Linear Regression

Nipun Batra and the teaching staff

January 25, 2024

IIT Gandhinagar

Setup

Linear Regression

- O/P is continuous in nature.

Linear Regression

- O/P is continuous in nature.
- Examples of linear systems:

Linear Regression

- O/P is continuous in nature.
- Examples of linear systems:
  - $F = ma$

Linear Regression

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

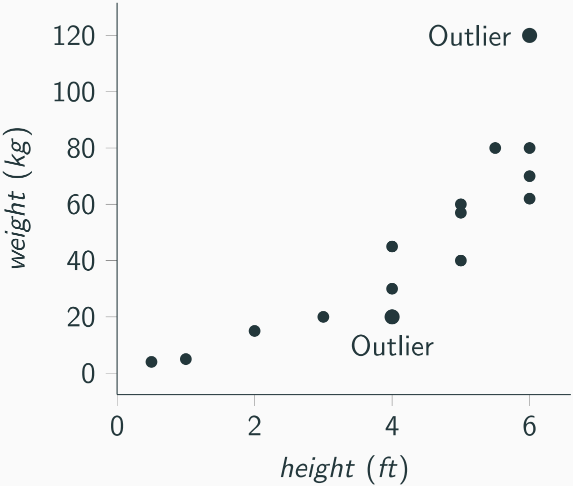
Task at hand

- TASK: Predict Weight = f(height)

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

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

Scatter Plot



Matrix representation of the expression

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

Matrix representation of the expression

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

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

Matrix representation of the expression

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

Matrix representation of the expression

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

$$W_{N\times 1} = X_{N\times 2}\theta_{2\times 1}$$

Matrix representation of the expression

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

$$W_{N\times 1} = X_{N\times 2}\theta_{2\times 1}$$

- θ_0 - Bias Term/Intercept Term

Matrix representation of the expression

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

$$W_{N\times 1} = X_{N\times 2}\theta_{2\times 1}$$

- θ_0 - Bias Term/Intercept Term
- θ_1 - Slope

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.

Examples in multiple dimensions.

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

Extension to multiple dimensions

In the previous example $y = f(x)$, where x is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

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

Extension to multiple dimensions

In the previous example y = f(x), where x is one-dimensional.

Examples in multiple dimensions.

One example is to predict the water demand of the IITGN campus

Demand = f(# occupants, Temperature)

Demand = Base Demand + K_1 * # occupants + K_2 * 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

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is
  $\hat{demand}_i = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$

Linear Relationship

We have

- $x_i = \begin{bmatrix} Temperature_i \\ \#Occupants_i \end{bmatrix}$
- Estimated demand for i^{th} sample is
  $\hat{demand_i} = \theta_0 + \theta_1 Temperature_i + \theta_2 Occupants_i$
- $\hat{demand_i} = x_i'^T \theta$

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 $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$

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 $x_i' = \begin{bmatrix} 1 \\ Temperature_i \\ \#Occupants_i \end{bmatrix} = \begin{bmatrix} 1 \\ x_i \end{bmatrix}$

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 θ_2 is likely to be positive

We can expect the following

- Demand increases, if # occupants increases, then θ_2 is likely to be positive
- Demand increases, if temperature increases, then θ_1 is likely to be positive

We can expect the following

- Demand increases, if # occupants increases, then θ_2 is likely to be positive
- Demand increases, if temperature increases, then θ_1 is likely to be positive
- Base demand is independent of the temperature and the # occupants, but, likely positive, thus θ_0 is likely positive.

