Lecture 9: MATH 342W: Introduction to Data Science and Machine Learning

Sergio E. Garcia Tapia*

February 27, 2025 (last updated March 14, 2025)

Recap

Last time, we saw that $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the column space of X (see Figure 1). We can express this as

$$\begin{aligned}\hat{\mathbf{y}} &= \underset{co\ell[X]}{\text{proj}}(\mathbf{y}) \\ &= X(X^\top X)^{-1}X^\top \mathbf{y} \\ &= H\mathbf{y}\end{aligned}$$

where H is the Hat matrix. What is $\underset{co\ell[X]}{\text{proj}}(\hat{\mathbf{y}})$? That is, what happens if we project $\hat{\mathbf{y}}$ onto the column space of X? Intuitively, it should be $\hat{\mathbf{y}}$, because $\hat{\mathbf{y}}$ is what we get from projecting $\mathbf{y}$ onto the column space of X, so $\hat{\mathbf{y}}$ is already in $co\ell[X]$. If this is the case, then we expect the following equalities to hold:

$$\begin{aligned}\underset{co\ell[X]}{\text{proj}}(\hat{\mathbf{y}}) &= \underset{co\ell[X]}{\text{proj}}(\underset{co\ell[X]}{\text{proj}}(\mathbf{y})) \\ &= H(H\mathbf{y}) \\ &= H^2\mathbf{y} \\ &= H\mathbf{y} \\ &= \hat{\mathbf{y}}\end{aligned}$$

Therefore, we conjecture that $H \cdot H = H$. Before we explore this idea, let's consider the null model again.

The Null Model in Multivariate Regression

In the null model g_0, we do not take any features into account, so $p = 0$. In this case, we expect g_0 to predict $\bar{y}$ for all inputs, as we saw when we first tackled OLS. Since $p = 0$ and there are n responses, the matrix X is of dimension $n \times (p+1)$ or $n \times 1$, where the

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

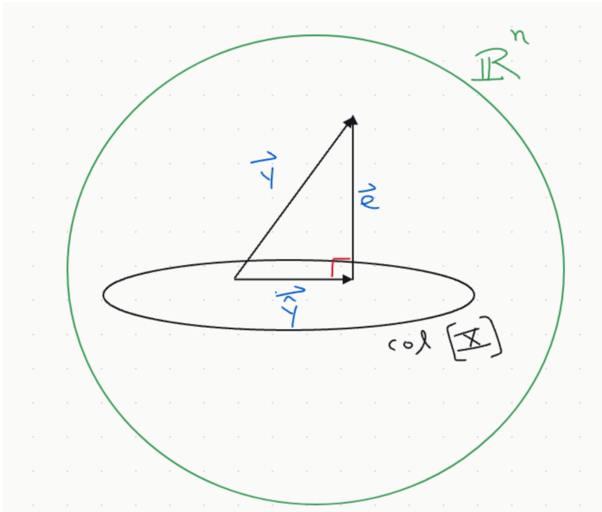

Figure 1: The prediction $\hat{\mathbf{y}}$ viewed as an orthogonal projection of the response vector $\mathbf{y}$ onto the column space of X.

first (and only) column has all 1's, like so:

$$X = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \vec{\mathbf{1}}_n$$

Now let's compute H:

$$\begin{aligned} H &= X(X^\top X)^{-1} X^\top \\ &= \vec{\mathbf{1}}_n (\vec{\mathbf{1}}_n^\top \vec{\mathbf{1}}_n)^{-1} \vec{\mathbf{1}}_n^\top \\ &= \vec{\mathbf{1}}_n \, (n)^{-1} \, \vec{\mathbf{1}}_n && (\vec{\mathbf{1}}_n^\top \vec{\mathbf{1}}_n = n) \\ &= \frac{1}{n} \vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^\top \end{aligned}$$

Notice that since $\vec{\mathbf{1}}_n$ is $n \times 1$ and $\vec{\mathbf{1}}_n^\top$ is $1 \times n$, the result of $\vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^\top$ is an $n \times n$ matrix. The product $\vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^\top$ is known as an *outer product*. It's easy to verify that the resulting matrix will have all 1's, so

$$\begin{aligned} H &= \frac{1}{n} \vec{\mathbf{1}}_n \vec{\mathbf{1}}_n^\top \\ &= \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \cdots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} && \text{(The matrix is } n \text{ by } n.\text{)} \\ &= \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \cdots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \end{aligned}$$

