mathbf{0}_{p+1} &= -2 \cdot X^\top \mathbf{y} + 2 \cdot X^\top X\mathbf{b} \\
X^\top X\mathbf{b} &= X^\top \mathbf{y}
\end{aligned}$$

Suppose for now that the matrix $X^\top X$ of size $(p+1) \times (p+1)$ is invertible. Then the solution is given by

$$\mathbf{b} = (X^\top X)^{-1} X^\top \mathbf{y} \tag{6}$$

For $X^\top X$ to be invertible, it must have full-rank. Since $X^\top X$ has $p+1$ columns, this means we need $\text{rank}(X^\top X) = p+1$, where the rank denotes the dimension of the range of $X^\top X$. We will show an equivalent condition for $X^\top X$ to have full rank, but first here's a theorem from linear algebra:

Theorem 1 (Fundamental Theorem of Liner Maps, or Rank-Nullity Theorem)**.** Let V and W be vector spaces, where V is finite-dimensional, and let $T : V \to W$ be a linear transformation. Then

$$\dim \text{range}(T) + \dim \text{null}(T) = \dim V$$

See [Axl23][1] for a proof of Theorem 1. Now we can show:

Theorem 2. Given any matrix $\mathbf{X}$ of dimension $n \times (p+1)$, we have

$$\text{rank}(X^\top X) = \text{rank}(X)$$

Proof. Our strategy is to show that X^T and X have the same null space, and then use Theorem 1. Let $\mathbf{v} \in \text{null}(X^\top X)$, so that $(X^\top X)\mathbf{v} = \mathbf{0}_{p+1}$. Multiply by $\mathbf{v}^\top$ on the left to get:

$$\begin{aligned} \mathbf{v}^\top X^\top X \mathbf{v} &= \mathbf{v}^\top \mathbf{0}_{p+1} \\ (X\mathbf{v})^\top (X\mathbf{v}) &= 0 \end{aligned}$$

Notice this says that the dot product of $X\mathbf{v}$ with itself is 0, which implies $X\mathbf{v} = \mathbf{0}_n$. Hence, $\mathbf{v} \in \text{null}(X)$, proving that $\text{null}(X^\top X) \subseteq \text{null}(X)$.
For the other direction, suppose now that $\mathbf{v} \in \text{null}X$, so that $X\mathbf{v} = \mathbf{0}_n$. Then multiplying on the left by $X^\top$ gives $X^\top X\mathbf{v} = X^\top \mathbf{0}_n = \mathbf{0}_{p+1}$, so $\mathbf{v} \in \text{null}(X^\top X)$, and hence $\text{null}(X) \subseteq \text{null}(X^\top X)$.
We conclude that $\text{null}(X^\top X) = \text{null}(X)$, and hence their dimensions are the same. Now we will apply Theorem 1 to both X and $X^\top X$. Note that the domain space for both X and $X^\top X$ is $\mathbb{R}^{p+1}$, so

$$\begin{aligned} \dim \text{range}(X) + \dim \text{null}(X) &= \dim \mathbb{R}^{p+1} \\ \dim \text{range}(X^\top X) + \dim \text{null}(X^\top X) &= \dim \mathbb{R}^{p+1} \end{aligned}$$

If we subtract both equations, we arrive at $\dim \text{range}(X) = \dim \text{range}(X^\top X)$. Since the dimension of the range is precisely the rank, we are done. □

[1] See Open Access edition at https://linear.axler.net/LADR4e.pdf

From Theorem 2, we see that

$$\begin{aligned} X^\top X \text{is invertible} &\iff \text{rank}(X^T X) = p+1 \\ &\iff \text{rank}(X) = p+1 \\ &\iff \text{all columns of } X \text{ are linearly independent.} \end{aligned}$$

What does this last fact mean? It means that *none of the* $p+1$ *features are "redundant"*. For example, if we measured height in feet and in meters, and included a feature for each measurement, then our columns would be linearly dependent, and X would not be full rank (and hence $X^\top X$ is not invertible). In practice, we do not worry much about this because we can feed our matrix into a computer that can determine the rank and compute the inverse for us.

Performance Metrics for Multivariate Ordinary Least Squares

Let's revisit the error measures for OLS:

$$\begin{aligned} SSE &:= \mathbf{e}^\top \mathbf{e} \\ MSE &:= \frac{1}{n-(p+1)} SSE \\ RMSE &:= \sqrt{MSE} \\ R^2 &:= 1 - \frac{SSE}{SST} \end{aligned}$$

The only one that has changed appearance is the MSE. Notice that if $p = 1$, then we recover the simple OLS formula for MSE. The quantity $p+1$ is known as the **number of degrees of freedom (df)**. That is, $df = \text{rank}(X) = p+1$. An interesting observation is that if p is high, then the MSE increases, so the average squared error is penalized for having more features. We will explore this further later on.

