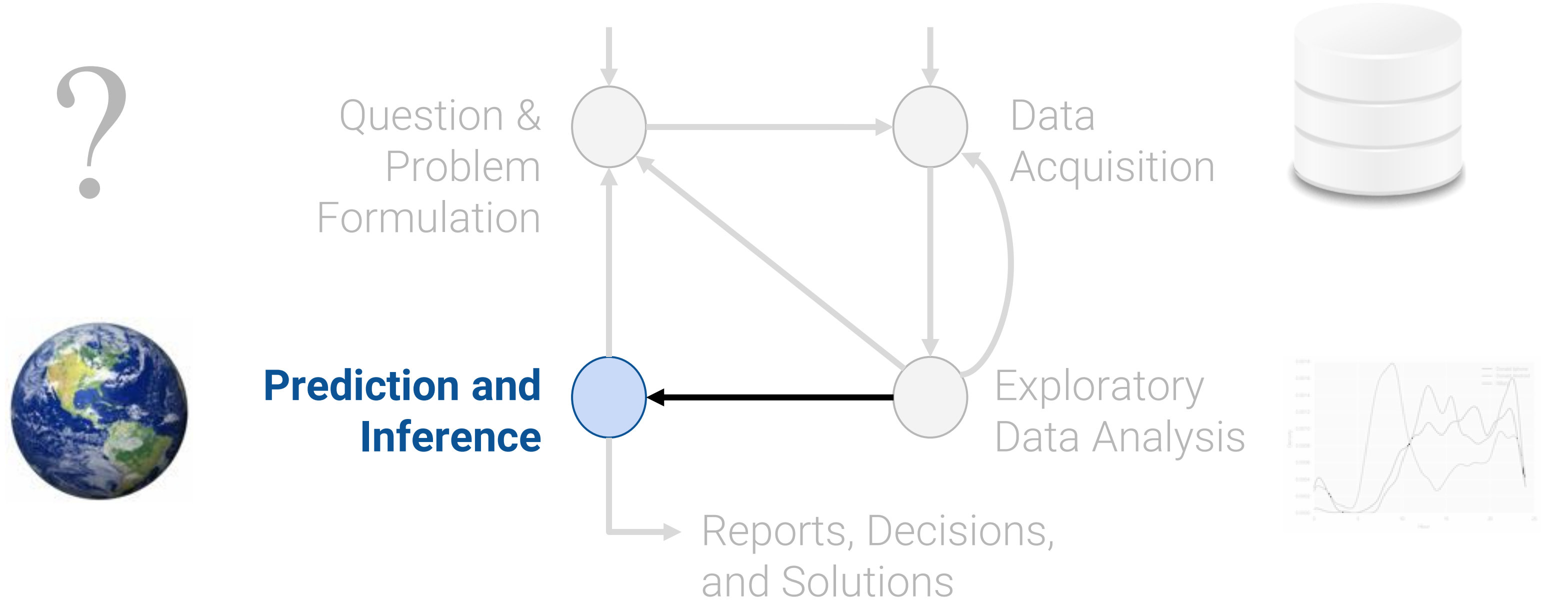
LECTURE 11

Ordinary Least Squares

Using linear algebra to derive the multiple linear regression model.

Plan for Next Few Lectures: Modeling



(today)

Modeling I: Intro to Modeling, Simple Linear Regression	→	Modeling II: Different models, loss functions, linearization	→	**Modeling III: Multiple Linear Regression**

Today's Roadmap

OLS Problem Formulation

- Multiple Linear Regression Model
- Mean Squared Error

Geometric Derivation

Performance: Residuals, Multiple R^2

OLS Properties

- Residuals
- The Bias/Intercept Term
- Existence of a Unique Solution

Multiple Linear Regression Model

OLS Problem Formulation

- **Multiple Linear Regression Model**
- Mean Squared Error

Geometric Derivation

Performance: Residuals, Multiple R^2

OLS Properties

- Residuals
- The Bias/Intercept Term
- Existence of a Unique Solution

Multiple Linear Regression

Define the **multiple linear regression** model:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

Predicted value of y

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \longrightarrow \boxed{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}} \longrightarrow \hat{y}$$

single **observation**
(p features)

single
prediction

NBA 2018-2019 Dataset

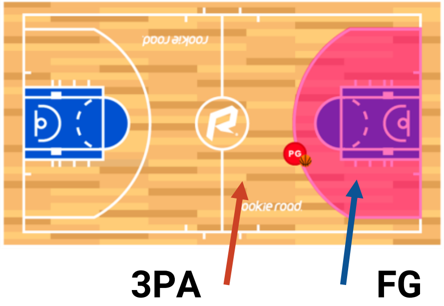
How many points does an athlete score per game?
PTS (average points/game)

To name a few factors:

- **FG**: average # 2 point field goals
- **AST**: average # of assists
- **3PA**: average # 3 point field goals attempted

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Rows correspond to individual players.



assist: a pass to a teammate that directly leads to a goal

Multiple Linear Regression Model

How many points does an athlete score per game?
PTS (average points/game)

To name a few factors:

- **FG**: average # 2 point field goals
- **AST**: average # of assists
- **3PA**: average # 3 point field goals attempted

3PA FG

assist: a pass to a teammate that directly leads to a goal

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
6	0.6	0.3	1.2	1.7

Rows correspond to individual players.

$$\boxed{\begin{matrix}0.4\\0.8\\1.5\end{matrix}} \rightarrow \boxed{\begin{bmatrix}\theta_0\\\theta_1\\\vdots\\\theta_p\end{bmatrix}} \rightarrow \boxed{\hat{y}}$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$
$$= \theta_0 + \theta_1 \underset{\text{FG}}{x_1} + \theta_2 \underset{\text{AST}}{x_2} + \theta_3 \underset{\text{3PA}}{x_3}$$

Today's Goal: Ordinary Least Squares

1. Choose a model	**Multiple Linear Regression**	
2. Choose a loss function	L2 Loss **Mean Squared Error (MSE)**	
3. Fit the model	Minimize average loss with ~~calculus~~ geometry	
4. Evaluate model performance	Visualize, ~~Root MSE~~ Multiple R^2	

In statistics, this model + loss is called **Ordinary Least Squares (OLS)**.

The solution to OLS are the minimizing loss for parameters $\hat{\theta}$, also called the **least squares estimate**.

Today's Goal: Ordinary Least Squares

1. Choose a model

Multiple Linear Regression

For each of our n data points:

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

3. Fit the model

Minimize average loss with ~~calculus~~ geometry

Linear Algebra!!

4. Evaluate model performance

Visualize,
~~Root MSE~~
Multiple R^2

From one feature to many features

Dataset for SLR

x	y
x_1	y_1
x_2	y_2
$\vdots$	$\vdots$
x_n	y_n

	FG	PTS
1	1.8	5.3
2	0.4	1.7
3	1.1	3.2
4	6.0	13.9
5	3.4	8.9
...	...	...

Dataset for Constant Model

y
y_1
y_2
$\vdots$
y_n

	PTS
1	5.3
2	1.7
3	3.2
4	13.9
5	8.9
...	...

Dataset for Multiple Linear Regression

$x_{:,1}$	$x_{:,2}$		$x_{:,p}$	y
x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
...	...	...	...	...

From one feature to many features

Dataset for Multiple Linear Regression

$x_{:1}$	$x_{:2}$		$x_{:p}$	y
x_{11}	x_{12}	$\ldots$	x_{1p}	y_1
x_{21}	x_{22}	$\ldots$	x_{2p}	y_2
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_{n1}	x_{n2}	$\ldots$	x_{np}	y_n

Observation i

$$\{x_{i1}, x_{i2}, \ldots x_{ip}, y_i\}$$

Feature 2

$$\{x_{12}, x_{22}, \ldots x_{n2}\}$$

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
...	...	...	...	...