Normal Equation

Generalized Linear Regression Format

- Assuming N samples for training

Generalized Linear Regression Format

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

Generalized Linear Regression Format

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

Generalized Linear Regression Format

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

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

Generalized Linear Regression Format

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

$$
\begin{bmatrix} \hat{y_1} \\ \hat{y_2} \\ \vdots \\ \hat{y_N} \end{bmatrix}_{N \times 1} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,M} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,M} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \dots & x_{N,M} \end{bmatrix}_{N \times (M+1)} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1) \times 1}
$$

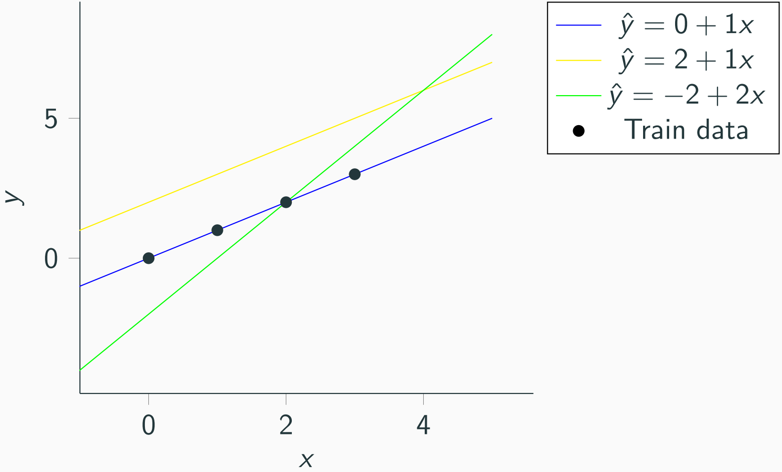
$$
\boxed{\hat{Y} = X\theta}
$$

Relationships between feature and target variables

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

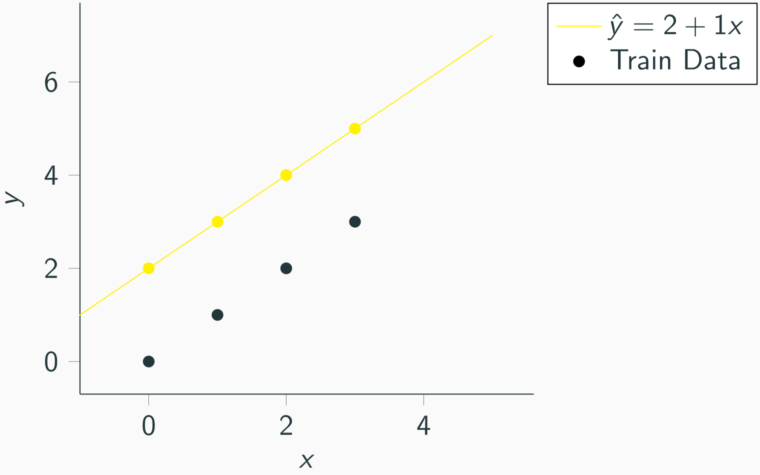
Relationships between feature and target variables

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



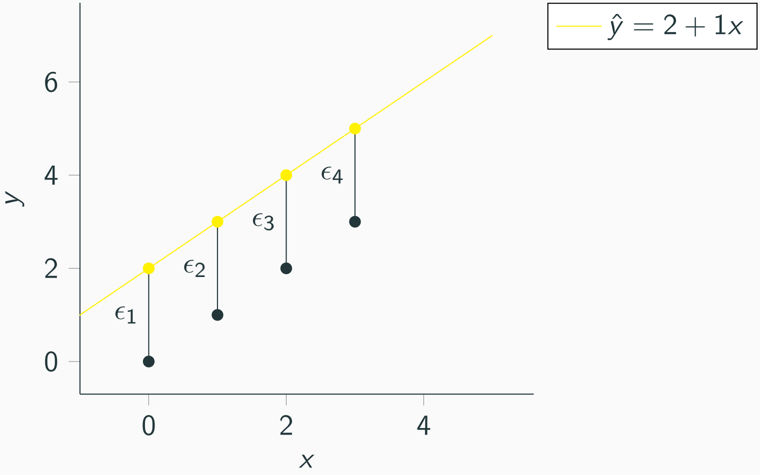
Relationships between feature and target variables

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 $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = x_i'^T \theta$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = x_i'^{T}\theta$
- ϵ_i denotes the error/residual for i^{th} sample