In particular, since every column is a scalar multiple of the first, the rank of this matrix is 1, i.e., $\text{rank}(H) = 1$, which is the number of columns in X (indeed, we saw last time that $\operatorname{rank}(H) = p+1$). Now we can compute the prediction:

$$\hat{\mathbf{y}} = H\mathbf{y} = \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \cdots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix} \mathbf{y} = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} y_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} y_i \end{bmatrix} = \bar{y} \cdot \vec{\mathbf{1}}_n$$

Hence, all n responses are predicted to be $\overline{y}$.

Properties of Orthogonal Projections

We have mentioned that H is an orthogonal projection. We will formally define what orthogonal projections are, and prove some useful properties that they satisfy.

Definition. A matrix P is an **orthogonal projection matrix** if and only if $\forall_{\mathbf{v},\mathbf{w}\in\mathbf{R}^n}$, we have

$$(\mathbf{v} - P\mathbf{v})^\top (P\mathbf{w}) = 0$$

That definition says that $P\mathbf{w}$ and $(\mathbf{v} - P\mathbf{v})$ are orthogonal for every $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$.

Theorem 1. A matrix P is an orthogonal projection if and only if the following two are both satisfied:

(i) **Symmetric**: $P^\top = P$.

(ii) **Idempotent**: $P^2 = P$ (it squares to itself).

Proof. We will prove the *if* direction ($\Longleftarrow$), and leave $\Longrightarrow$ as an exercise. Thus, we are assuming that $P^\top = P$ and $P^2 = P$. We have to show that it satisfies the definition of an orthogonal projection. Let $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Then

$$\begin{aligned}
(\mathbf{v} - P\mathbf{v})^\top P\mathbf{w} &= (\mathbf{v}^\top - \mathbf{v}^\top P^\top)P\mathbf{w} \\
&= \mathbf{v}^\top P\mathbf{w} - \mathbf{v}^\top P^\top \cdot P\mathbf{w} \\
&= \mathbf{v}^\top P\mathbf{w} - \mathbf{v}^\top P \cdot P\mathbf{w} && \text{(Symmetry: } P^\top = P) \\
&= \mathbf{v}^\top P\mathbf{w} - \mathbf{v}^\top P^2\mathbf{w} \\
&= \mathbf{v}^\top P\mathbf{w} - \mathbf{v}^\top P\mathbf{w} && \text{(Idempotency: } P^2 = P) \\
&= 0
\end{aligned}$$

□

Let's verify H satisfies the conditions of Theorem 1. First, we will check that $H^\top = H$:

$$\begin{aligned}
H^\top &= \left(X(X^\top X)^{-1}X^\top\right)^\top \\
&= (X^\top)^\top[(X^\top X)^{-1}]^\top X^\top && \text{(by } (AB)^\top = B^\top A^\top) \\
&= X[(X^\top X)^\top]^{-1}X^\top && \text{(by } (A^{-1})^\top = (A^\top)^{-1}) \\
&= X[X^\top X]^{-1}X^\top && \text{(since } (X^\top X)^\top = X^\top X) \\
&= H
\end{aligned}$$

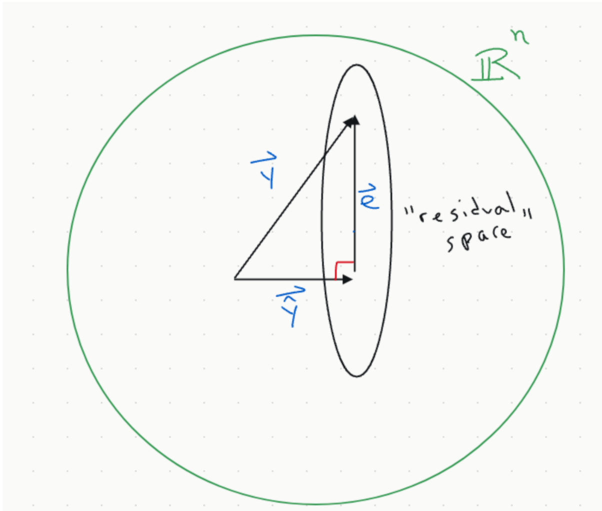

Figure 2: The residual $\mathbf{e}$ viewed as an orthogonal projection of the response vector $\mathbf{y}$ onto the "residual space".