Model

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$\begin{cases} \hat{y}_1 = \theta_0 + \theta_1 x_{11} + \theta_2 x_{12} + \cdots + \theta_p x_{1p} \\ \hat{y}_2 = \theta_0 + \theta_1 x_{21} + \theta_2 x_{22} + \cdots + \theta_p x_{2p} \\ \vdots \\ \hat{y}_n = \theta_0 + \theta_1 x_{n1} + \theta_2 x_{n2} + \cdots + \theta_p x_{np} \end{cases}$$

The **dot product (or inner product)** is a vector operation that

- can only be carried out on two vectors of the **same length**
- sums up the products of the corresponding entries of the two vectors, and
- returns a single number

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$$\begin{aligned} \boxed{\vec{u} \cdot \vec{v}} &= \vec{u}^\top \vec{v} = \vec{v}^\top \vec{u} \\ &= 1 \cdot 1 + 2 \cdot 1 + 3 \cdot 1 \\ &= 6 \end{aligned}$$

"u dot v"

Sidenote (not in scope): we can interpret dot product geometrically:

- It is the product of three things: the **magnitude** of both vectors, and the **cosine** of the angles between them. $\vec{u} \cdot \vec{v} = ||\vec{u}|| \cdot ||\vec{v}|| \cdot \cos\theta$
- Another interpretation: 3Blue1Brown

Vector Notation

$$\hat{y} = \theta_0 + \underbrace{\theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p}_{}$$

This part looks a little like a dot product...

We want to collect all the θ_i's into a single vector

$$= \boxed{\theta_0} + \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \theta_0 \cdot 1 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

😱What about this one???

Vector Notation

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

$$= \theta_0 \cdot 1 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$$

We want to collect all the θ_i's into a single vector

$$= \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \cdot \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = x^\top \theta$$

bias term, intercept term

Matrix Notation

$$\begin{cases} \hat{y}_1 = \theta_0 + \theta_1 x_{11} + \theta_2 x_{12} + \cdots + \theta_p x_{1p} \\ \hat{y}_2 = \theta_0 + \theta_1 x_{21} + \theta_2 x_{22} + \cdots + \theta_p x_{2p} \\ \vdots \\ \hat{y}_n = \theta_0 + \theta_1 x_{n1} + \theta_2 x_{n2} + \cdots + \theta_p x_{np} \end{cases}$$

$$\begin{cases} \hat{y}_1 = x_1^\top \theta & \text{where } x_1^\top = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \text{is datapoint/observation 1} \\ \hat{y}_2 = x_2^\top \theta & \text{where } x_2^\top = \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \text{is datapoint/observation 2} \\ \vdots \\ \hat{y}_n = x_n^\top \theta & \text{where } x_n^\top = \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{is datapoint/observation n} \end{cases}$$

Matrix Notation

$$\begin{cases} \hat{y}_1 = x_1^\top \theta & \text{where } x_1^\top = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \text{is datapoint/observation 1} \\ \hat{y}_2 = x_2^\top \theta & \text{where } x_2^\top = \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \text{is datapoint/observation 2} \\ \vdots & \\ \hat{y}_n = x_n^\top \theta & \text{where } x_n^\top = \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \text{is datapoint/observation n} \end{cases}$$

	FG	AST	3PA	PTS
1	1.8	0.6	4.1	5.3
2	0.4	0.8	1.5	1.7
3	1.1	1.9	2.2	3.2
4	6.0	1.6	0.0	13.9
5	3.4	2.2	0.2	8.9
...	...	...	...	...

For data point/observation 2, we have

$$x_2 = \begin{bmatrix} 1 \\ 0.4 \\ 0.8 \\ 1.5 \end{bmatrix} \quad y_2 = 1.7 \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

$$\begin{aligned} \hat{y}_2 &= x_2^\top \theta \\ &= \theta_0 + \theta_1 \cdot 0.4 + \theta_2 \cdot 0.8 + \theta_3 \cdot 1.5 \end{aligned}$$

Dimension check

$x_2 \in \mathbb{R}^4$ or $\mathbb{R}^{(p+1)}$

$\theta \in \mathbb{R}^4$ or $\mathbb{R}^{(p+1)}$

$y_2 \in \mathbb{R}$ $\quad \hat{y}_2 \in \mathbb{R}$

also called scalars

Matrix Notation

$$\hat{y}_1 = \boxed{\begin{matrix} [1 & x_{11} & x_{12} & \dots & x_{1p}] \\ [1 & x_{21} & x_{22} & \dots & x_{2p}] \\ & & \vdots & & \\ [1 & x_{n1} & x_{n2} & \dots & x_{np}] \end{matrix}}$$

$$\hat{y}_1 = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \theta = x_1^T \theta$$

$$\hat{y}_2 = \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \theta = x_2^T \theta$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$\hat{y}_n = \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \theta = x_n^T \theta$$

n row vectors, each with dimension **(p+1)**

Expand out each datapoint's (transposed) input

Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \boxed{\begin{matrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \end{bmatrix} \\ \begin{bmatrix} 1 & x_{21} & x_{22} & \dots & x_{2p} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \end{matrix}} \, \theta$$

n row vectors, each with dimension **(p+1)**

Vectorize predictions and parameters to encapsulate all n equations into a single matrix equation.

Matrix Notation

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \boxed{\mathbb{X}}\ \theta$$

Design matrix with dimensions n x (p + 1)

The Design Matrix $\mathbb{X}$

We can use linear algebra to represent our predictions of all n data points at once.

One step in this process is to stack all of our input features together into a **design matrix**:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

What do the **rows** and **columns** of the design matrix represent in terms of the observed data?

	Field Goals	Assists	3-Point Attempts	
Bias	**FG**	**AST**	**3PA**	**PTS**
1	1.8	0.6	4.1	5.3
1	0.4	0.8	1.5	1.7
1	1.1	1.9	2.2	3.2
1	6.0	1.6	0.0	13.9
1	3.4	2.2	0.2	8.9
...	...	...	...	...
1	4.0	0.8	0.0	11.5
1	3.1	0.9	0.0	7.8
1	3.6	1.1	0.0	8.9
1	3.4	0.8	0.0	8.5
1	3.8	1.5	0.0	9.4

Example design matrix
708 rows x (3+1) cols

The Design Matrix $\mathbb{X}$

We can use linear algebra to represent our predictions of all n data points at once.

One step in this process is to stack all of our input features together into a **design matrix**:

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

A **row** corresponds to one **observation**, e.g., all (p+1) features for datapoint 3

A **column** corresponds to a **feature**, e.g. feature 1 for all n data points

Special all-ones feature often called the **bias/intercept**

Bias	FG (Field Goals)	AST (Assists)	3PA (3-Point Attempts)	PTS
1	1.8	0.6	4.1	5.3
1	0.4	0.8	1.5	1.7
1	1.1	1.9	2.2	3.2
1	6.0	1.6	0.0	13.9
1	3.4	2.2	0.2	8.9
...	...	...	...	...
1	4.0	0.8	0.0	11.5
1	3.1	0.9	0.0	7.8
1	3.6	1.1	0.0	8.9
1	3.4	0.8	0.0	8.5
1	3.8	1.5	0.0	9.4

Example design matrix
708 rows x (3+1) cols

The Multiple Linear Regression Model using Matrix Notation

We can express our linear model on our entire dataset as follows:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}$$

$$\boxed{\hat{\mathbb{Y}} = \mathbb{X}\theta}$$

Prediction vector $\mathbb{R}^n$

Design matrix $\mathbb{R}^{n \times (p+1)}$

Parameter vector $\mathbb{R}^{(p+1)}$

Note that our **true output** is also a vector:

$$\mathbb{Y} \in \mathbb{R}^n$$

Linear in Theta

An expression is "**linear in theta**" if it is a **linear combination** of parameters $\theta = [\theta_0, \theta_1, \ldots, \theta_p]$

.

1. $\hat{y} = \theta_0 + \theta_1(2) + \theta_2(4 \cdot 8) + \theta_3(\log 42)$

2. $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 x_3 + \theta_3 . \log(x_4)$

3. $\hat{y} = \theta_0 + \theta_1 x_1 + \log(\theta_2) x_2 + \theta_3 \theta_4$

4. $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 5 & 6 & 7 \\ 1 & 8 & 9 & 0 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

5. $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Which of these expressions are linear in theta?

slido

Which of the following expressions are linear in theta?

ⓘ Start presenting to display the poll results on this slide.

Linear in Theta

An expression is "**linear in theta**" if it is a **linear combination** of parameters $\theta = [\theta_0, \theta_1, \dots, \theta_p]$

1. $$\hat{y} = \theta_0 + \theta_1(2) + \theta_2(4 \cdot 8) + \theta_3(\log 42)$$
$$= \begin{bmatrix} 1 & 2 & 4.8 & \log(42) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

2. $$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 x_3 + \theta_3 . \log(x_4)$$
$$= \begin{bmatrix} 1 & x_1 & x_2 x_3 & \log(x_4) \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

3. $$\hat{y} = \theta_0 + \theta_1 x_1 + \log(\theta_2) x_2 + \theta_3 \theta_4$$
✕

4. $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 1 & 5 & 6 & 7 \\ 1 & 8 & 9 & 0 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

5. $$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

"**Linear in theta**" means the expression can separate into a matrix product of two terms: **a vector of thetas**, and a matrix/vector not involving thetas.