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = x_i'^T \theta$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = x_i'^T \theta$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$

Error terms

- $y_i = \hat{y}_i + \epsilon_i$ where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- y_i denotes the ground truth for i^{th} sample
- $\hat{y}_i$ denotes the prediction for i^{th} sample, where $\hat{y}_i = x_i'^T \theta$
- ϵ_i denotes the error/residual for i^{th} sample
- θ_0, θ_1: The parameters of the linear regression
- $\epsilon_i = y_i - \hat{y}_i$
- $\epsilon_i = y_i - (\theta_0 + x_i \times \theta_1)$

Good fit

- $|\epsilon_1|$, $|\epsilon_2|$, $|\epsilon_3|$, ... should be small.

Good fit

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

Good fit

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

Normal Equation

Normal Equation

$$Y = X\theta + \epsilon$$

Normal Equation

$$Y = X\theta + \epsilon$$

To Learn: θ

Normal Equation

$$Y = X\theta + \epsilon$$

To Learn: θ

Objective: minimize $\epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_N^2$

Normal Equation

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

Normal Equation

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

Objective: Minimize $\epsilon^T \epsilon$

Derivation of Normal Equation

$$\begin{aligned}
\epsilon &= y - X\theta \\
\epsilon^T &= (y - X\theta)^T = y^T - \theta^T X^T \\
\epsilon^T \epsilon &= (y^T - \theta^T X^T)(y - X\theta) \\
&= y^T y - \theta^T X^T y - y^T X\theta + \theta^T X^T X\theta \\
&= y^T y - 2y^T X\theta + \theta^T X^T X\theta
\end{aligned}$$

This is what we wish to minimize

Minimizing the objective function

$$\frac{\partial \epsilon^T \epsilon}{\partial \theta} = 0 \tag{1}$$

- $\dfrac{\partial}{\partial \theta} y^T y = 0$
- $\dfrac{\partial}{\partial \theta}(-2y^T X\theta) = (-2y^T X)^T = -2X^T y$
- $\dfrac{\partial}{\partial \theta}(\theta^T X^T X\theta) = 2X^T X\theta$

Substitute the values in the top equation

Normal Equation derivation

$$0 = -2X^T y + 2X^T X\theta$$

$$X^T y = X^T X\theta$$

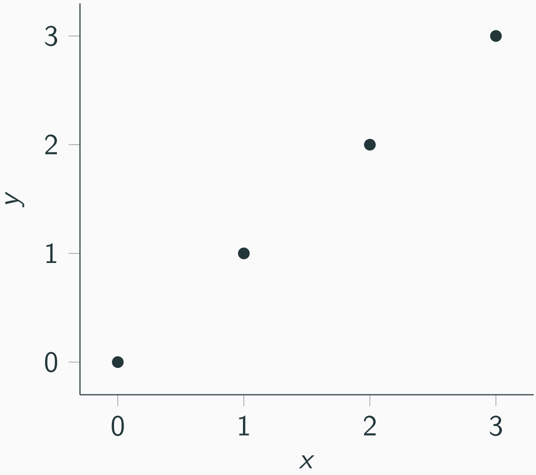
$$\boxed{\hat{\theta}_{OLS} = (X^T X)^{-1} X^T y}$$

Worked out example

x	y
0	0
1	1
2	2
3	3

Given the data above, find θ_0 and θ_1.

Scatter Plot



Worked out example

$$\begin{aligned}
X &= \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \\
X^T &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \\
X^T X &= \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}
\end{aligned} \tag{2}$$

Given the data above, find θ_0 and θ_1.

