Lecture 3:

Least Squares

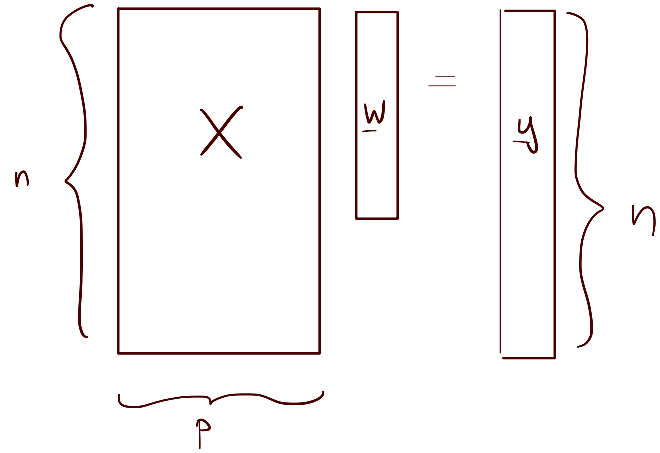
Given: vector of labels $\underline{y} \in \mathbb{R}^n$

matrix of features $X \in \mathbb{R}^{n \times p}$

Want: vector of weights $\underline{w} \in \mathbb{R}^p$

Assume: $n \geq p$

$\text{Rank}(X) = p$ (X has p linearly independent columns)



$\Longleftarrow$ n equations
p unknowns

Let $\underline{x}_i \in \mathbb{R}^p$ be feature vector of i^{th} sample; $= (i^{th} \text{ row of } X)^T$

Let $\underline{X}_j \in \mathbb{R}^n$ be j^{th} feature for all samples $= j^{th}$ col of X

If $\hat{\underline{y}} = X\underline{w}$, then we have a system of n linear equations;

i^{th} equation:

$$y_i = w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}$$

$$= \sum_{j=1}^{p} w_j x_{ij} = \langle \underline{w}, \underline{x}_i \rangle$$

↳ i^{th} row of X (transposed)

Side Note

(we might also consider $\hat{y} = \underline{\underline{w_0}} + w_1 x_{01} + w_2 x_{02} + \quad + w_p x_{0p}$

this can be done implicitly by letting $\underline{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0p} \end{bmatrix}$, $\underline{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p+1}$

now our model is $\hat{y} = \langle \underline{w}, \underline{x}_0 \rangle$, same as before!)

In general, $y \neq X\underline{w}$ for any $\underline{w}$ (because of modeling errors, noise)

Define residual $r_i = r_i(\underline{w}) = y_i - \langle \underline{w}, \underline{x}_i \rangle = y_i - \hat{y}_i \Rightarrow \underline{r} = [r_1\ r_2\ \cdots\ r_n]^T$

LEAST SQUARES ESTIMATION:

find $\underline{w}$ to minimize $\underbrace{\sum_{i=1}^{n} |r_i(\underline{w})|^2}_{\langle r, r \rangle = \|\underline{r}\|^2}$

Why least squares?

1) magnify effect of large errors

2) makes math easy

3) nice geometric interpretation

4) coincides with modeling $y = X\underline{w} + \underline{\varepsilon}$, $\underline{\varepsilon}$ = Gaussian noise (later)

SPOILER ALERT!

best $\underline{w}$ (that minimizes sum of squared errors) is

$$\hat{\underline{w}} = (X^T X)^{-1} X^T y$$

and corresponding predicted labels are $\hat{y} = X\hat{\underline{w}} = X(X^T X)^{-1} X^T y$

Today: when is this $\hat{\underline{w}}$ valid?

what is $(X^T X)^{-1}$?

where does this formula come from?