Mean Squared Error

Today's Goal: Ordinary Least Squares

1. Choose a model

Multiple Linear Regression

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

2. Choose a loss function

L2 Loss

Mean Squared Error (MSE)

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

More Linear Algebra!!

3. Fit the model

Minimize average loss with ~~calculus~~ geometry

4. Evaluate model performance

Visualize,
~~Root MSE~~
Multiple R^2

[Linear Algebra] Vector Norms and the L2 Vector Norm

The **norm** of a vector is some measure of that vector's **size/length**.

- The two norms we need to know for Data 100 are the L_1 and L_2 norms (sound familiar?).
- Today, we focus on L_2 norm. We'll define the L_1 norm another day.

For the n-dimensional vector $\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, the **L2 vector norm** is

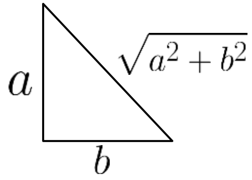
$$||\vec{x}||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n} (x_i^2)}$$

[Linear Algebra] The L2 Norm as a Measure of Length

The L2 vector norm is a generalization of the Pythagorean theorem into n dimensions.

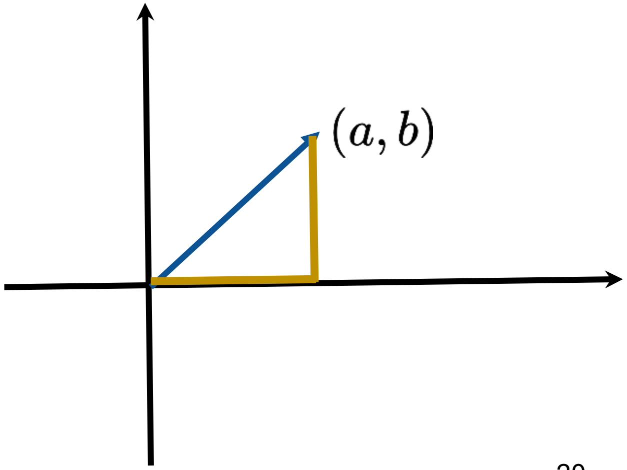
$$||\vec{x}||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n}(x_i^2)}$$

In $\mathbb{R}^2$



It can therefore be used as a measure of **length** of a vector

- The vector on the right has length $||\vec{v}||_2 = \sqrt{a^2 + b^2}$

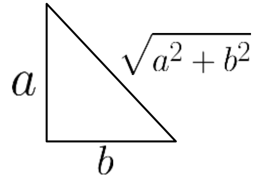


[Linear Algebra] The L2 Norm as a Measure of Distance

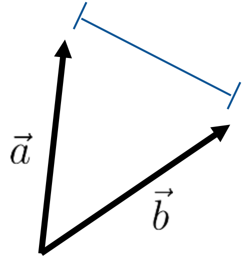
The L2 vector norm is a generalization of the Pythagorean theorem into *n* dimensions.

$$||\vec{x}||_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\sum_{i=1}^{n}(x_i^2)}$$

In $\mathbb{R}^2$



In $\mathbb{R}^n$



It can also be used as a measure of **distance** between two vectors.

- For n-dimensional vectors $\vec{a}, \vec{b}$, their distance is $||\vec{a} - \vec{b}||_2$.

Note: The square of the L2 norm of a vector is the sum of the squares of the vector's elements:

$$(||\vec{x}||_2)^2 = \sum_{i=1}^{n} x_i^2$$

Looks like Mean Squared Error!!

Mean Squared Error with L2 Norms

We can rewrite mean squared error as a squared L2 norm:

$$R(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$= \frac{1}{n}||\mathbb{Y} - \hat{\mathbb{Y}}||_2^2$$

With our linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$:

$$\boxed{R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2}$$

Ordinary Least Squares

The **least squares estimate** $\hat{\theta}$ is the parameter that **minimizes** the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

How should we interpret the OLS problem?

A. Minimize the mean squared error for the linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$

B. Minimize the **distance** between true and predicted values $\mathbb{Y}$ and $\hat{\mathbb{Y}}$

C. Minimize the **length** of the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}} = \begin{bmatrix} y_1 - \hat{y_1} \\ y_2 - \hat{y_2} \\ \vdots \\ y_n - \hat{y_n} \end{bmatrix}$

D. All of the above

E. Something else

slido

How should we interpret the OLS problem?

ⓘ Start presenting to display the poll results on this slide.

Ordinary Least Squares

The **least squares estimate** $\hat{\theta}$ is the parameter that **minimizes** the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

How should we interpret the OLS problem?

A. Minimize the mean squared error for the linear model $\hat{\mathbb{Y}} = \mathbb{X}\theta$

B. Minimize the **distance** between true and predicted values $\mathbb{Y}$ and $\hat{\mathbb{Y}}$

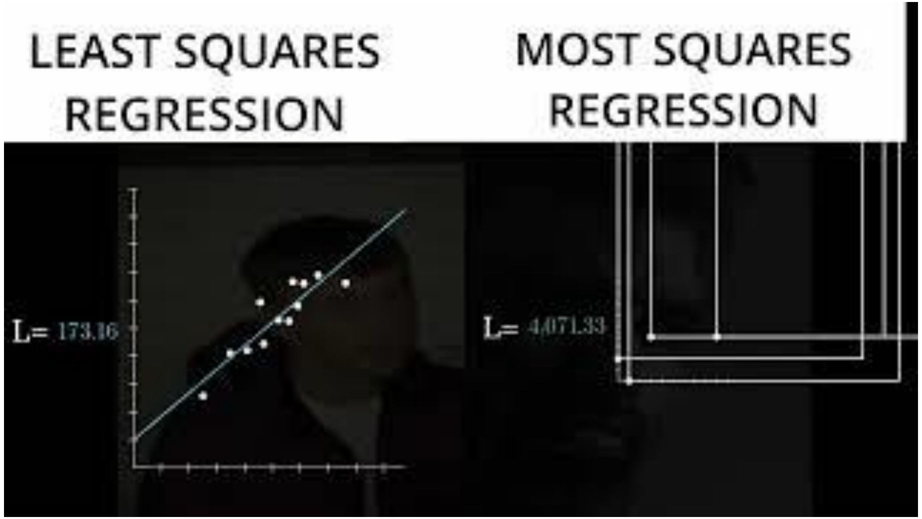
C. Minimize the **length** of the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}} = \begin{bmatrix} y_1 - \hat{y_1} \\ y_2 - \hat{y_2} \\ \vdots \\ y_n - \hat{y_n} \end{bmatrix}$

Important for today

D. All of the above

E. Something else

Interlude


LEAST SQUARES REGRESSION
MOST SQUARES REGRESSION
L= 173.16
L= 4,071.33

Geometric Derivation

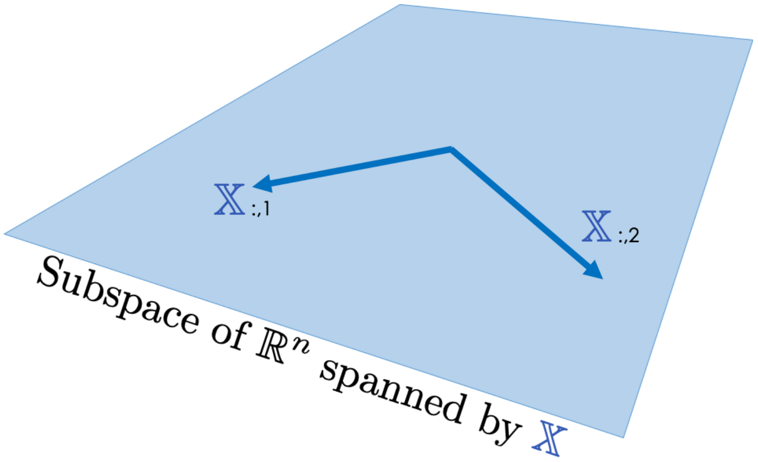
Lecture 11, Data 100 Summer 2023

Today's Goal: Ordinary Least Squares

1. Choose a model	Multiple Linear Regression	$\hat{\mathbb{Y}} = \mathbb{X}\theta$
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	$R(\theta) = \frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2$
3. Fit the model	Minimize average loss with ~~calculus~~ geometry	The calculus derivation requires matrix calculus (out of scope, but here's a link if you're interested). Instead, we will derive $\hat{\theta}$ using a **geometric argument**.
4. Evaluate model performance	Visualize, ~~Root MSE~~ Multiple R^2	

The set of all possible linear combinations of the columns of $\mathbb{X}$ is called the **span** of the columns of $\mathbb{X}$ (denoted $span(\mathbb{X})$), also called the **column space**.

