

# Linear Regression

---

Nipun Batra and the teaching staff

IIT Gandhinagar

September 3, 2025

# Table of Contents

1. Setup
2. Normal Equation
3. Basis Expansion
4. Geometric Interpretation
5. Dummy Variables and Multicollinearity
6. Practice and Review

# Setup

# Linear Regression

- Output is continuous in nature.
- Examples of linear systems:
  - $F = ma$
  - $v = u + at$

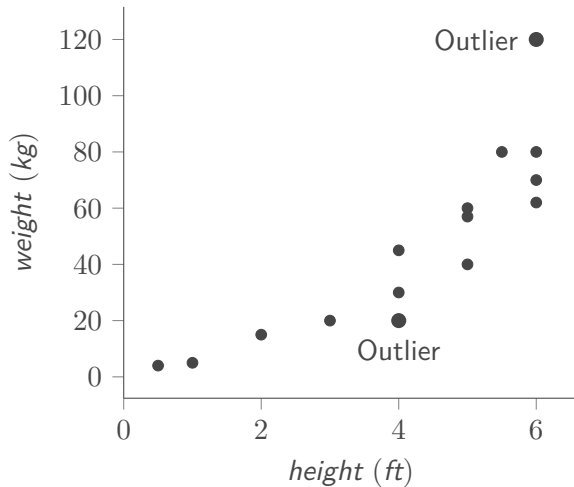
# Task at hand

- TASK: Predict  $\text{Weight} = f(\text{height})$

Height	Weight
3	29
4	35
5	39
2	20
6	41
7	?
8	?
1	?

The first part of the dataset is the training points. The latter ones are testing points.

# Scatter Plot



## Matrix representation of the expression

- $weight_1 \approx \theta_0 + \theta_1 \cdot height_1$
- $weight_2 \approx \theta_0 + \theta_1 \cdot height_2$
- $weight_N \approx \theta_0 + \theta_1 \cdot height_N$

$$weight_i \approx \theta_0 + \theta_1 \cdot height_i$$

## Matrix representation of the expression

$$\begin{bmatrix} weight_1 \\ weight_2 \\ \dots \\ weight_N \end{bmatrix} = \begin{bmatrix} 1 & height_1 \\ 1 & height_2 \\ \dots & \dots \\ 1 & height_N \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\hat{\mathbf{y}}_{n \times 1} = \mathbf{X}_{n \times d} \boldsymbol{\theta}_{d \times 1}$$

- $\theta_0$  - Bias Term/Intercept Term
- $\theta_1$  - Slope



## Extension to multiple dimensions

- In the previous example  $y = f(x)$ , where  $x$  is one-dimensional
- Now consider examples in multiple dimensions
- Example: Predict the water demand of the IITGN campus
- Mathematical representation:

$$\text{Demand} = f(\# \text{ occupants, Temperature})$$

- Linear form:

$$\text{Demand} = \text{Base Demand} + K_1 * \# \text{ occupants} + K_2 * \text{Temperature}$$

# Intuition

We hope to:

- Learn  $f$ :  $Demand = f(\#occupants, Temperature)$
- From training dataset
- To predict the condition for the testing set

# Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for  $i^{th}$  sample is  
 $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand}_i = x_i'^T \theta$
- where  $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$
- and  $x_i' = \begin{bmatrix} 1 \\ Temperature_i \\ \#Occupants_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$
- Notice the transpose in the equation! This is because  $x_i$  is a column vector

## We can expect the following

- Demand increases, if  $\#$  occupants increases, then  $\theta_2$  is likely to be positive
- Demand increases, if temperature increases, then  $\theta_1$  is likely to be positive
- Base demand is independent of the temperature and the  $\#$  occupants, but, likely positive, thus  $\theta_0$  is likely positive.