Least Squares Estimates

We see that the least square estimate is given by Equation 6. Therefore, if $\mathbf{x}_* \in \mathbb{R}^{p+1}$ (recall we pre-pend a 1) is an incoming observation for which we would like to predict a response, we can do so via a dot product:

$$\hat{y}_* = g(\mathbf{x}_*) = \mathbf{x}_*^\top \mathbf{b}$$

More generally, if we have n_* observations, we can create an $n_* \times (p+1)$ matrix X_* similar to X (first column is $\vec{\mathbf{1}}_*$) where the ith row contains the ith observation, and we can predict for all values at once using matrix multiplication:

$$\hat{\mathbf{y}}_* = \begin{bmatrix} \hat{y}_{1,*} \\ \vdots \\ \hat{y}_{n_*,*} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{*,1}^\top \mathbf{b} \\ \vdots \\ \mathbf{x}_{*,n}^\top \mathbf{b} \end{bmatrix} = X_* \mathbf{b}$$

What if we don't have any features? That is, we have $p = 0$ and n observations, and the $n \times (p+1)$ matrix X only has the first column of 1's. In this case, $X = \vec{\mathbf{1}}_n$. Since $\mathbf{b}$ is

length $p+1$ and $p=0$, it has a single entry, and hence

$$\begin{aligned}\mathbf{b} &= b_0 \\ &= (X^\top X)^{-1} X^\top \mathbf{y} \\ &= (\vec{\mathbf{1}}_n^\top \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^\top \mathbf{y} \\ &= (n)^{-1} \sum_{i=1}^{n} y_i \\ &= \bar{y}\end{aligned}$$

Hence, the **null model** is given by

$$g_0(\mathbf{x}_*) = \bar{y} \quad \text{for all } \mathbf{x}_* \in \{1\} \times \mathbb{R}^p,$$

and it is a least squares estimate because we obtained it by using the least squares formula.

The Hat Matrix

In linear algebra we learn that a useful (and correct) way to think about matrix multiplication $A\mathbf{u}$ is as a linear combination of the columns of A using the entries in $\mathbf{u}$ as coefficients. For example if A is an $n \times p$ matrix, and $A_{\cdot,1}, \ldots, A_{\cdot,p}$ are the p columns of A, then

$$A\mathbf{u} = u_1 A_{\cdot,1} + \cdots + u_p A_{\cdot,p}$$

In particular, $A\mathbf{u} \in col(A) = \text{span}(A_{\cdot,1}, \ldots, A_{\cdot,p})$. Similarly, in the equation for the predictions $\hat{\mathbf{y}} = X\mathbf{b}$, we see that $\hat{\mathbf{y}}$ is a linear combinations of the $p+1$ columns of X. We are assuming that X is a $n \times (p+1)$ full rank matrix with $p+1 < n$, so full rank means that $\dim col(X) = p+1$. Therefore, it follows that $\hat{\mathbf{y}}$ belongs to a $(p+1)$-dimensional subspace of $\mathbb{R}^n$.

Let's expand the expression for $\hat{\mathbf{y}}$ using Equation 6:

$$\begin{aligned}\hat{\mathbf{y}} &= X\mathbf{b} \\ &= X(X^\top X)^{-1} X^\top \mathbf{y} \\ &= \underbrace{(X(X^\top X)^{-1} X^\top)}_{H} \mathbf{y} \\ &= H\mathbf{y} \end{aligned} \tag{7}$$

The $n \times n$ matrix H is called the **Hat matrix**. It turns out that $\text{rank}(H) = p+1$, which we can argue as follows.

Suppose that $\mathbf{v} \in \text{null}(H)$, so that $H\mathbf{v} = 0$. By associativity of matrix multiplication, this means $X((X^\top X)X^{-1}\mathbf{v}) = 0$. Since X has full rank $(p+1)$, this means it is injective, and hence $(X^\top X)X^{-1}\mathbf{v} = 0$. Similarly, associativity says this is equivalent to $(X^\top X)(X^\top \mathbf{v}) = 0$. By Theorem 2, $X^\top X$ also has rank $p+1$, so it is full rank and hence injective. Thus, we must have $X^\top \mathbf{v} = 0$. In other words, $\text{null}(H) \subseteq \text{null}(X^\top)$. It's easy to see that $\text{null}(X^\top) \subseteq \text{null}(H)$ by definition of H, so their null spaces are of the same size and hence $\dim \text{null}(H) = \dim \text{null}(X^\top)$. Since $\text{rank}(X^\top) = \text{rank}(X) = p+1$, we can invoke Theorem 1 to write

$$\begin{aligned}\text{rank}(X^\top) + \dim \text{null}(X^\top) &= n \\ \text{rank}(H) + \dim \text{null}(H) &= n\end{aligned}$$