Span

The span of a set of vectors $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p \in \mathbb{R}^n$ is the set of vectors that can be written as a weighted sum of the $\underline{X}_j$'s:

$$\mathrm{Span}(\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p) = \left\{ \underline{y} \in \mathbb{R}^n : \underline{y} = \sum_{i=1}^{p} w_i \underline{X}_i \text{ for some } w_1, \ldots, w_p \in \mathbb{R} \right\}$$

If $X = [\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_p]$, then $\mathrm{range}(X) := \mathrm{span}(\text{cols of } X) = \mathrm{span}(\underline{x}_1, \ldots, \underline{x}_p)$

Ex. $\underline{X}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\underline{X}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. $\mathrm{span}(\underline{X}_1, \underline{X}_2)$ = vectors of form $\begin{bmatrix} \alpha \\ \beta \\ 0 \end{bmatrix}$ for some α, β

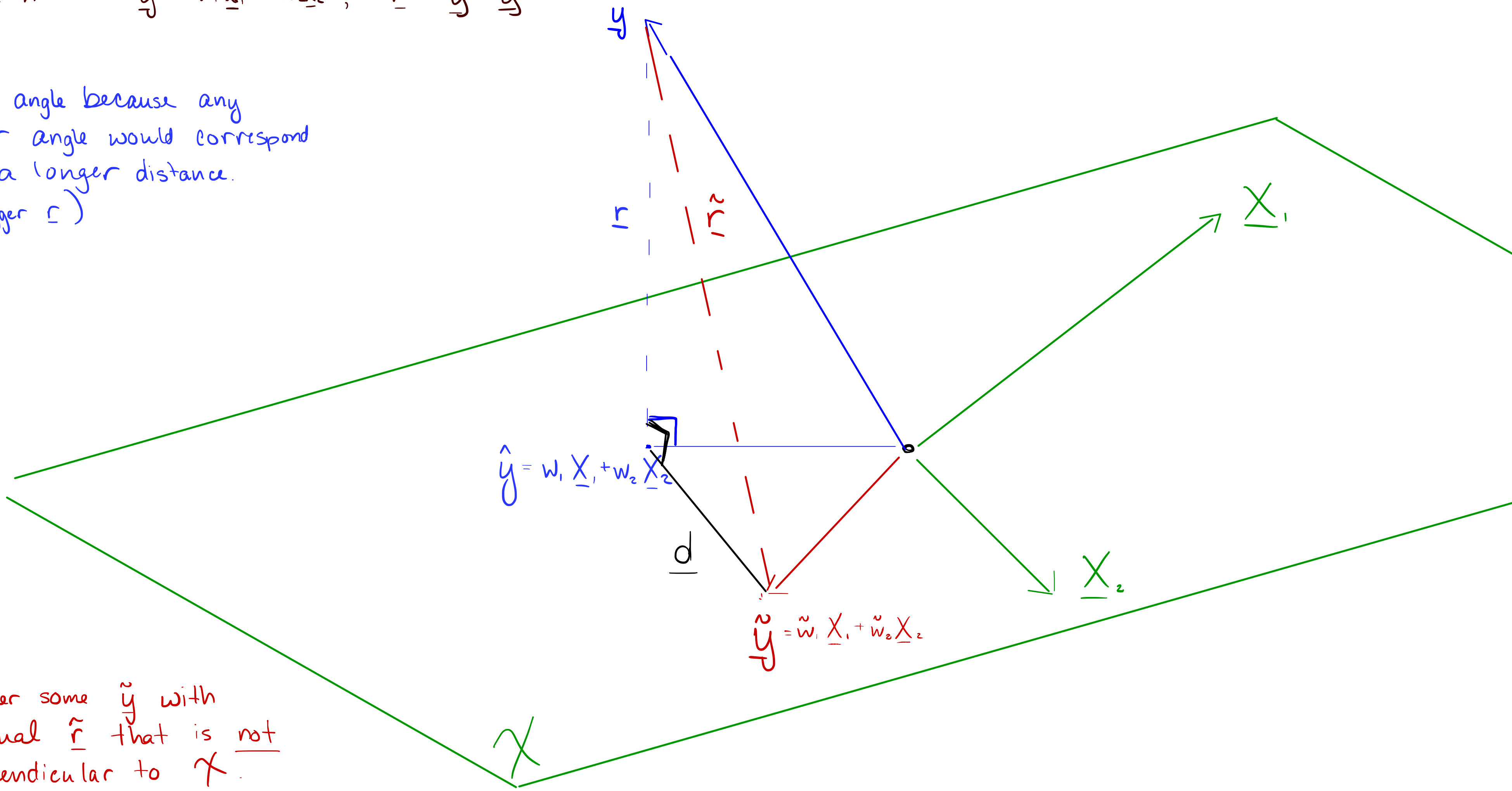
i.e. vectors with zero in 3^{rd} coordinate

Geometry of Least Squares

for general p, $\hat{y} = w_1 \underline{X}_1 + w_2 \underline{X}_2 + \cdots + w_p \underline{X}_p$

$p = 2$. $n = 3$. $\hat{\underline{y}} = w_1 \underline{X}_1 + w_2 \underline{X}_2$, $\underline{r} = \underline{y} - \hat{\underline{y}}$

Right angle because any other angle would correspond to a longer distance. (bigger $\underline{r}$)



Consider some $\tilde{y}$ with residual $\tilde{\underline{r}}$ that is <u>not</u> perpendicular to $\mathcal{X}$.