- Intuitively, this is all of the vectors you can "reach" using the columns of $\mathbb{X}$.
- If each column of $\mathbb{X}$ has length n, $span(\mathbb{X})$ is a subspace of $\mathbb{R}^n$.



Approach 1: So far, we've thought of our model as horizontally stacked predictions per datapoint:

$$\begin{array}{c} n \\ \\ \end{array}\underset{1}{\begin{bmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{bmatrix}} = \begin{bmatrix} \text{——} \; x_1^T \; \text{——} \\ \text{——} \; x_2^T \; \text{——} \\ \vdots \\ \text{——} \; x_n^T \; \text{——} \end{bmatrix} \underset{1}{\begin{bmatrix} | \\ \theta \\ | \end{bmatrix}}\begin{array}{c} p+1 \\ \\ \end{array}$$

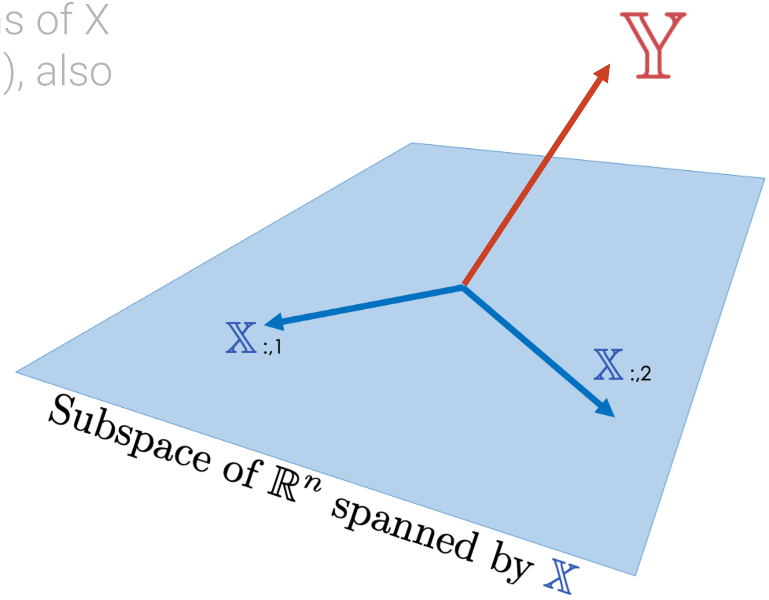
Approach 2: However, it is helpful sometimes to think of matrix-vector multiplication as performed by columns. We can also think of $\hat{\mathbb{Y}}$ as a **linear combination of feature vectors**, scaled by **parameters**.

$$\begin{array}{c} n \\ \\ \end{array}\underset{1}{\begin{bmatrix} | \\ \hat{\mathbb{Y}} \\ | \end{bmatrix}} = \begin{array}{c} n \\ \\ \end{array}\underset{p+1}{\begin{pmatrix} | & | \\ \mathbb{X}_{:,1} & \mathbb{X}_{:,2} \\ | & | \end{pmatrix}} \underset{1}{\begin{bmatrix} | \\ \theta \\ | \end{bmatrix}}\begin{array}{c} p+1 \\ \\ \end{array} = \theta_1 \begin{matrix} | \\ \mathbb{X}_{:,1} \\ | \end{matrix} + \theta_2 \begin{matrix} | \\ \mathbb{X}_{:,2} \\ | \end{matrix}$$

Prediction is a Linear Combination of Columns

The set of all possible linear combinations of the columns of X is called the **span** of the columns of X (denoted $span(\mathbb{X})$), also called the **column space**.

- Intuitively, this is all of the vectors you can "reach" using the columns of X.
- If each column of X has length n, $span(\mathbb{X})$ is a subspace of $\mathbb{R}^n$.



Our prediction $\hat{\mathbb{Y}} = \mathbb{X}\theta$ is a **linear combination** of the columns of $\mathbb{X}$. Therefore $\hat{\mathbb{Y}} \in span(\mathbb{X})$.

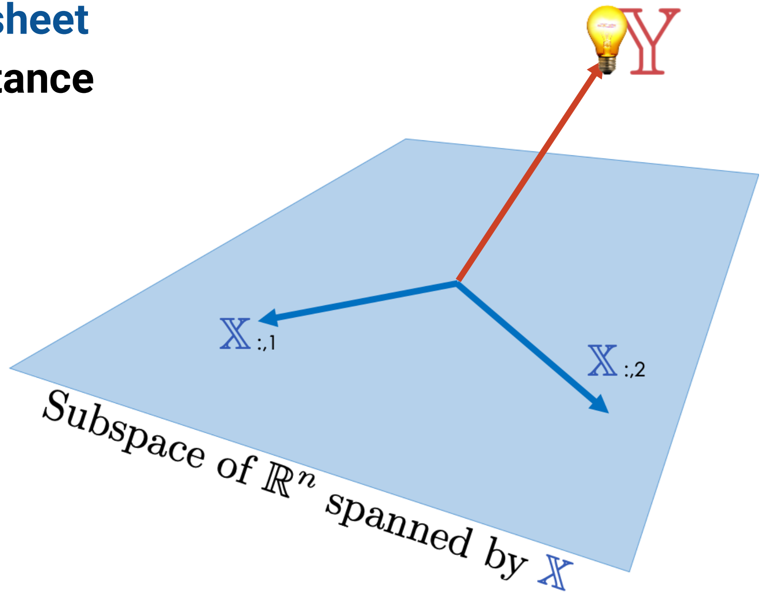
Interpret: Our linear prediction $\hat{\mathbb{Y}}$ will be in $span(\mathbb{X})$, even if the true $\mathbb{Y}$ values might not be.

Goal: Find the vector in $span(\mathbb{X})$ that is **closest** to $\mathbb{Y}$.

A thought experiment

If you’re a human being who can only stand on the **blue sheet of paper**, and you need to get as close as possible in **distance** to the **light bulb** located at the tip of the **red** arrow.

Where do you stand on the blue sheet?

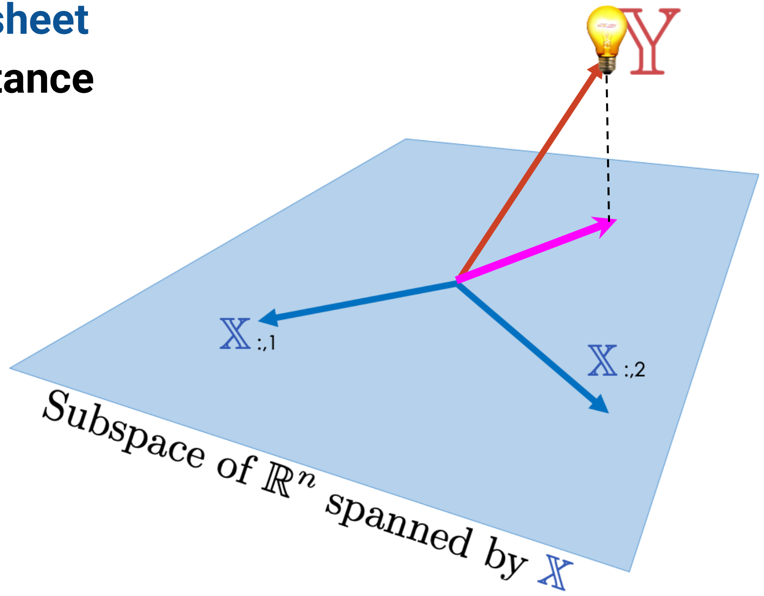


A thought experiment

If you’re a human being who can only stand on the **blue sheet of paper**, and you need to get as close as possible in **distance** to the **light bulb** located at the tip of the **red** arrow.

Where do you stand on the blue sheet?