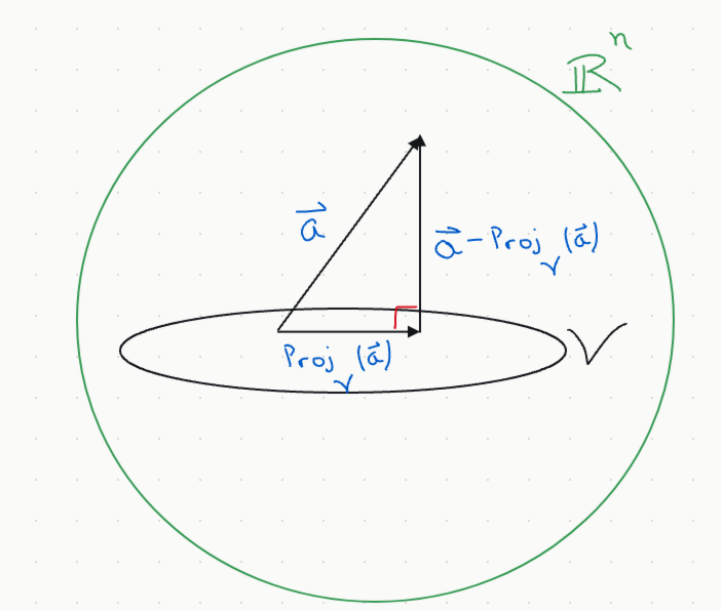

Figure 1: Orthogonal projection of a vector $\mathbf{a}$ in $\mathbb{R}^n$ onto a k-dimensional subspace V of $\mathbb{R}$.

Subtracting both equations, we arrive at $\text{rank}(H) = p+1$. Again, this hinges on X being $n \times (p+1)$, of rank $p+1$, and $p+1 < n$. In particular, H is not invertible. Later on, we consider the case where $p+1 > n$.

Orthogonal Projections

Let $\mathcal{V}$ be a k-dimensional subspace of $\mathbb{R}^n$, where $k \leq n$. Then

$$\mathcal{V} = \text{span}(\mathbf{v}_1, \mathbf{v}_2 \ldots, \mathbf{v}_k)$$

for $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, and the vectors form a linearly independent list. Let $\mathbf{a} \in \mathbb{R}^n$, and consider the projection $\mathbf{a}$ onto V, denoted $\text{proj}_{\mathcal{V}}(\mathbf{a})$, which is a vector in $\mathcal{V}$ such that the difference $\mathbf{a} - \text{proj}_V(\mathbf{a})$ is perpendicular to $\mathcal{V}$. Think of $\text{proj}_{\mathcal{V}}(\mathbf{a})$ as the shadow of $\mathbf{a}$ along $\mathcal{V}$. See Figure 1. This is called an **orthogonal projection**.

Since $\text{proj}_{\mathcal{V}}(\mathbf{a}) \in \mathcal{V}$, there are coefficients $w_1, \ldots, w_k$ such that

$$\text{proj}_{\mathcal{V}}(\mathbf{a}) = w_1\mathbf{v}_1 + w_2\mathbf{v}_2 + \cdots + w_k\mathbf{v}_k$$

Let V denote a matrix whose columns are the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$ that we said span $\mathcal{V}$ (in fact they form a basis), and $\mathbf{w}$ be a k-length vector from the w_i coefficients:

$$V = \begin{bmatrix} \vdots & \cdots & \vdots \\ \mathbf{v}_1 & \cdots & \mathbf{v}_k \\ \vdots & \cdots & \vdots \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}$$

Then we can express the projection as a matrix multiplication:

$$\text{proj}_{\mathcal{V}}(\mathbf{a}) = V\mathbf{w} \tag{8}$$

Moreover, since $\mathbf{a} - \text{proj}_{\mathcal{V}}(\mathbf{a}) \in \mathcal{V}^{\perp}$ (the orthogonal complement), it follows that for all $\mathbf{v}_j$ we have

$$\begin{aligned} 0 &= \mathbf{v}_j^{\top}(\mathbf{a} - \text{proj}_{\mathcal{V}}(\mathbf{a})) \\ &= \mathbf{v}_j^{\top}(\mathbf{a} - V\mathbf{w}) \end{aligned}$$

Since the equation above holds for all j, we can write it as a matrix multiplication:

$$V^{\top}(\mathbf{a} - V\mathbf{w}) = \mathbf{0}_k$$

To see why, note that since we think of $\mathbf{v}_j$ as a column vector, the jth row of $V^{\top}$ is $\mathbf{v}_j^{\top}$, and product of $V^{\top}$ and $(\mathbf{a} - V\mathbf{w})$ is precisely the product of each row of $V^{\top}$ with $\mathbf{a} - V\mathbf{w}$. Now we can expand and solve for $\mathbf{w}$:

$$\begin{aligned} V^{\top}\mathbf{a} - V^{\top}V\mathbf{w} &= \mathbf{0}_k \\ V^{\top}V\mathbf{w} &= V^{\top}\mathbf{a} \\ \mathbf{w} &= (V^{\top}V)^{-1}V^{\top}\mathbf{a} \end{aligned} \tag{9}$$