Then $\|\tilde{r}\|^2 = \|r\|^2 + \|d\|^2$ by Pythagorean Thm

and so $\|\tilde{r}\|^2 > \|r\|^2$ so weights corresponding to $\tilde{y}$ cannot be optimal (they don't make $\|\tilde{r}\|^2$ as small as possible)

$\mathcal{X}$ = space of all vectors $\hat{y}$ that can be written as $\hat{\underline{y}} = \alpha \underline{X}_1 + \beta \underline{X}_2$ for some $\alpha, \beta \in \mathbb{R}$. called "span" of cols of X. $\underline{y}$ may not be in this space

$\hat{\underline{w}}$ = "argument $\underline{w}$ that minimizes" $\sum_{i=1}^{n} r_i^2(\underline{w})$

$$= \underset{\underline{w}}{\arg\min} \sum_{i=1}^{n} r_i^2(\underline{w}) = \underset{\underline{w}}{\arg\min} \langle r, r \rangle = \underset{w}{\arg\min} \|r\|_2^2$$

Let $\underline{r} := r(\hat{\underline{w}}) = \begin{bmatrix} r_1(\hat{\underline{w}}) \\ \vdots \\ r_n(\hat{\underline{w}}) \end{bmatrix}$

We know that $\underline{r} = \underline{y} - X\hat{\underline{w}}$ is perpendicular/orthogonal to span of columns of X

This implies $\underline{X}_i^T \underline{r} = 0$ for each column i of X

Two vectors $\underline{u}, \underline{v}$ are orthogonal if $\langle \underline{u}, \underline{v} \rangle = 0$

$\Longrightarrow$ $X^T \underline{r} = \underline{0}$ $\Leftarrow$ vector of zeros

$\Longrightarrow$ $X^T(\underline{y} - X\hat{\underline{w}}) = \underline{0}$

$\Longrightarrow$ $\hat{w}$ is solution to linear system of equations $X^T\underline{y} = X^T X \underline{w}$

WHAT CAN WE SAY ABOUT $\underline{w}$ SATISFYING $X^T\underline{y} = X^T X \underline{w}$?

- does it exist ?
- is it unique ?

↳ "yes" to both when columns of X are <u>linearly independent</u>

Consider following linear systems. For each, how many solutions are there? (zero, one, or many) If one or more solutions exist, find one or more. Why do different cases have different numbers of solutions?

a) $3w_1 + 2w_2 = 1$
$3w_1 + w_2 = 0$

$\Rightarrow w_2 = -3w_1$
$-w_2 + 2w_2 = w_2 = 1$
$\Rightarrow w_1 = -\frac{1}{3} \Rightarrow$ one soln

rank(X) = 2

$$\underbrace{\begin{bmatrix} 3 & 2 \\ 3 & 1 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}}_{w} = \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix}}_{y}$$

---

b) $$\begin{bmatrix} 3 & 2 \\ 3 & 1 \\ 3 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

$\Rightarrow$ 1 soln: $x_1 = -\frac{1}{3}$, $x_2 = 1$

---

c) $3x_1 + 2x_2 = 1$
$3x_1 + x_2 = 0$
$2x_1 + 2x_2 = 2$

$x_1 = -\frac{1}{3}$, $x_2 = 1$

$2(-\frac{1}{3}) + 2(1) \neq 2$

$\Rightarrow$ 0 solutions

$$\begin{bmatrix} 3 & 2 \\ 3 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

---

d) $3x_1 + x_2 = 0$

∞ solutions

---

e) $3x_1 + x_2 = 1$
$6x_1 + 2x_2 = 2$

$$\Longleftrightarrow \begin{bmatrix} 3 & 1 \\ 6 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$\Rightarrow \infty$ solns

rank-1 matrix!