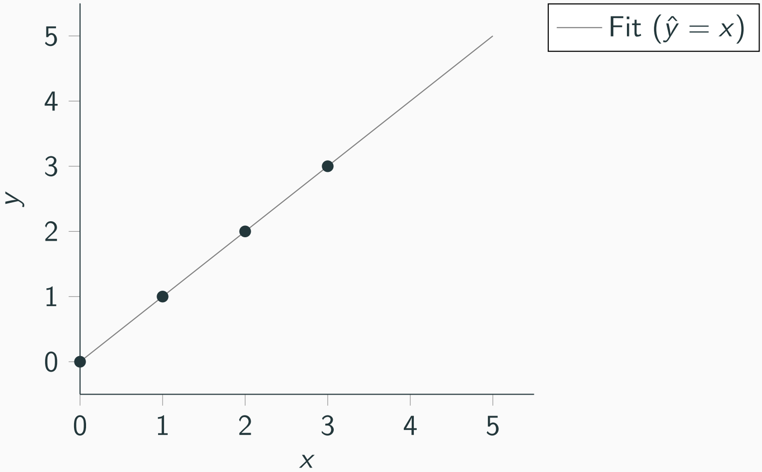
Worked out example

$$
\begin{aligned}
(X^T X)^{-1} &= \frac{1}{20}\begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \\
X^T y &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}
\end{aligned} \tag{3}
$$

Worked out example

$$
\begin{aligned}
\theta &= (X^T X)^{-1}(X^T y) \\
\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} &= \frac{1}{20}\begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}\begin{bmatrix} 6 \\ 14 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\end{aligned} \tag{4}
$$

Scatter Plot

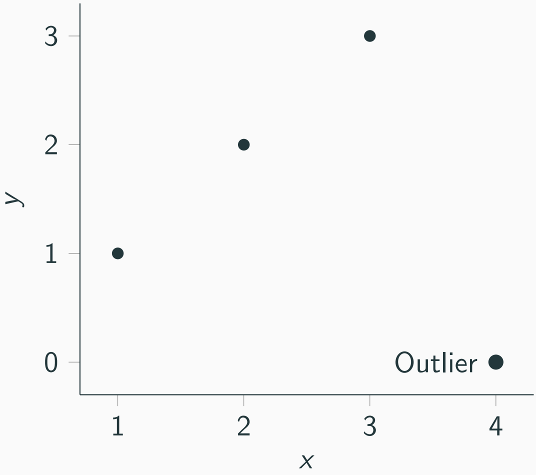


Effect of outlier

x	y
1	1
2	2
3	3
4	0

Compute the θ_0 and θ_1.

Scatter Plot



Worked out example

$$
\begin{aligned}
X &= \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \\
X^T &= \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \\
X^T X &= \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}
\end{aligned} \tag{5}
$$

Given the data above, find θ_0 and θ_1.

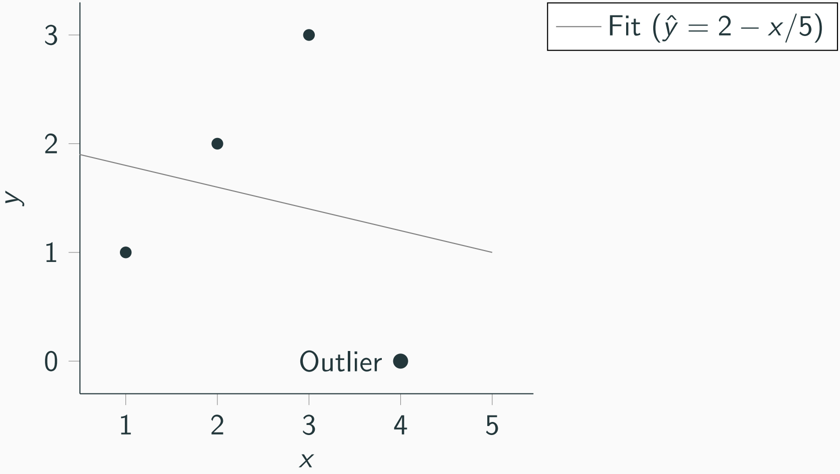
Worked out example

$$
\begin{aligned}
(X^T X)^{-1} &= \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix} \\
X^T y &= \begin{bmatrix} 6 \\ 14 \end{bmatrix}
\end{aligned} \tag{6}
$$