# Normal Equation

# Generalized Linear Regression Format

- Assuming  $N$  samples for training
- # Features =  $M$

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix}_{N \times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N \times (M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1) \times 1}$$

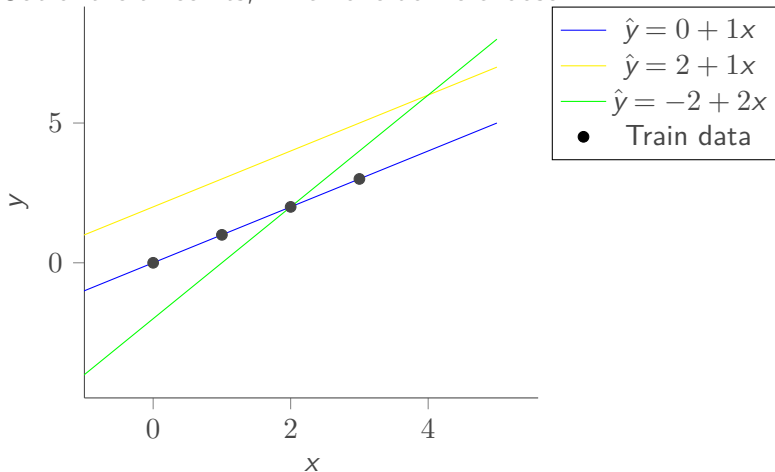
$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$$

# Relationships between feature and target variables

- There could be different  $\theta_0, \theta_1 \dots \theta_M$ . Each of them can represents a relationship.
- Given multiples values of  $\theta_0, \theta_1 \dots \theta_M$  how to choose which is the best?
- Let us consider an example in 2d

# Relationships between feature and target variables

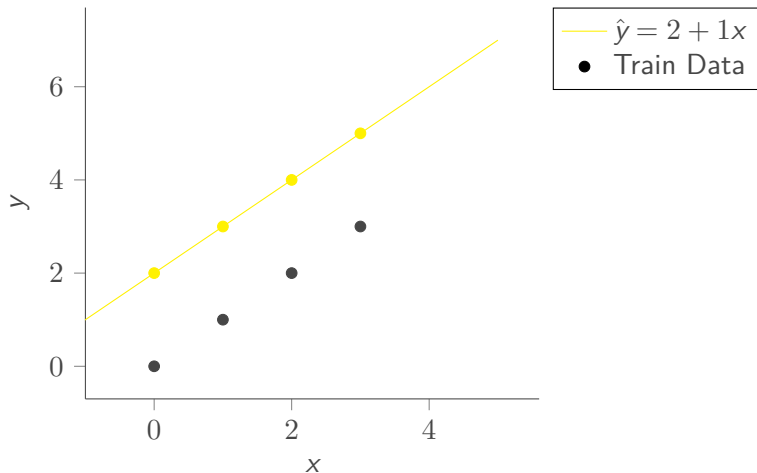
Out of the three fits, which one do we choose?





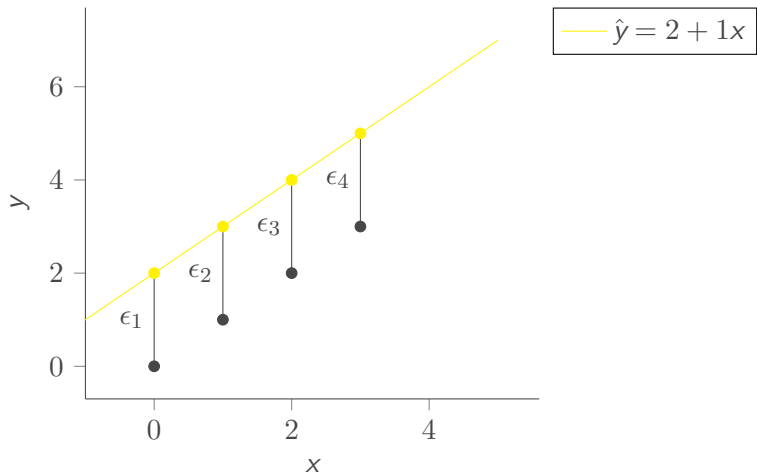
# Relationships between feature and target variables

We have  $\hat{y} = 2 + 1x$  as one relationship.



# Relationships between feature and target variables

How far is our estimated  $\hat{y}$  from ground truth  $y$ ?