if X is $n \times p$, $p \leq n$, and rank$(X) < p$,
then $X^T X$ not invertible AND no unique solution

Linear Independence

A collection of vectors $\underline{V}_1, \underline{V}_2, \ldots, \underline{V}_p \in \mathbb{R}^n$ is linearly independent when

$$\sum_{i=1}^{p} \alpha_i \underline{V}_i = 0 \quad \text{if and only if} \quad \alpha_i = 0 \text{ for } i = 1, 2, \ldots, p$$

That is, any weighted sum of the vectors is nonzero unless all the weights are zero

Ex $\quad n = 3, \; p = 2 \quad \underline{V}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad V_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \Rightarrow$ yes, linearly undependent (LI)

note $\quad \alpha_1 \underline{V}_1 + \alpha_2 \underline{V}_2 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_2 \end{bmatrix}$ This can only be the zero vector if $\alpha_1 = \alpha_2 = 0$

Ex $\quad n = 3, \; p = 3 \quad \underline{V}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad V_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad V_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \Rightarrow$ yes, linearly undependent (LI)

$\alpha_1 \underline{V}_1 + \alpha_2 \underline{V}_2 + \alpha_3 \underline{V}_3 = \begin{bmatrix} \alpha_1 \\ \alpha_2 + \alpha_3 \\ \alpha_2 \end{bmatrix}$ this $= 0$ only if $\alpha_1 = \alpha_2 = \alpha_3 = 0$

Ex $n=3$, $p=4$ $\quad \underline{V}_1 = \begin{bmatrix}1\\0\\0\end{bmatrix}$, $\underline{V}_2 = \begin{bmatrix}0\\1\\1\end{bmatrix}$, $\underline{V}_3 = \begin{bmatrix}0\\1\\0\end{bmatrix}$, $\underline{V}_4 = \begin{bmatrix}1\\0\\1\end{bmatrix}$

note $\quad V_4 = \underline{V}_1 + \underline{V}_2 - \underline{V}_3$ (we can write one vector as linear combination (weighted sum) of others This implies linear dependence

$$\alpha_1 \underline{V}_1 + \alpha_2 \underline{V}_2 + \alpha_3 \underline{V}_3 + \alpha_4 \underline{V}_4 = \begin{bmatrix}\alpha_1 + \alpha_4\\ \alpha_2 + \alpha_3\\ \alpha_2 + \alpha_4\end{bmatrix} \longrightarrow \text{if } \alpha_1 = -\alpha_4 = \alpha_2 = -\alpha_3, \text{ then}$$

$$\alpha_1 \underline{V}_1 + \alpha_2 \underline{V}_2 + \alpha_3 \underline{V}_3 + \alpha_4 \underline{V}_4 = 0$$

$\Longrightarrow$ NOT linearly independent

Linear independence $\Longrightarrow$ $n \geq p$

$p > n$ $\Longrightarrow$ Linear dependence

Matrix rank number of linearly independent columns = # linearly independent rows

if $X^T = [\underline{x}_1 \quad \underline{x}_2 \quad \quad \underline{x}_n] \in \mathbb{R}^{p \times n}$, then $\text{rank}(X) \leq \min(p, n)$

Ex $\quad X = \begin{bmatrix}1 & 0\\0 & 1\\0 & 1\end{bmatrix} \Rightarrow \text{rank}(X) = 2$

Ex $\quad X = \begin{bmatrix}1 & 0 & 1\\0 & 1 & 1\\0 & 1 & 1\end{bmatrix} \Rightarrow \text{rank}(X) = 2$

Matrix Inverse

For a square matrix A, its inverse A^{-1} is a square matrix that satisfies:

$$AA^{-1} = A^{-1}A = I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$$

Ex. $A = \begin{bmatrix} 1/4 & 0 \\ 0 & 2 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} 4 & 0 \\ 0 & 1/2 \end{bmatrix}$

Ex. $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \Rightarrow A^{-1} = \begin{bmatrix} -2 & 1 \\ 3/2 & -1/2 \end{bmatrix}$

Not all matrices have inverses.