Worked out example

$$
\begin{aligned}
\theta &= (X^T X)^{-1}(X^T y) \\
\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} &= \begin{bmatrix} 2 \\ (-1/5) \end{bmatrix}
\end{aligned} \tag{7}
$$

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

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

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 $\hat{s} = f(t, t^2)$

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?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

1. $\hat{s} = \theta_0 + \theta_1 * t$ is linear
2. Is $\hat{s} = \theta_0 + \theta_1 * t + \theta_2 * t^2$ linear?

[1] `https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression`

A big caveat: Linear in what?![1]

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?

[1] `https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression`

A big caveat: Linear in what?![1]

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 $\hat{s} = \theta_0 + \theta_1 * t + \mathrm{e}^{\theta_2} * t$ linear?

[1] `https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression`

A big caveat: Linear in what?![1]

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 $\hat{s} = \theta_0 + \theta_1 * t + \mathrm{e}^{\theta_2} * t$ linear?
5. All except #4 are linear models!

[1] https://stats.stackexchange.com/questions/8689/what-does-linear-stand-for-in-linear-regression

A big caveat: Linear in what?![1]

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

[1] 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.
- $\phi : \mathbb{R}^D \to \mathbb{R}^K$ 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 $\sigma(x) = \frac{1}{1+e^{-x}}$

Geometric Interpretation

Linear Combination of Vectors

Let $v_1, v_2, v_3, \ldots, v_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.

Linear Combination of Vectors

Let $v_1, v_2, v_3, \ldots, v_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.
A linear combination of $v_1, v_2, v_3, \ldots, v_i$ is of the following form

Linear Combination of Vectors

Let $v_1, v_2, v_3, \ldots, v_i$ be vectors in $\mathbb{R}^D$, where D denotes the dimensions.
A linear combination of $v_1, v_2, v_3, \ldots, v_i$ is of the following form

$$\alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \cdots + \alpha_i v_i$$

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

Span of vectors

Let $v_1, v_2, \ldots, v_i$ be vectors in $\mathbb{R}^D$, with D dimensions.

Span of vectors

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

Span of vectors

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

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

Span of vectors

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

$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \ldots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \ldots, v_i$.

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 $\text{SPAN}\{v_1, v_2, \dots, v_i\}$

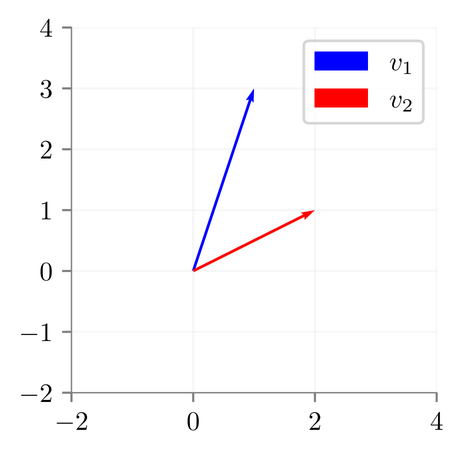
$$\{\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_i v_i \quad | \quad \alpha_1, \alpha_2, \dots, \alpha_i \in \mathbb{R}\}$$

It is the set of all vectors that can be generated by linear combinations of $v_1, v_2, \dots, v_i$.

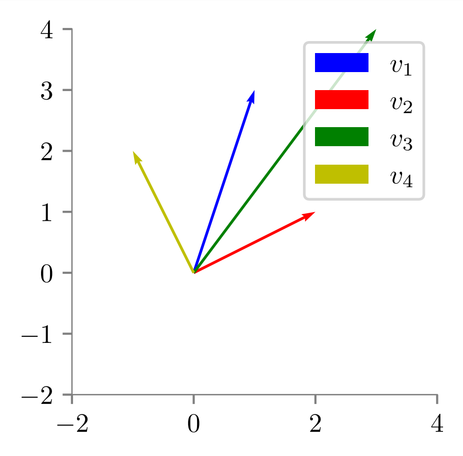
If we stack the vectors $v_1, v_2, \dots, v_i$ as columns of a matrix V, then the span of $v_1, v_2, \dots, v_i$ is given as $V\alpha$ where $\alpha \in \mathbb{R}^i$