Next, let's check that $H^2 = H$:

$$\begin{aligned}
H^2 &= H \cdot H \\
&= (X(X^\top X)^{-1}X^\top)(X(X^\top X)^{-1}X^\top) \\
&= X(X^\top X)^{-1} \underbrace{[(X^\top X)(X^\top X)^{-1}]}_{I_{p+1}} X^\top \\
&= X(X^\top X)^{-1}X^\top \\
&= H
\end{aligned}$$

Thus, H is indeed an orthogonal projection matrix.

Orthogonal Projection onto Residual Space

Now let's revisit an idea we mentioned last time, in which we said that $I_n - H$ is also an orthogonal projection matrix. This time, however, it is a projection onto the residual space (see Figure 2). To justify this, we must argue as in the case for H, by showing $I_n - H$ satisfies both conditions of Theorem 1. We will leverage the idempotency and symmetry of H:

(i) **Symmetric**: We must show $(I_n - H)^\top = (I_n - H)$:

$$(I_n - H)^\top = I_n^\top - H^\top = I_n - H$$

Note that the identify matrix is indeed symmetric, and we've used the fact that H is, too.

(ii) **Idempotent**: We must show $(I_n - H)^2 = (I_n - H)$:

$$\begin{aligned}
(I_n - H)^2 &= (I_n - H)(I_n - H) \\
&= I_n \cdot I_n - I_n \cdot H - H \cdot I_n + H^2 \\
&= I_n - H - H + H && (H \text{ is idempotent}) \\
&= I_n - H
\end{aligned}$$

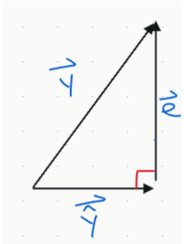

Figure 3: Orthogonal decomposition of response $\mathbf{y}$ into prediction $\hat{\mathbf{y}}$ and residual error $\mathbf{e}$.

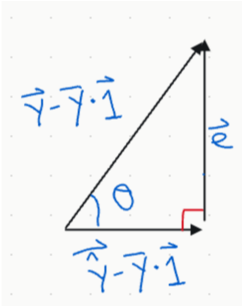

Figure 4: Orthogonal decomposition of mean-control response $\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n$ into $\hat{\mathbf{y}} - \bar{y} \cdot \vec{\mathbf{1}}_n$ and residual error $\mathbf{e}$.

SSR, R^2, and Geometric Interpretations

Consider again Figure 1. Since $\hat{\mathbf{y}}$ and $\mathbf{e}$ are orthogonal, note that Pythagorean's Theorem says that

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\mathbf{e}\|^2$$

See Figure 3. We will use this geometric intuition in a moment. Let's attempt to come up with a fit for $\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n$ (referred to as the *mean-control response*) by projecting it onto $co\ell[X]$:

$$\begin{aligned}
\operatorname*{proj}_{co\ell[X]}(\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n) &= H(\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n) \\
&= H\mathbf{y} - \bar{y} \cdot H\vec{\mathbf{1}}_n \\
&= \hat{\mathbf{y}} - \bar{y} \cdot \vec{\mathbf{1}}_n && (\vec{\mathbf{1}}_n \in co\ell[X] \implies \vec{\mathbf{1}}_n \in co\ell[H])
\end{aligned}$$

Now

$$(\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n) - (\hat{\mathbf{y}} - \bar{y} \cdot \vec{\mathbf{1}}_n) = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{e}$$

In particular, the two vectors being subtracted above are orthogonal (see Figure 4). Therefore, by Pythagorean's Theorem

$$\begin{aligned}
\|\mathbf{y} - \bar{y} \cdot \vec{\mathbf{1}}_n\|^2 &= \|\hat{\mathbf{y}} - \bar{y} \cdot \vec{\mathbf{1}}_n\|^2 + \|\mathbf{e}\|^2 \\
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2 \\
SST &= SSR + SSE \qquad (1)
\end{aligned}$$

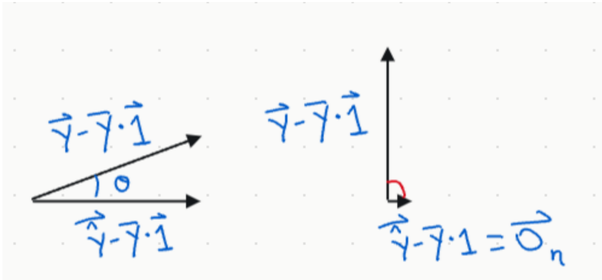

Figure 5: Depictions related to mean-control response in two cases: (i) high R^2, and (ii) $\theta = 90°$, where we do not beat g_0.