# Error terms

- $y_i = \hat{y}_i + \epsilon_i$  where  $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- **Critical Assumption:**  $\epsilon_i$  are independent and identically distributed (i.i.d.)
- $y_i$  denotes the ground truth for  $i^{th}$  sample
- $\hat{y}_i$  denotes the prediction for  $i^{th}$  sample, where  $\hat{y}_i = \mathbf{x}_i^\top \boldsymbol{\theta}$
- $\epsilon_i$  denotes the error/residual for  $i^{th}$  sample
- $\theta_0, \theta_1$ : The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$
- $\epsilon_i = y_i - (\theta_0 + x_i \cdot \theta_1)$

## Good fit

- $|\epsilon_1|, |\epsilon_2|, |\epsilon_3|, \dots$  should be small.
- minimize  $\epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_N^2$  -  $L_2$  Norm
- minimize  $|\epsilon_1| + |\epsilon_2| + \dots + |\epsilon_n|$  -  $L_1$  Norm

# Normal Equation

- Model specification:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

- To Learn:  $\boldsymbol{\theta}$
- Objective: minimize  $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$

# Normal Equation

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix}$$

Objective: Minimize  $\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon}$

# Derivation of Normal Equation

$$\boldsymbol{\epsilon} = \mathbf{y} - \mathbf{X}\boldsymbol{\theta}$$

$$\boldsymbol{\epsilon}^\top \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})$$

$$= \mathbf{y}^\top \mathbf{y} - 2\mathbf{y}^\top \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$$

This is what we wish to minimize

## Minimizing the objective function

$$\frac{\partial \epsilon^\top \epsilon}{\partial \theta} = 0$$

- $\frac{\partial}{\partial \theta} \mathbf{y}^\top \mathbf{y} = 0$
- $\frac{\partial}{\partial \theta} (-2\mathbf{y}^\top \mathbf{X}\theta) = -2\mathbf{X}^\top \mathbf{y}$
- $\frac{\partial}{\partial \theta} (\theta^\top \mathbf{X}^\top \mathbf{X}\theta) = 2\mathbf{X}^\top \mathbf{X}\theta$

Substitute the values in the top equation



## Normal Equation derivation

$$\mathbf{0} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}$$

$$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta}$$

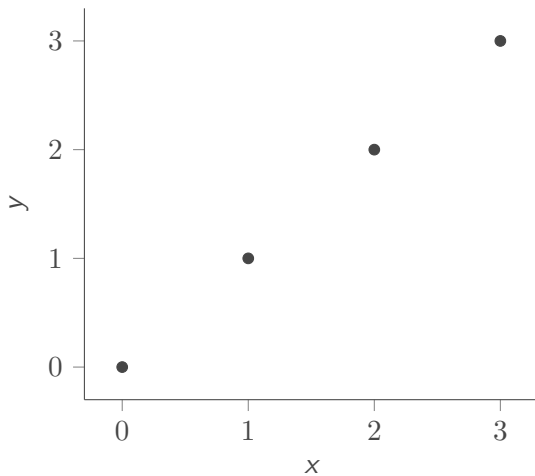
$$\hat{\boldsymbol{\theta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

## Worked out example

x	y
0	0
1	1
2	2
3	3

Given the data above, find  $\theta_0$  and  $\theta_1$ .

# Scatter Plot



## Worked out example

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$$

Given the data above, find  $\theta_0$  and  $\theta_1$ .

## Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}$$

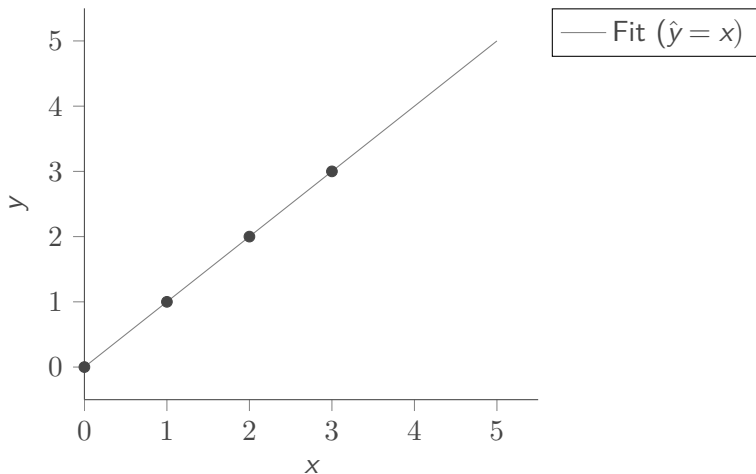
$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Worked out example