Example

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



Example

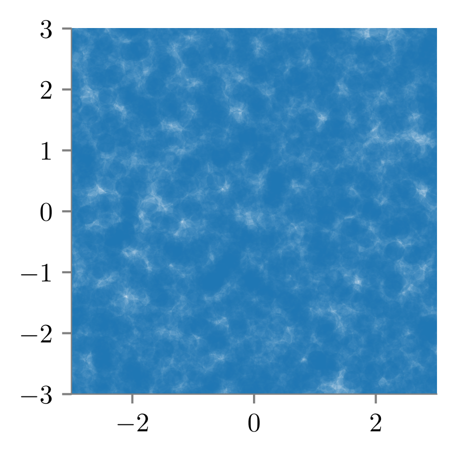


We have $v_3 = v_1 + v_2$

We have $v_4 = v_1 - v_2$

Example

Simulating the above example in python using different values of α_1 and α_2



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

Example

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

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. $x = 3y$?

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

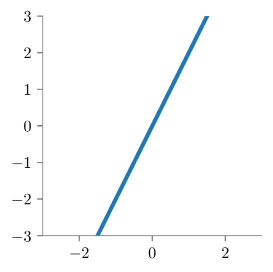
Example

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 $y = 2x$

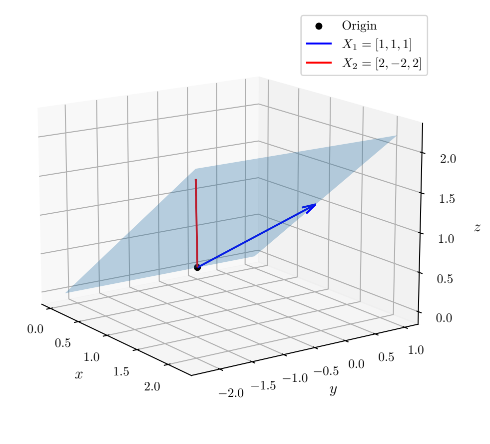


Example

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

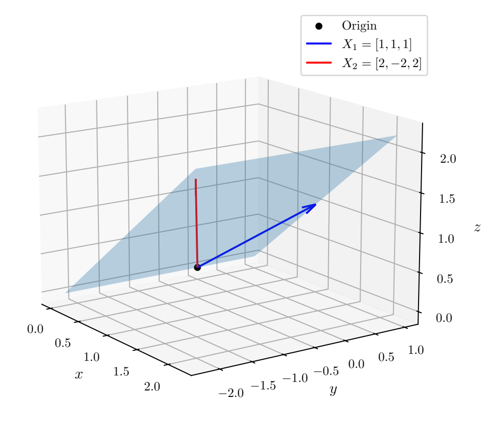
Example

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



Example

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



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

Geometric Interpretation

Consider X and y as follows.

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

- We are trying to learn θ for $\hat{y} = X\theta$ such that $||y - \hat{y}||_2$ is minimised

Geometric Interpretation

Consider X and y as follows.

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

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

Geometric Interpretation

Consider X and y as follows.