Definition. Given a n real numbers $(y_i)_{i=1}^n$ with mean $\bar{y}$ and associated predictions $(\hat{y}_i)_{i=1}^n$, the SSR is given by

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Using trigonometry, we can see that

$$\begin{aligned}\cos^2\theta &= \frac{\|\hat{\mathbf{y}} - \bar{y}\cdot\vec{\mathbf{1}}_n\|^2}{\|\mathbf{y} - \bar{y}\cdot\vec{\mathbf{1}}_n\|^2} \\ &= \frac{SSR}{SST} && \text{(by Equation 1)} \\ &= \frac{SST - SSE}{SST} \\ &= 1 - \frac{SSE}{SST} \\ &= R^2\end{aligned}$$

Since $\cos^2\theta \in [0,1]$, we see that $R^2 \in [0,1]$. Let's think about what this means. If your projection $\hat{\mathbf{y}}$ is close to the response vector $\mathbf{y}$, then the angle between them is small. On the other hand, if $\theta = 90°$, then there is no fit. In the latter case, $\hat{\mathbf{y}} - \bar{y}\cdot\vec{\mathbf{1}}$ is the zero vector, so $\hat{\mathbf{y}} = \bar{y}\vec{\mathbf{1}}_n$, and hence we did not do better than the null model g_0.

Reviewing Ignorance Error

Recall that ignorance error comes from the fact that the features (the x's) do not give sufficient information about the true drivers (the z's). We mentioned that a way to address that is by adding more features (increase p).

Let $X' = [X \mid \mathbf{x}_{\cdot\text{new}}]$, where we append a new column corresponding to a new feature that we measure. Then we expect that the error will decrease, and hence the angle between $\mathbf{y}$ and $\hat{\mathbf{y}}$ correspondingly decreases (see Figure 6). In the case where $p = 1$, we have $X = [\vec{\mathbf{1}} \mid \mathbf{x}_{\cdot,1}]$, so $col[X]$ is a 2-dimensional plane. See Figure 7.

Figure 6: Adding a new feature to X. We expect θ to decrease.

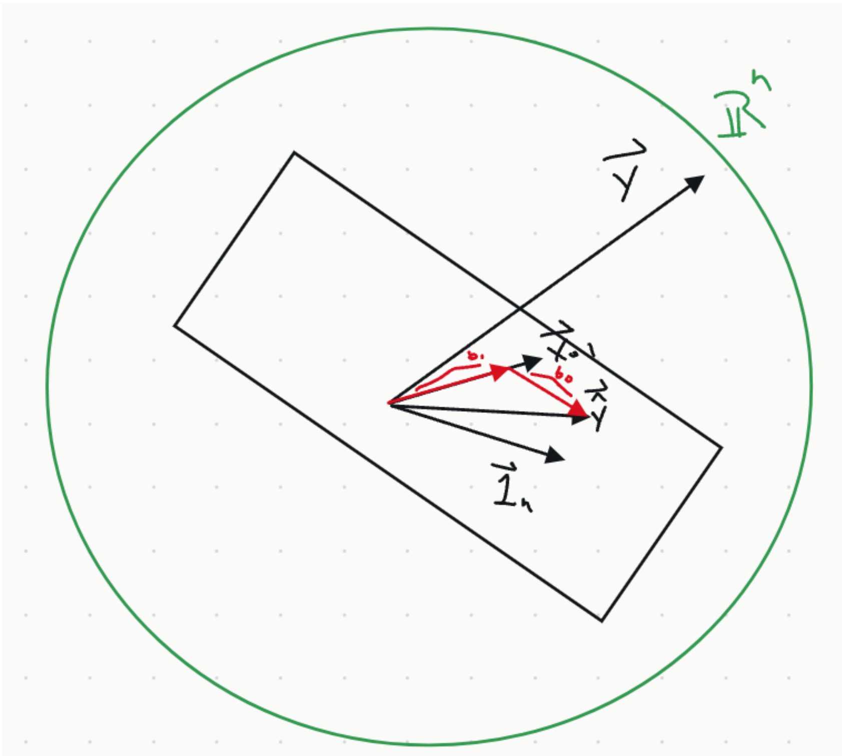

Figure 7: Illustration of least square when using 1 feature. Depicted is the portion of $\mathbf{y}$ along $\mathbf{x}_{\cdot 1}$ (the weight b_1) and the portion of $\mathbf{y}$ along $\vec{\mathbf{1}}_n$ (the intercept b_0)