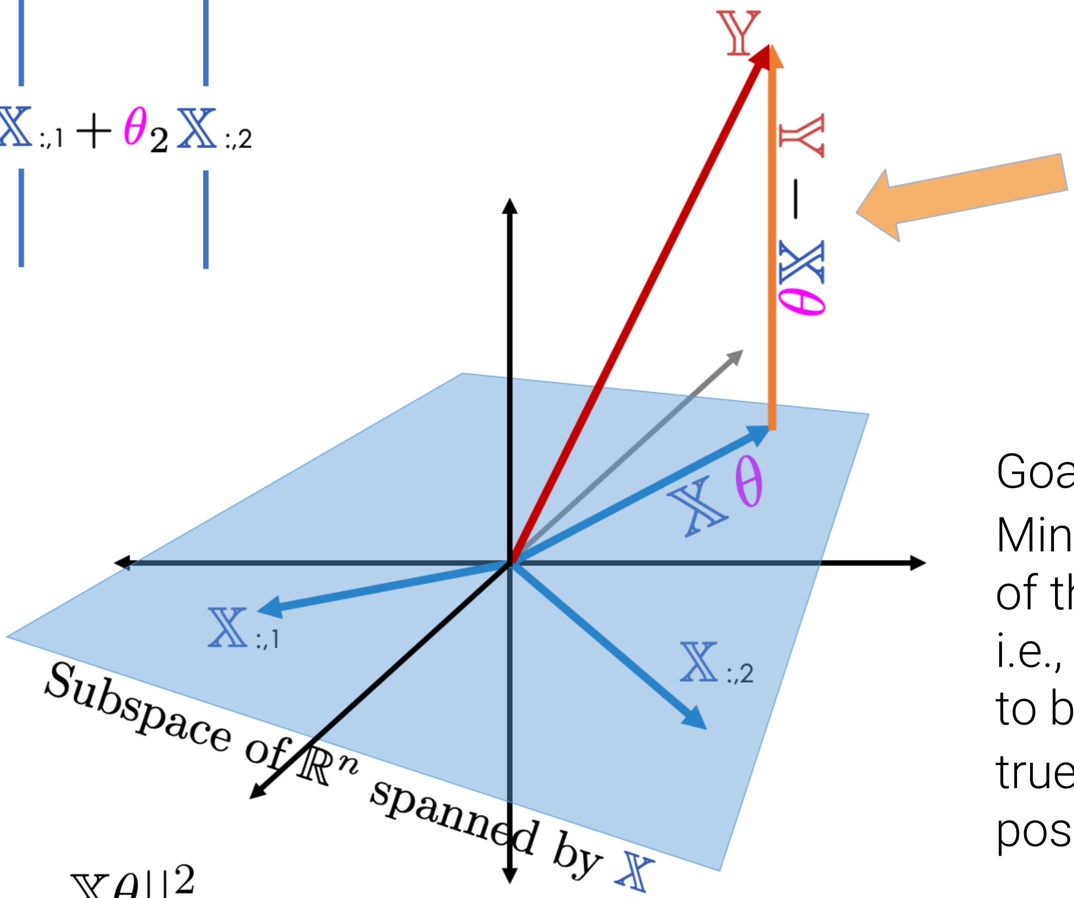
Right below the lightbulb - that’s the closest you can get because you can’t travel vertically!



$$\hat{\mathbb{Y}} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

n

1



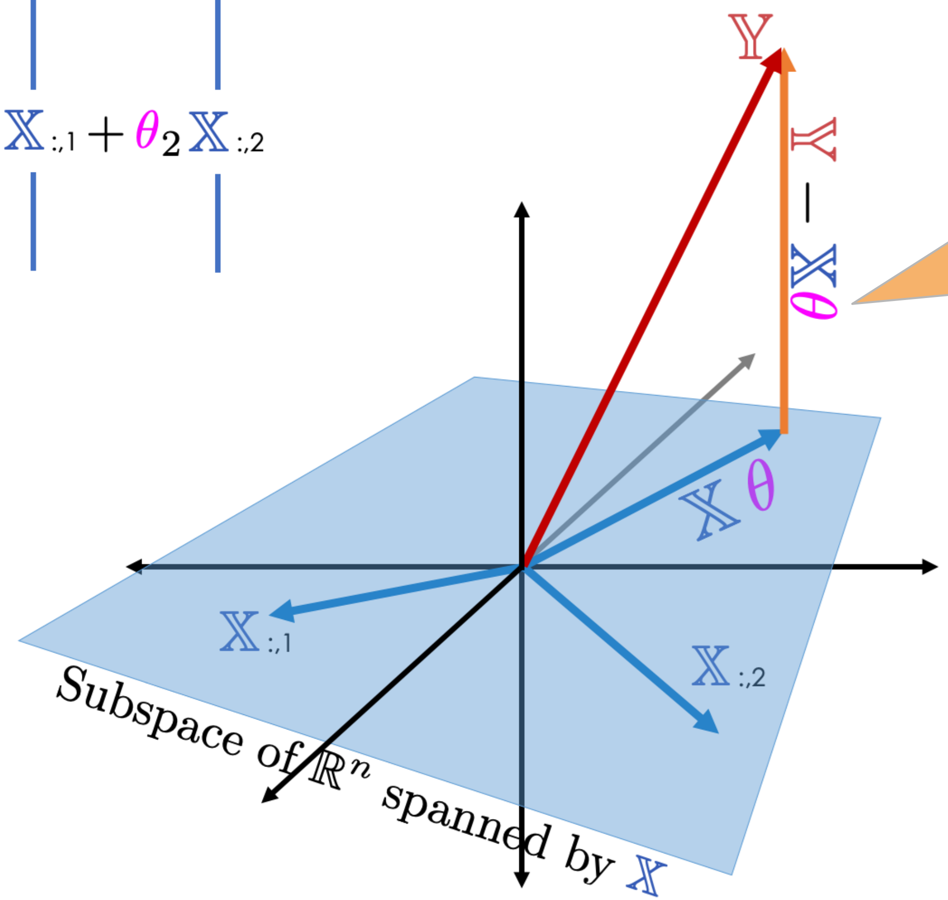
This is the residual vector, $e = \mathbb{Y} - \hat{\mathbb{Y}}$.

Goal:

Minimize the L_2 norm of the residual vector. i.e., get the predictions $\hat{\mathbb{Y}}$ to be "as close" to our true $\mathbb{Y}$ values as possible.

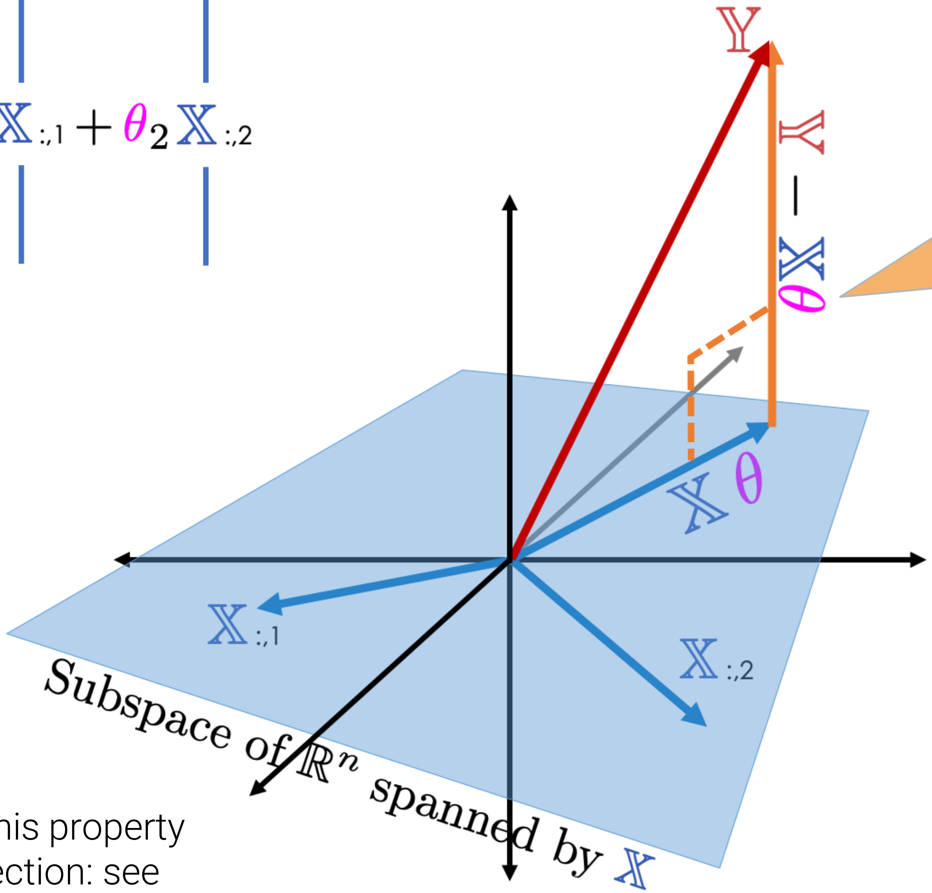
$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

$$\hat{\mathbb{Y}} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$



How do we minimize this distance – the norm of the residual vector (squared)?