The last step requires that $V^{\top}V$ be invertible, and we argue that it is. To see why, recall that by Theorem 2, $\text{rank}[V^{\top}V] = \text{rank}[V]$, and since V is made up of k linearly independent vectors, $\mathbf{v}_1, \ldots, \mathbf{v}_k$, its rank is k. Hence, $\text{rank}[V^{\top}V] = k$, and it has size $k \times k$, so it is invertible.

Now that we have determined the coefficients $\mathbf{w}$ used to write $\text{proj}_{\mathcal{V}}(\mathbf{a})$ in terms of the basis of $\mathcal{V}$, we can combine Equations 8 and 9:

$$\text{proj}_{\mathcal{V}}(\mathbf{a}) = V(V^{\top}V)^{-1}V^{\top}\mathbf{a} \tag{10}$$

The matrix $V(V^{\top}V)^{-1}V^{\top}$ is called an **orthogonal projection matrix**. This resembles the Hat matrix in Equation 7. Indeed, this implies that the Hat matrix H is an orthogonal projection matrix, and given the expression $\hat{\mathbf{y}} = H\mathbf{y}$, it says that $\hat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of H, which in turn is the column space of X:

$$\hat{\mathbf{y}} = \underset{col[X]}{\text{proj}}(\mathbf{y})$$

See Figure 2. Recall the residual is defined as $\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}}$. Using Equation 7, we can write

$$\mathbf{e} := \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - H\mathbf{y} = (I - H)\mathbf{y} \tag{11}$$

Though we will not show it this time, the matrix $I - H$ is also an orthogonal projection matrix. This time, we are projecting onto the "residual space". See Figure 3. In fact, a theorem in linear algebra says that the space $col[X] \cap (col[X])^{\perp} = \{0\}$, and their sum is the entire space, so we have the direct sum:

$$col[X] \oplus (col[X])^{\perp} = \mathbb{R}^n$$

See Chapter 6 in [Axl23]. In particular, we imagine that if we put the bases for both of these space together, we get a fuller matrix as in Figure 4.

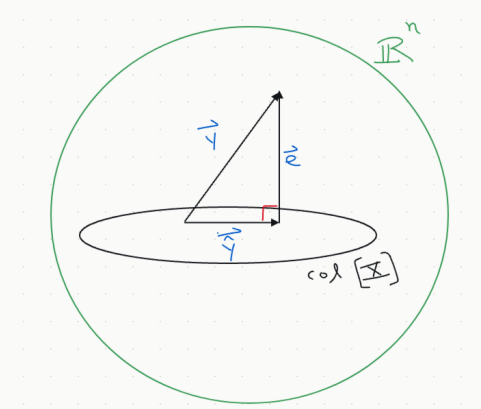

Figure 2: The prediction $\hat{\mathbf{y}}$ viewed as an orthogonal projection of the response vector $\mathbf{y}$ onto the column space of X.

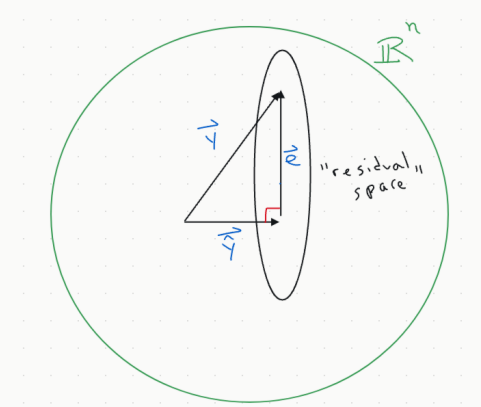

Figure 3: The residual $\mathbf{e}$ viewed as an orthogonal projection of the response vector $\mathbf{y}$ onto the "residual space".

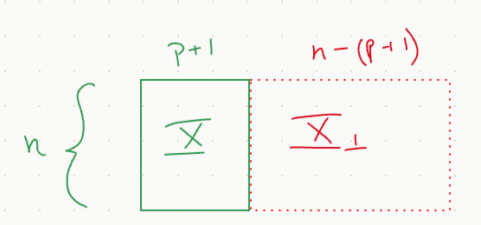

Figure 4: Matrix from putting the bases of $col[X]$ and $(col[X])^{\perp}$ together.

References

[Axl23] Sheldon Axler. *Linear Algebra Done Right*. 4th ed. Springer, 2023. ISBN: 9783031410253.