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{y})$$

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 14 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

# Scatter Plot



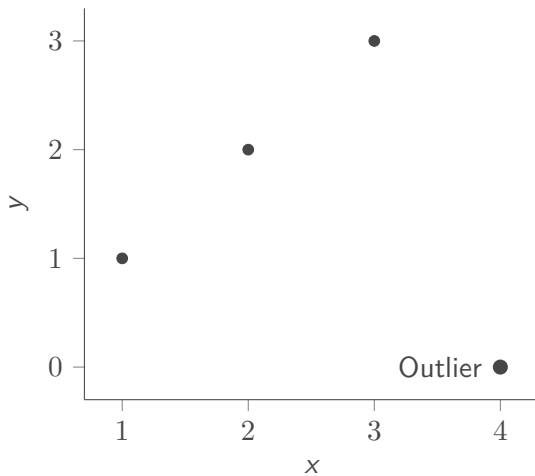
## Effect of outlier

x	y
1	1
2	2
3	3
4	0

Compute the  $\theta_0$  and  $\theta_1$ .



# Scatter Plot



## Worked out example

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}$$

$$\mathbf{X}^\top = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$\mathbf{X}^\top \mathbf{X} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

Given the data above, find  $\theta_0$  and  $\theta_1$ .

## Worked out example

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

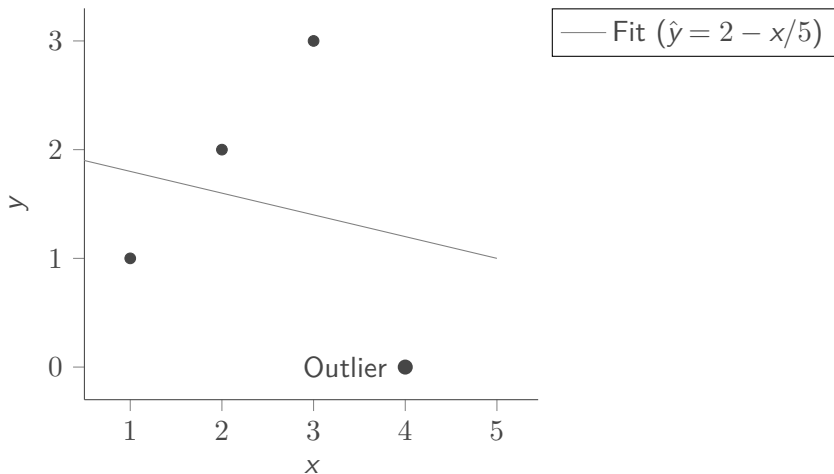
$$\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

## Worked out example

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1}(\mathbf{X}^\top \mathbf{y})$$

$$\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} = \begin{bmatrix} 2 \\ (-1/5) \end{bmatrix}$$

# Scatter Plot



# Basis Expansion

# Variable Transformation

Transform the data, by including the higher power terms in the feature space.

t	s
0	0
1	6
3	24
4	36

The above table represents the data before transformation

# Variable Transformation

Add the higher degree features to the previous table

t	$t^2$	s
0	0	0
1	1	6
3	9	24
4	16	36

- The above table represents the data after transformation
- Now, we can write  $\hat{s} = f(t, t^2)$
- Other transformations:  $\log(x)$ ,  $x_1 \times x_2$



## A big caveat: Linear in what?!<sup>1</sup>

1.  $\hat{s} = \theta_0 + \theta_1 * t$  is linear
2. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$  linear?
3. Is  $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2 + \theta_3 * \cos(t^3)$  linear?
4. Is  $\hat{s} = \theta_0 + \theta_1 * t + e^{\theta_2} * t$  linear?
5. All except #4 are linear models!
6. Linear refers to the relationship between the parameters that you are estimating ( $\theta$ ) and the outcome