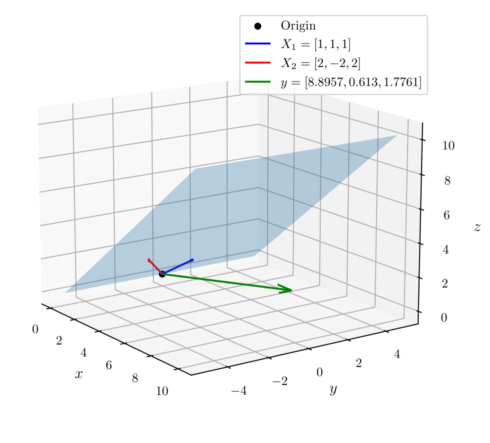
$$\mathrm{X} = \begin{pmatrix} 1 & 2 \\ 1 & -2 \\ 1 & 2 \end{pmatrix}, \quad \mathrm{y} = \begin{pmatrix} 8.8957 \\ 0.6130 \\ 1.7761 \end{pmatrix}$$

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

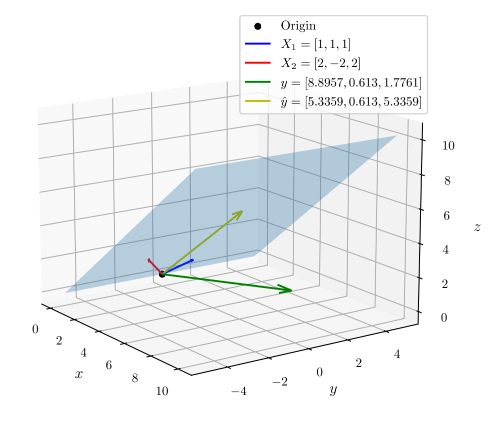
$$\underset{\hat{y} \in SPAN\{\bar{x_1}, \bar{x_2}, \ldots, \bar{x_D}\}}{\arg\min} ||y - \hat{y}||_2$$

Geometric Interpretation

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

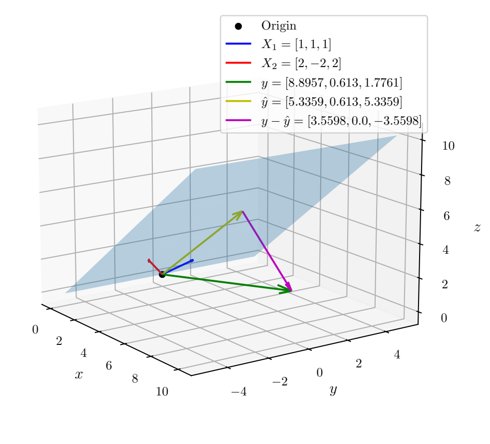


Geometric Interpretation



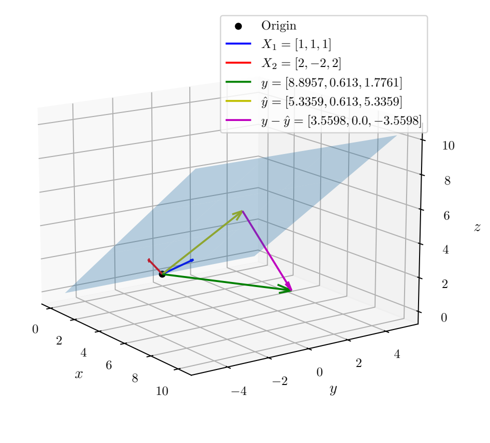
- We seek a $\hat{y}$ in the span of the columns of X such that it is closest to y

Geometric Interpretation



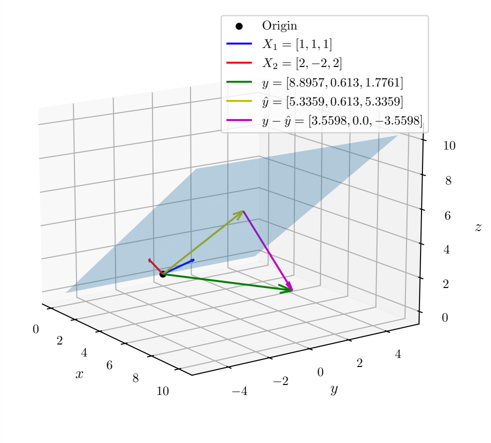
- This happens when $y - \hat{y} \perp x_j \forall j$ or $x_j^T (y - \hat{y}) = 0$

Geometric Interpretation



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

Geometric Interpretation



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