Specifically: $\boxed{\text{A only has an inverse if it is \underline{full rank}}}$

Ex: $A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, $A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$ have no inverse.

$$\boxed{X \in \mathbb{R}^{n \times p},\ n \geq p,\ \text{rank}(X) = p \Rightarrow \text{rank}(X^T X) = p \Rightarrow X^T X \text{ has inverse}}$$

Recall from earlier:

$\hat{\underline{w}}$ = "argument $\underline{w}$ that minimizes" $\sum_{i=1}^{n} r_i^2(\underline{w}) = \underset{\underline{w}}{\text{argmin}} \sum_{i=1}^{n} r_i^2(\underline{w})$

Let $\hat{\underline{r}} := r(\hat{\underline{w}}) = \begin{bmatrix} r_1(\hat{\underline{w}}) \\ \vdots \\ r_n(\hat{\underline{w}}) \end{bmatrix}$

We know that $\hat{\underline{r}} = \underline{y} - X\hat{\underline{w}}$ is perpendicular / orthogonal to span of columns of X

This implies $\underline{X}_i^T \hat{\underline{r}}(\hat{w}) = 0$ for each column i of X

Two vectors $\underline{u}, \underline{v}$ are orthogonal if $\langle \underline{u}, \underline{v} \rangle = 0$

$\Longrightarrow X^T \hat{\underline{r}}(\hat{w}) = \underline{0}$ ⇐ vector of zeros

$\Longrightarrow X^T(\underline{y} - X\hat{\underline{w}}) = \underline{0}$

$\Longrightarrow \hat{\underline{w}}$ is solution to linear system of equations $X^T\underline{y} = X^T X \underline{w}$

So if $X^T X$ is invertible (i.e. if $X^T X$ is full-rank, which occurs if $\text{rank}(X) = p < n$) then there is a unique solution:

$$\hat{\underline{w}} = (X^T X)^{-1} X^T \underline{y}$$

$$\Longrightarrow \hat{\underline{y}} = X\hat{\underline{w}}$$

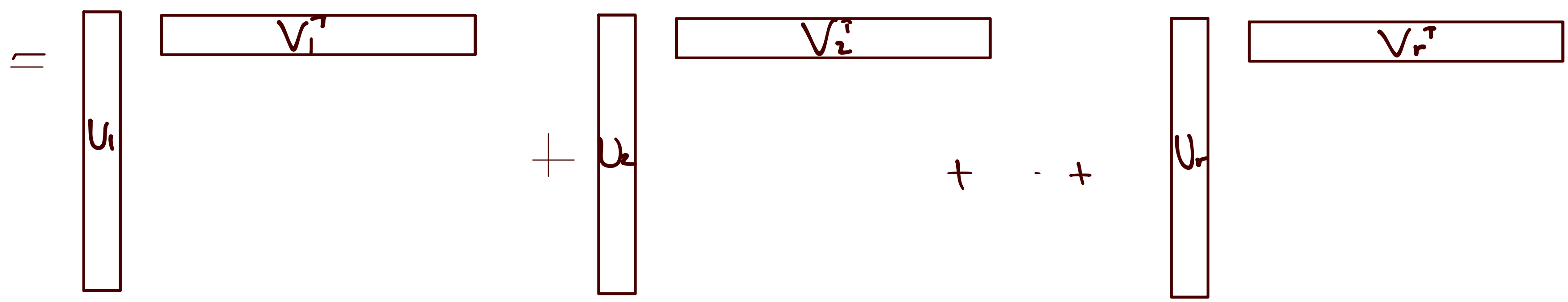
$$= X(X^T X)^{-1} X^T \underline{y}$$

$P_X := X(X^T X)^{-1} X^T$ is called a Projection Matrix because $P_X \underline{y}$ projects $\underline{y}$ onto range(X)

Ex $\quad X = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \begin{bmatrix} v_1 & v_2 & v_3 \end{bmatrix} = \begin{bmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 \\ u_2 v_1 & u_2 v_2 & u_2 v_3 \\ u_3 v_1 & u_3 v_2 & u_3 v_3 \end{bmatrix} \longrightarrow \text{Rank}(X) = 1$

Recall the outer product representation of matrix product

$$UV = \begin{bmatrix} | & | & & | \\ U_1 & U_2 & & U_r \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & V_1^T & - \\ - & V_2^T & - \\ & \vdots & \\ - & V_r^T & - \end{bmatrix}$$



$= \sum_{k=1}^{r} U_k V_k^T =$ sum of rank-1 matrices $\Longrightarrow \text{rank}(UV) = r$ if U_k's are LI and V_k's are LI