---

<sup>1</sup><https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression>

# Basis Functions

- Linear regression only refers to linear in the parameters
- We can perform an arbitrary nonlinear transformation  $\phi(x)$  of the inputs  $x$  and then linearly combine the components of this transformation.
- $\phi : \mathbb{R}^D \rightarrow \mathbb{R}^K$  is called the basis function

# Basis Functions

Some examples of basis functions:

- Polynomial basis:  $\phi(x) = \{1, x, x^2, x^3, \dots\}$
- Fourier basis:  $\phi(x) = \{1, \sin(x), \cos(x), \sin(2x), \cos(2x), \dots\}$
- Gaussian basis:  
$$\phi(x) = \{1, \exp(-\frac{(x-\mu_1)^2}{2\sigma^2}), \exp(-\frac{(x-\mu_2)^2}{2\sigma^2}), \dots\}$$
- Sigmoid basis:  $\phi(x) = \{1, \sigma(x - \mu_1), \sigma(x - \mu_2), \dots\}$  where  
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

**Notebook:** [basis.html](#)

Interactive examples and visualizations of different basis functions

# Geometric Interpretation

# Linear Combination of Vectors

- Let  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  be vectors in  $\mathbb{R}^D$ , where  $D$  denotes the dimensions
- A linear combination of  $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_i$  is of the following form:

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \alpha_3 \mathbf{v}_3 + \dots + \alpha_i \mathbf{v}_i$$

where  $\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_i \in \mathbb{R}$

# Span of vectors

- Let  $v_1, v_2, \dots, v_i$  be vectors in  $\mathbb{R}^D$ , with  $D$  dimensions
- The span of  $v_1, v_2, \dots, v_i$  is denoted by  $\text{SPAN}\{v_1, v_2, \dots, v_i\}$ :

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_i v_i \mid \alpha_1, \alpha_2, \dots, \alpha_i \in \mathbb{R}\}$$

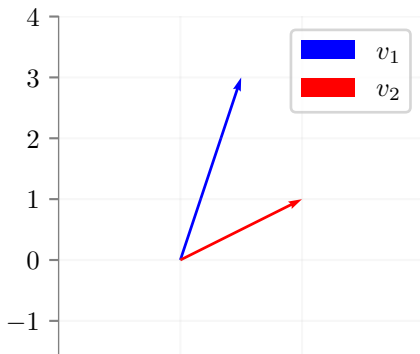
- It is the set of all vectors that can be generated by linear combinations of  $v_1, v_2, \dots, v_i$
- If we stack the vectors  $v_1, v_2, \dots, v_i$  as columns of a matrix  $V$ , then the span of  $v_1, v_2, \dots, v_i$  is given as  $V\alpha$  where  $\alpha \in \mathbb{R}^i$

## Example

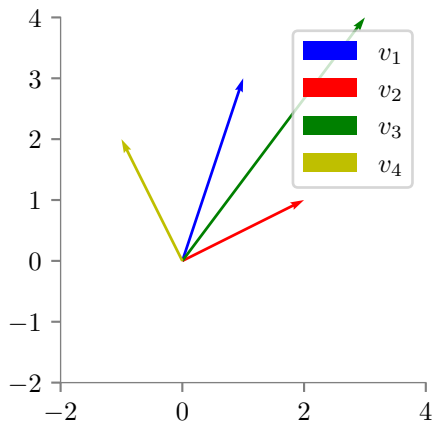
Find the span of  $\left(\begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \end{bmatrix}\right)$