$$\underbrace{\hat{\mathbb{Y}}}_{n \times 1} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$

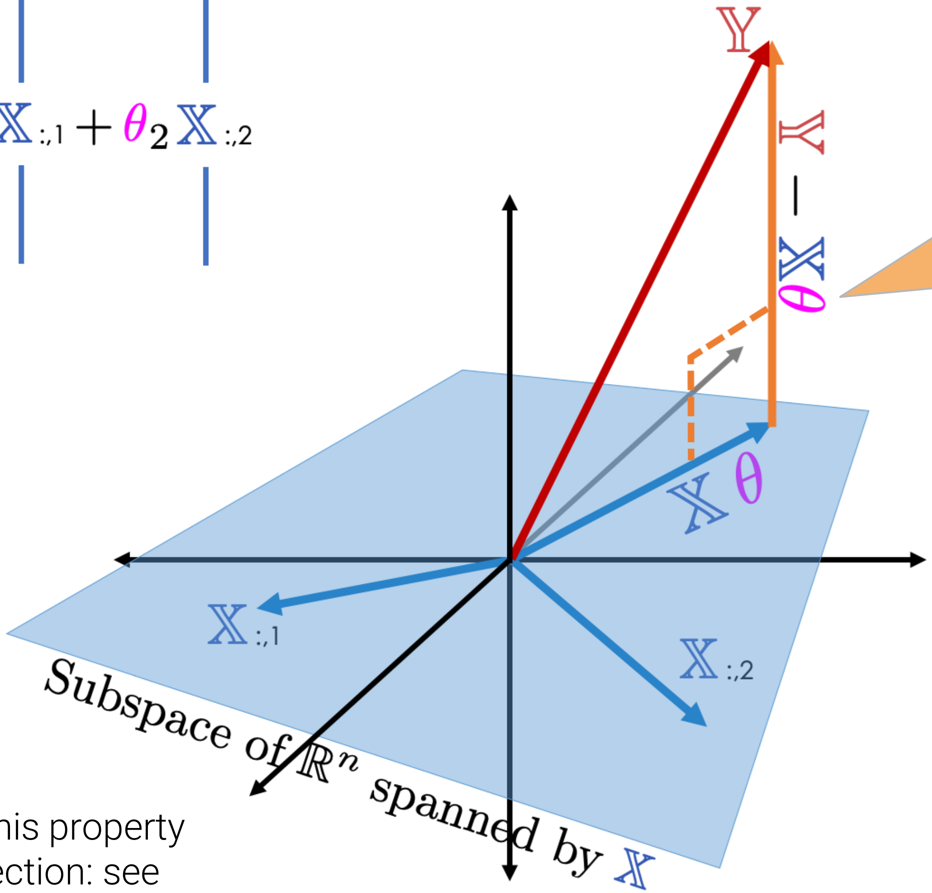


How do we minimize this distance – the norm of the residual vector (squared)?

The vector in $span(\mathbb{X})$ that is closest to $\mathbb{Y}$ is the **orthogonal projection** of $\mathbb{Y}$ onto $span(\mathbb{X})$.

We will not prove this property of orthogonal projection: see Khan Academy.

$$\hat{\mathbb{Y}} = \theta_1 \mathbb{X}_{:,1} + \theta_2 \mathbb{X}_{:,2}$$



How do we minimize this distance – the norm of the residual vector (squared)?

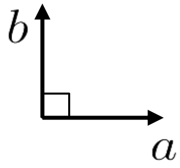
The vector in $span(\mathbb{X})$ that is closest to $\mathbb{Y}$ is the **orthogonal projection** of $\mathbb{Y}$ onto $span(\mathbb{X})$.

Thus, we should choose the θ that makes the residual vector **orthogonal** to $span(\mathbb{X})$.

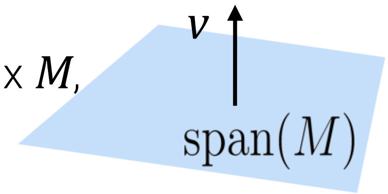
We will not prove this property of orthogonal projection: see Khan Academy.

[Linear Algebra] Orthogonality

1. Vector a and Vector b are **orthogonal** if and only if their dot product is 0: $a^T b = 0$
This is a generalization of the notion of two vectors in 2D being perpendicular.



2. A vector v is **orthogonal** to $\text{span}(M)$, the span of the columns of a matrix M, if and only if v is orthogonal to **each column** in M.



Let's express **2** in matrix notation. Let $v \in \mathbb{R}^{n \times 1}$, $M \in \mathbb{R}^{n \times d}$ where $M = \begin{bmatrix} | & | & & | \\ M_{:1} & M_{:2} & \vdots & M_{:d} \\ | & | & & | \end{bmatrix}$

$$\begin{aligned} M_{:1}^{\top} v &= 0 \\ M_{:2}^{\top} v &= 0 \\ &\vdots \\ M_{:d}^{\top} v &= 0 \end{aligned} \quad \Rightarrow \quad \begin{bmatrix} M_{:1}^{\top} v \\ M_{:2}^{\top} v \\ \vdots \\ M_{:d}^{\top} v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Rightarrow \quad \underbrace{M^T}_{M^T \in \mathbb{R}^{d \times n}} v = \underbrace{\vec{0}}_{\text{zero vector}}$$

v is orthogonal to each column of M, $M_{:j} \in \mathbb{R}^n$

zero vector (d-length vector full of 0s).

Ordinary Least Squares Proof

The **least squares estimate** $\hat{\theta}$ is the parameter θ that minimizes the objective function $R(\theta)$:

$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

Design matrix

$$M^T v = \vec{0}$$

Residual vector

Equivalently, this is the $\hat{\theta}$ such that the residual vector $\mathbb{Y} - \mathbb{X}\hat{\theta}$ is orthogonal to $span(\mathbb{X})$

Definition of orthogonality of $\mathbb{Y} - \mathbb{X}\hat{\theta}$ to $span(\mathbb{X})$ (0 is the $\vec{0}$ vector)

$$\mathbb{X}^T\left(\mathbb{Y} - \mathbb{X}\hat{\theta}\right) = 0$$

Rearrange terms

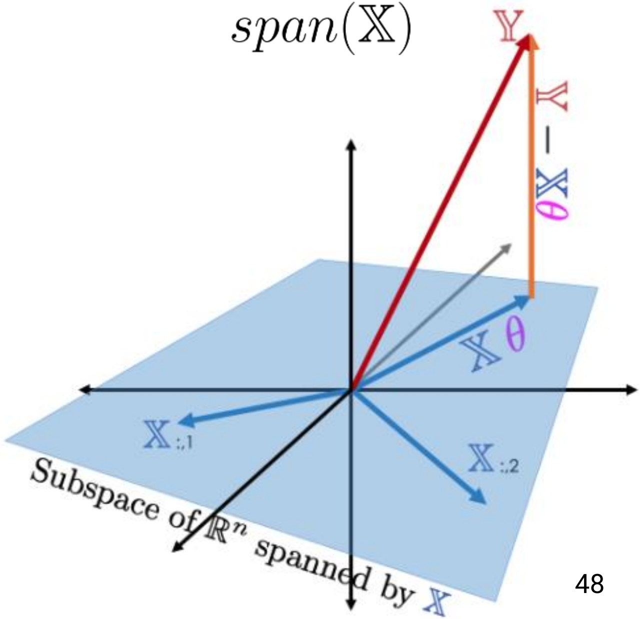
$$\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}\hat{\theta} = 0$$

The **normal equation**

$$\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$$

If $\mathbb{X}^T\mathbb{X}$ is invertible

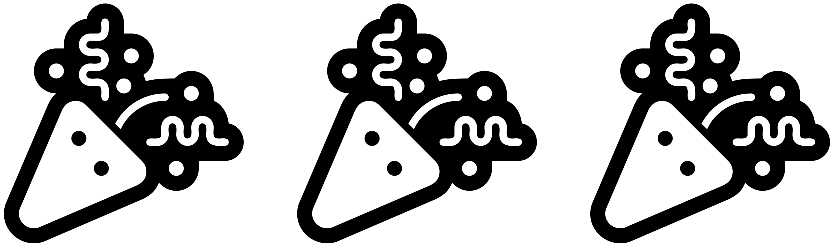
$$\boxed{\hat{\theta} = \left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^T\mathbb{Y}}$$



$$\hat{\theta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top \mathbb{Y}$$

This result is so important that it deserves its own slide.