Eigenvectors and Eigenvalues of H

The following material is outside the scope of this class, in the sense that you are not required to know it. Nevertheless, we will explore the concepts.

Recall that X has a column of 1's and p columns of features:

$$X = \begin{bmatrix} \vec{\mathbf{1}}_n & \mathbf{x}_{\cdot,1} & \cdots & \mathbf{x}_{\cdot,p} \end{bmatrix}$$

In particular, the columns of X are clearly in the column space of X. Recall that H is the orthogonal projection matrix onto the column space of X. This means that

$$H\vec{\mathbf{1}}_n = \vec{\mathbf{1}}_n, \quad H\mathbf{x}_{\cdot,1} = \mathbf{x}_{\cdot,1}, \quad \cdots, \quad H\mathbf{x}_{\cdot,p} = \mathbf{x}_{\cdot,p}$$

One simple way to verify this is as follows:

$$\begin{aligned} HX &= (X(X^\top X)^{-1}X^\top)X \\ &= X[(X^\top X)^{-1}(X^\top X)] \\ &= X \cdot I_{p+1} \\ &= X \end{aligned}$$

We conclude that $\lambda = 1$ is an eigenvalue of H, and the eigenspace associated with $\lambda = 1$ is spanned by $\mathbf{1}, \mathbf{x}_{\cdot,1}, \ldots, \mathbf{x}_{\cdot,p}$. Since H is symmetric (i.e., self-adjoint), the spectral theorem guarantees that it has an eigendecomposition (*diagonalization*) (see [Axl23]). Thus, we can write

$$H = P^{-1}DP$$

where

$$P = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{bmatrix}, \quad D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Here, D is a diagonal matrix consisting of the eigenvalues of H, and P is an invertible matrix whose columns are the eigenvectors of H. We have already argued that

$$\mathbf{v}_1 = \mathbf{1}_n, \quad \mathbf{v}_2 = \mathbf{x}_{\cdot,1}, \quad \ldots, \quad \mathbf{v}_{p+1} = \mathbf{x}_{\cdot,p}.$$
$$\lambda_1 = \lambda_2 = \cdots = \lambda_{p+1} = 1$$

What about the remaining $n - (p+1)$ eigenvectors? Note that if a vector belongs to $\mathit{col}[X]^\perp$ (the orthogonal complement of $\mathit{col}[X]$ or equivalently the residual space), then H maps it to 0. Therefore, the remaining $n - (p+1)$ eigenvalues of H are all zero, and the eigenvectors associated with the 0 eigenvalue span $\mathit{col}[X]^\perp$.

$$P = \begin{bmatrix} \vec{\mathbf{1}} & \mathbf{x}_{\cdot,1} & \cdots & \mathbf{x}_{\cdot,p} & \mathbf{x}_{\perp,\cdot,1} & \cdots & \mathbf{x}_{\perp,\cdot,n-(p+1)} \end{bmatrix}$$

$$D = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \ddots & 1 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & 0 & \ddots & \vdots \\ \vdots & \cdots & \cdots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \ddots & 0 \end{bmatrix}$$

One last fact is related to the trace of H. Recall this is the sum of the diagonal entries. We'll leverage the diagonalization:

$$\begin{aligned}
\sum_{i=1}^{n} h_{i,i} &= \mathrm{tr}[H] \\
&= \mathrm{tr}[P^{-1}DP] \\
&= \mathrm{tr}[PP^{-1}D] && (\mathrm{tr}[ABC] = \mathrm{tr}[CAB] = \mathrm{tr}[BCA]) \\
&= \mathrm{tr}[D] \\
&= p + 1 \\
&= \mathrm{rank}(X)
\end{aligned}$$

References

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