**Notebook:** [geometric-linear-regression.html](#)

Interactive geometric visualization of vector spans and linear regression



## Example



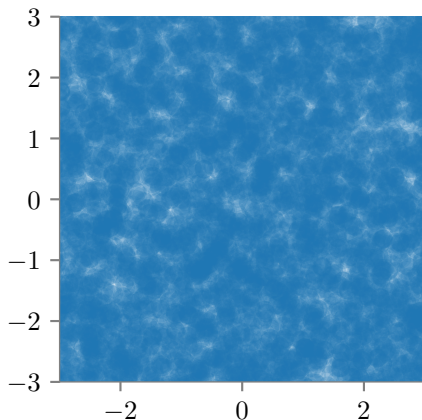
We have  $v_3 = v_1 + v_2$

We have  $v_4 = v_1 - v_2$



## Example

Simulating the above example in python using different values of  $\alpha_1$  and  $\alpha_2$

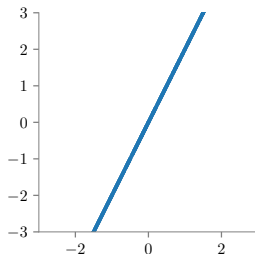


$$\text{Span}((v_1, v_2)) \in \mathcal{R}^2$$

## Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right)$

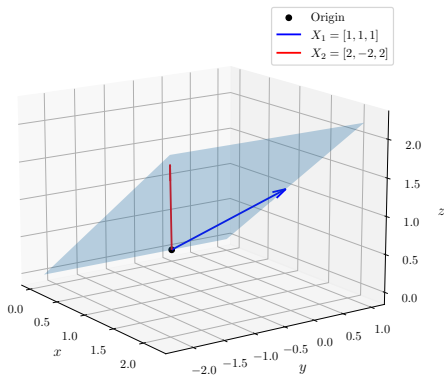
- Can we obtain a point  $(x, y)$  s.t.  $x = 3y$ ?
- No
- Span of the above set is along the line  $y = 2x$



# Example

Find the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$

- Visualization:



- The span is the plane  $z = x$  or  $x_3 = x_1$

# Geometric Interpretation

Consider  $\mathbf{X}$  and  $\mathbf{y}$  as follows.

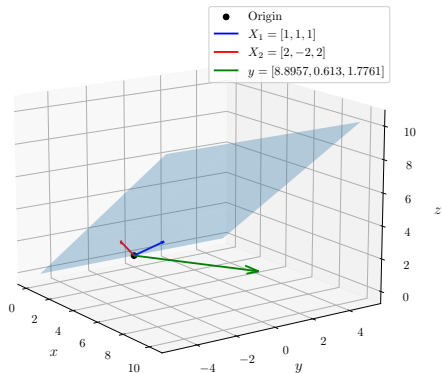
$$\mathbf{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

- We are trying to learn  $\boldsymbol{\theta}$  for  $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\theta}$  such that  $\|\mathbf{y} - \hat{\mathbf{y}}\|_2$  is minimised
- Consider the two columns of  $\mathbf{X}$ . Can we write  $\mathbf{X}\boldsymbol{\theta}$  as the span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$ ?
- We wish to find  $\hat{\mathbf{y}}$  such that

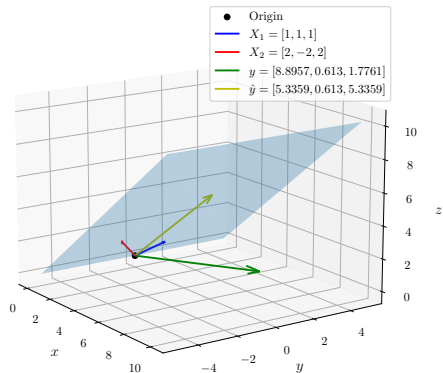
$$\arg \min_{\hat{\mathbf{y}} \in \text{SPAN}\{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_D\}} \|\mathbf{y} - \hat{\mathbf{y}}\|_2$$

# Geometric Interpretation

Span of  $\left( \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -2 \\ 2 \end{bmatrix} \right)$

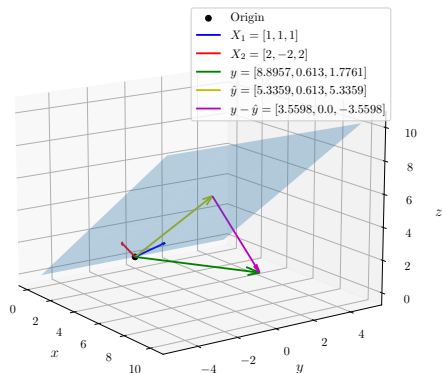


# Geometric Interpretation



- We seek a  $\hat{\mathbf{y}}$  in the span of the columns of  $\mathbf{X}$  such that it is closest to  $\mathbf{y}$