It is the **least squares estimate** and the solution to the normal equation $\mathbb{X}^T \mathbb{X} \hat{\theta} = \mathbb{X}^T \mathbb{Y}$



$$\hat{\theta} = (\mathbb{X}^{\top}\mathbb{X})^{-1}\mathbb{X}^{\top}\mathbb{Y}$$

This result is so important that it deserves its own slide.

It is the **least squares estimate** and the solution to the normal equation $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$

Least Squares Estimate

1. Choose a model	Multiple Linear Regression	$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	$$R(\theta) = \frac{1}{n}\|\|\mathbb{Y} - \mathbb{X}\theta\|\|_2^2$$
3. Fit the model	Minimize average loss with ~~calculus~~ geometry	$$\hat{\theta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}$$
4. Evaluate model performance	Visualize, ~~Root MSE~~ Multiple R^2	

Performance

OLS Problem Formulation

- Multiple Linear Regression Model
- Mean Squared Error

Geometric Derivation

Performance: Residuals, Multiple R^2

OLS Properties

- Residuals
- The Bias/Intercept Term
- Existence of a Unique Solution

Least Squares Estimate

1. Choose a model	Multiple Linear Regression	$\hat{\mathbb{Y}} = \mathbb{X}\theta$
2. Choose a loss function	L2 Loss Mean Squared Error (MSE)	$R(\theta) = \frac{1}{n}\vert\vert\mathbb{Y} - \mathbb{X}\theta\vert\vert_2^2$
3. Fit the model	Minimize average loss with ~~calculus~~ geometry	$\hat{\theta} = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top\mathbb{Y}$
4. Evaluate model performance	Visualize, ~~Root MSE~~ Multiple R^2	

Demo

Multiple Linear Regression

$$\hat{\mathbb{Y}} = \mathbb{X}\theta$$

Prediction vector $\mathbb{R}^n$

Design matrix $\mathbb{R}^{n \times (p+1)}$

Parameter vector $\mathbb{R}^{(p+1)}$

Note that our **true output** is also a vector:

$$\mathbb{Y} \in \mathbb{R}^n$$

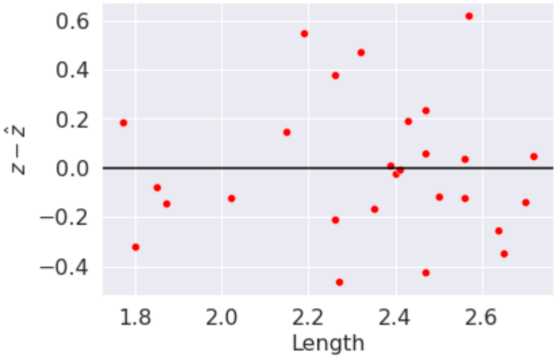
$$R(\theta) = \frac{1}{n}||\mathbb{Y} - \mathbb{X}\theta||_2^2$$

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

[Visualization] Residual Plots

Simple linear regression

Plot residuals vs
the single feature x.

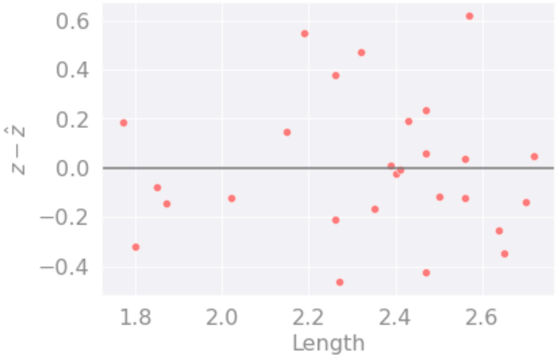


Compare

[Visualization] Residual Plots

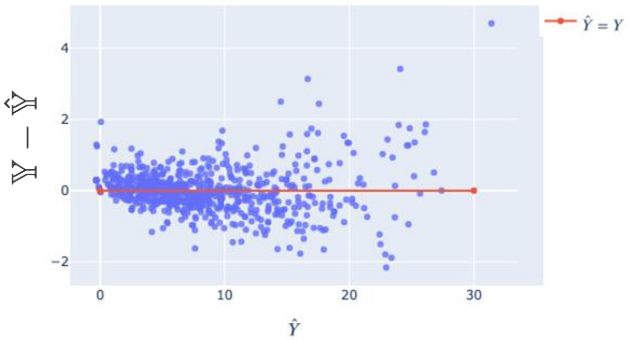
Simple linear regression

Plot residuals vs
the single feature x.



Multiple linear regression

Plot residuals vs
fitted (predicted) values $\hat{y}$



Compare

See notebook

Same interpretation as before (Data 8 textbook):

- A good residual plot shows no pattern.
- A good residual plot also has a similar vertical spread throughout the entire plot. Else (heteroscedasticity), the accuracy of the predictions is not reliable.

[Metrics] Multiple R^2

Compare

Simple linear regression

<u>Error</u>

RMSE $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

<u>Linearity</u>

Correlation coefficient, r

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Multiple linear regression

<u>Error</u>

RMSE $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$

<u>Linearity</u>

Multiple R^2, also called the **coefficient of determination**

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

Compare

[Metrics] Multiple R^2

We define the **multiple R^2** value as the **proportion of variance** or our **fitted values** (predictions) $\hat{y}$ to our true values y .

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

Also called the **correlation of determination**.

R^2 ranges from 0 to 1 and is effectively
“the proportion of variance that the **model explains**.”

For OLS with an intercept term (e.g. $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$),
$R^2 = [r(y, \hat{y})]^2$ is equal to the square of correlation between y, $\hat{y}$.

- For SLR, $R^2 = r^2$, the correlation between ***x***, y.

[Metrics] Multiple R^2

$$\text{predicted PTS} = 3.98 + 2.4 \cdot \text{AST}$$

$R^2 = 0.457$

$$\text{predicted PTS} = 2.163 + 1.64 \cdot \text{AST} + 1.26 \cdot \text{3PA}$$

$R^2 = 0.609$

Compare

Simple linear regression

Error

RMSE

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Linearity

Correlation coefficient, r

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Multiple linear regression

Error

RMSE

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Linearity

Multiple R^2, also called the **coefficient of determination**

$$R^2 = \frac{\text{variance of fitted values}}{\text{variance of } y} = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2}$$

As we add more features, our fitted values tend to become closer and closer to our actual y values. Thus, R^2 increases.

- The SLR **model** (AST only) explains 45.7% of the variance in the true y.
- The AST & 3PA **model** explains 60.9%.

Adding more features doesn't always mean our model is better, though! We will see why after the midterm.

OLS Properties

OLS Problem Formulation

- Multiple Linear Regression Model
- Mean Squared Error

Geometric Derivation

Performance: Residuals, Multiple R^2

OLS Properties

- Residuals
- The Bias/Intercept Term
- Existence of a Unique Solution

Residual Properties

When using the optimal parameter vector, our residuals $e = \mathbb{Y} - \mathbb{X}\hat{\theta}$ are orthogonal to $span(\mathbb{X})$.

$$\mathbb{X}^T e = 0$$

Proof: First line of our OLS estimate proof (slide).

For all linear models:

Since our predicted response $\hat{\mathbb{Y}}$ is in $span(\mathbb{X})$ by definition, $\hat{\mathbb{Y}}^T e = 0$, and hence it is orthogonal to the residuals.

For all linear models with an **intercept term**, $\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_p x_p$, the **sum of residuals is zero**.

$$\sum_{i=1}^{n} e_i = 0$$

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ 1 & x_{31} & x_{32} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

You will prove both properties in homework.

(Proof hint) $\mathbb{1}^T e = 0$

Properties When Our Model Has an Intercept Term

For all linear models with an **intercept term**, the **sum of residuals is zero**.

$$\sum_{i=1}^{n} e_i = 0$$ (previous slide)

- This is the real reason why we don't directly use residuals as loss.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

- This is also why positive and negative residuals will cancel out in any residual plot where the (linear) model contains an intercept term, even if the model is terrible.