# Geometric Interpretation



- This happens when  $y - \hat{y} \perp x_j \forall j$  or  $x_j^\top (y - \hat{y}) = 0$
- $\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0$
- $\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\boldsymbol{\theta}$  or  $\hat{\boldsymbol{\theta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$

# Dummy Variables and Multicollinearity



# Multi-collinearity

- There can be situations where inverse of  $\mathbf{X}^\top \mathbf{X}$  is not computable
- This condition arises when the  $|\mathbf{X}^\top \mathbf{X}| = 0$  (determinant is zero)
- **Example:** Perfect multicollinearity

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} \quad (1)$$

- The matrix  $\mathbf{X}$  is not full rank (rank = 2, not 3)
- Notice: Column 3 = 2 × Column 2 (perfect linear dependence)
- **Cannot uniquely solve for  $\theta$ !**

# Multi-collinearity: Definition and Problems

## Definition: Multicollinearity

Arises when predictor variables/features in  $\mathbf{X}$  can be expressed as a linear combination of others

### Types:

- **Perfect:** Exact linear relationship (determinant = 0)
- **High:** Strong but not perfect correlation (determinant  $\approx 0$ )

### Problems caused:

- Unstable coefficient estimates (same data  $\rightarrow$  different  $\theta$ )
- High variance: Small changes in data  $\rightarrow$  Large changes in coefficients
- Can't interpret individual feature importance

# Why Multicollinearity Causes Instability

**The core problem:** Multiple parameter combinations give identical results

## Example: Simple Example

If  $x_2 = 2x_1$  exactly, then:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$y = \theta_0 + \theta_1 x_1 + \theta_2 (2x_1)$$

$$y = \theta_0 + (\theta_1 + 2\theta_2)x_1$$

- Many  $(\theta_1, \theta_2)$  pairs give same prediction
- $(\theta_1 = 1, \theta_2 = 0)$  and  $(\theta_1 = 3, \theta_2 = -1)$  both work!
- Small noise “chooses” randomly between solutions
- Result: Wildly different coefficients for same data

# Numerical Example: Why Coefficients Go Wild

**Dataset:** House prices with sq\_ft and sq\_m (perfectly correlated)

Price (\$k)	sq_ft	sq_m
200	2000	186
300	3000	279
400	4000	372

**What happens with tiny noise?**

- **Clean data:**  $(\theta_1, \theta_2) = (0.1, 0)$
- **Add 0.1% noise:**  $(\theta_1, \theta_2) = (-2.5, 28.0)$
- Same predictions, completely different coefficients!
- Which feature is “important”? Impossible to say!

## Key Points:

**Solutions:** Drop one variable, or use regularization (Ridge/Lasso)

# Dummy Variables: The Problem

**Example:** Pollution in Delhi = P

- Model specification:

$$P = \theta_0 + \theta_1 * \#Vehicles + \theta_2 * \text{Wind speed} + \theta_3 * \text{Wind Direction}$$

- But, wind direction is a categorical variable
- **Naive approach:** {N:0, E:1, W:2, S:3 }
- **Problem:** This incorrectly implies  $S > W > E > N$  (meaningless ordering!)
- Model assumes: S is “3 times better” than N for reducing pollution

# One-Hot Encoding (N-1 Variables)

**Correct approach:** Use binary indicators for each category

**N-1 encoding (recommended)**

Wind Direction	Is North?	Is East?	Is West?
North	1	0	0
East	0	1	0
West	0	0	1
South	0	0	0

## Key Points:

South is the **reference category** - all others are compared to it

# Why Not N Variables?

## Full encoding (problematic):

Wind	Is N?	Is E?	Is W?	Is S?
North	1	0	0	0
East	0	1	0	0
West	0	0	1	0
South	0	0	0	1

### Important: Multicollinearity Problem!

Notice:  $Is\_N + Is\_E + Is\_W + Is\_S = 1$  (always!)  
One column is perfectly predictable from the others

# N-1 vs N Encoding: The Dummy Variable Trap

## Why N-1 encoding is better:

- **N encoding problem:** Perfect multicollinearity
- Mathematical relationship:  $ls_S = 1 - (ls_N + ls_E + ls_W)$
- Matrix  $\mathbf{X}^T \mathbf{X}$  becomes non-invertible
- No unique solution exists!

### Example: The Dummy Variable Trap

Always use N-1 dummy variables for N categories.  
The omitted category becomes the **baseline/reference**.



# Binary Encoding

N	00
E	01
W	10
S	11

- W and S are related by one bit
- This introduces dependencies between them, and this can cause confusion in classifiers

# Interpreting Dummy variables

Gender	height
F	...
F	...
F	...
M	...
M	...

Encoding

Is Female	height
1	...
1	...
1	...
0	...
0	...

## Interpreting Dummy Variables

Is Female	height
1	5
1	5.2
1	5.4
0	5.8
0	6

- Model:  $height_i = \theta_0 + \theta_1 * (\text{Is Female}) + \epsilon_i$
- We get  $\theta_0 = 5.9$  and  $\theta_1 = -0.7$
- $\theta_0 = \mathbf{5.9}$ : Average height of males (reference category)
- $\theta_1 = \mathbf{-0.7}$ : Difference between female and male heights
- **Female height**  $= \theta_0 + \theta_1 = 5.9 + (-0.7) = 5.2$
- **Male height**  $= \theta_0 = 5.9$
- So  $\theta_1 = \text{Avg}(\text{female}) - \text{Avg}(\text{male}) = 5.2 - 5.9 = -0.7$

## Alternative Encoding: +1/-1 Scheme

Instead of 0/1, we could use +1/-1:

### Example: +1/-1 Encoding

$$x_i = \begin{cases} +1 & \text{if female} \\ -1 & \text{if male} \end{cases}$$

- Model:  $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$
- For females:  $y_i = \theta_0 + \theta_1 \cdot (+1) = \theta_0 + \theta_1$
- For males:  $y_i = \theta_0 + \theta_1 \cdot (-1) = \theta_0 - \theta_1$

### Key Points: Interpretation

- $\theta_0$  = overall average height across all people
- $\theta_1$  = half the difference between female and male heights

## Summary: Categorical Variable Encodings

Method	Good?	Variables	Issue
Ordinal (0,1,2,3)	No	1	Implies fake ordering
Full One-Hot	No	N	Multicollinearity
N-1 One-Hot	Yes	N-1	Recommended
Binary Encoding	Maybe	$\log(N)$	Artificial relationships
+1/-1 Encoding	Yes	1*	Only for 2 categories

### Definition: Best Practice

Use **N-1 one-hot encoding** for categorical variables.  
Choose the most common category as reference.

# Practice and Review

## Pop Quiz: Linear Regression

1. What is the geometric interpretation of least squares?
2. When does the normal equation have a unique solution?
3. How do polynomial features help with non-linear relationships?
4. What are the assumptions behind linear regression?

# Critical Assumptions of Linear Regression

**Before using linear regression, verify these assumptions:**

- **Linearity:** Relationship between  $x$  and  $y$  is linear
- **Independence:** Observations are independent of each other
- **Homoscedasticity:** Error variance is constant across all values of  $x$
- **Normality:** Errors are normally distributed (for inference)
- **No Multicollinearity:** Features are not highly correlated

**Violation Consequences:**

- Biased coefficient estimates
- Invalid confidence intervals
- Poor prediction performance



# Key Takeaways

- **Linear Model:** Assumes linear relationship between features and target
- **Least Squares:** Minimizes sum of squared residuals
- **Normal Equation:** Closed-form solution when  $\mathbf{X}^\top \mathbf{X}$  is invertible
- **Geometric View:** Projection onto column space of design matrix
- **Feature Engineering:** Basis expansion enables non-linear modeling
- **Foundation:** Building block for more complex models