It follows from the property above that for linear models with intercepts, the average predicted y value is equal to the average true y value.

$$\bar{y} = \overline{\hat{y}}$$

These properties are true when there is an intercept term, and not necessarily when there isn't.

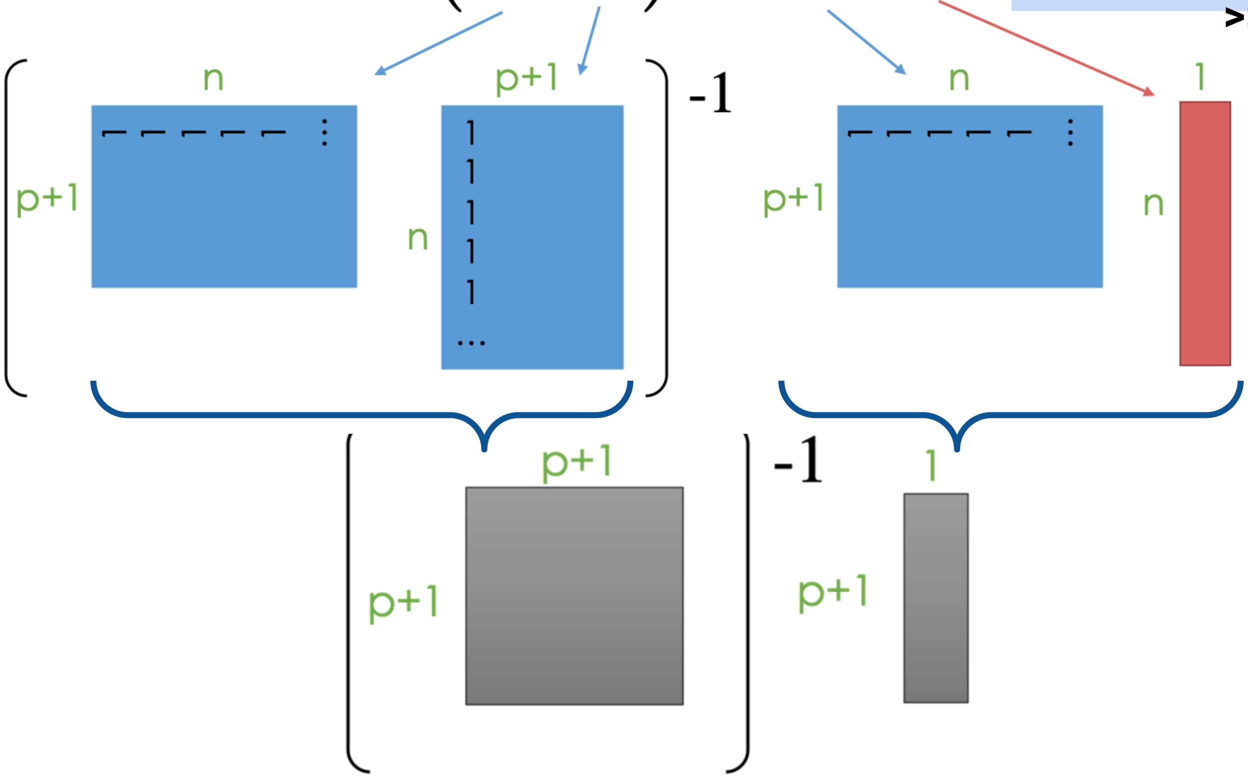
Does a Unique Solution Always Exist?

	Model	Estimate	Unique?
Constant Model + MSE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = mean(y) = \bar{y}$	**Yes**. Any set of values has a unique mean.
Constant Model + MAE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = median(y)$	**Yes**, if odd. **No**, if even. Return average of middle 2 values.
Simple Linear Regression + MSE	$\hat{y} = \theta_0 + \theta_1 x$	$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ $\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$	**Yes**. Any set of non-constant* values has a unique mean, SD, and correlation coefficient.
Ordinary Least Squares (Linear Model + MSE)	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$	???

Understanding The Solution Matrices

In most settings,
 # observations **#**
features
n
>> **p**

$$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

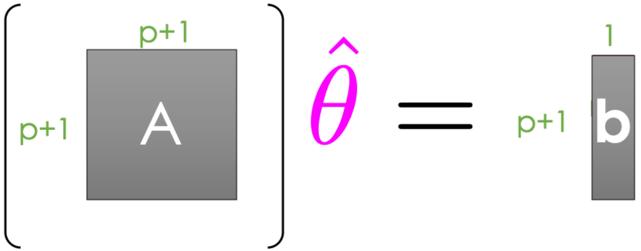


Understanding The Solution Matrices

In practice, instead of directly inverting matrices, we can use more efficient numerical solvers to directly solve a system of linear equations.

The **Normal Equation**:

$$\mathbb{X}^T \mathbb{X} \hat{\theta} = \mathbb{X}^T \mathbb{Y}$$



Note that at least one solution always exists:

Intuitively, we can always draw a line of best fit for a given set of data, but there may be multiple lines that are "equally good". (Formal proof is beyond this course.)

Uniqueness of a Solution: Proof

Claim

The Least Squares estimate $\hat{\theta}$ is **unique** if and only if $\mathbb{X}$ is **full column rank**.

Proof

- The solution to the normal equation $\mathbb{X}^T\mathbb{X}\hat{\theta} = \mathbb{X}^T\mathbb{Y}$ is the least square $\hat{\theta}$ estimate .

- $\hat{\theta}$ has a **unique** solution if and only if the square matrix $\mathbb{X}^T\mathbb{X}$ is **invertible**, which happens if and only if $\mathbb{X}^T\mathbb{X}$ is full rank.
  - The **rank** of a square matrix is the max **# of linearly independent columns** it contains.
  - $\mathbb{X}^T\mathbb{X}$ has shape (p +1) x (p + 1), and therefore has max rank p + 1.

- $\mathbb{X}^T\mathbb{X}$ and $\mathbb{X}$ **have the same rank** (proof out of scope).

- Therefore $\mathbb{X}^T\mathbb{X}$ has rank p + 1 if and only if $\mathbb{X}$ has rank p + 1 (full column rank).

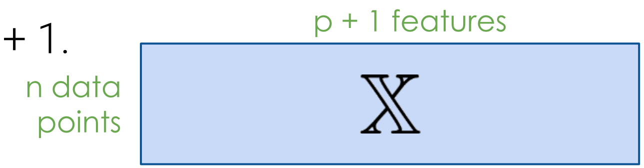
Uniqueness of a Solution: Interpretation

Claim:

The Least Squares estimate $\hat{\theta}$ is **unique** if and only if $\mathbb{X}$ is **full column rank**.

When would we **not** have unique estimates?

1. If our design matrix $\mathbb{X}$ is "**wide**":
   - (property of rank) If n < p, rank of $\mathbb{X}$ = min(n, p + 1) **<** p + 1.
   - In other words, if we have way more features than observations, then $\hat{\theta}$ is not unique.
   - Typically we have n >> p so this is less of an issue.



2. If we our design matrix $\mathbb{X}$ has features that are **linear combinations of other features**.
   - By definition, rank of $\mathbb{X}$ is number of linearly independent columns in $\mathbb{X}$.
   - Example: If "Width", "Height", and "Perimeter" are all columns,
     - Perimeter = 2 * Width + 2 * Height → $\mathbb{X}$ is not full rank.
   - Important with one-hot encoding (to discuss in later).

Does a Unique Solution Always Exist?

	Model	Estimate	Unique?
Constant Model + MSE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = mean(y) = \bar{y}$	**Yes**. Any set of values has a unique mean.
Constant Model + MAE	$\hat{y} = \theta_0$	$\hat{\theta}_0 = median(y)$	**Yes**, if odd. **No**, if even. Return average of middle 2 values.
Simple Linear Regression + MSE	$\hat{y} = \theta_0 + \theta_1 x$	$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$ $\hat{\theta}_1 = r\frac{\sigma_y}{\sigma_x}$	**Yes**. Any set of non-constant* values has a unique mean, SD, and correlation coefficient.
Ordinary Least Squares (Linear Model + MSE)	$\hat{\mathbb{Y}} = \mathbb{X}\theta$	$\hat{\theta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$	**Yes**, if $\mathbb{X}$ is full col rank (all cols lin independent, #datapoints>> #features)

Lecture 11

Ordinary Least Squares