Northeastern University
Khoury College of
Computer Sciences

Review - Linear Algebra & Calculus | Linear Regression

DS 4400 | Machine Learning and Data Mining I

Zohair Shafi

Spring 2026

Monday | January 12, 2026

Today’s Outline

1. Linear Algebra
2. Calculus
3. Simple Linear Regression

Today’s Outline

1. **Linear Algebra**
2. Calculus
3. Simple Linear Regression

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

This is a system of m equations with n unknown parameters

$$a_{11}x_1 + a_{12}x_2 = b_1$$

$$a_{21}x_1 + a_{22}x_2 = b_2$$

$$a_{31}x_1 + a_{32}x_2 = b_3$$

$$a_{41}x_1 + a_{42}x_2 = b_4$$

Linear Algebra

Systems of Linear Equations

- Consider the equation $Ax = b$
  - $A \in \mathbb{R}^{m \times n}$
  - $x \in \mathbb{R}^{n \times 1}$
  - $b \in \mathbb{R}^{m \times 1}$

This is a system of m equations with n unknown parameters

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 &= b_1 \\
a_{21}x_1 + a_{22}x_2 &= b_2 \\
a_{31}x_1 + a_{32}x_2 &= b_3 \\
a_{41}x_1 + a_{42}x_2 &= b_4
\end{aligned}
\longrightarrow
\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \\ a_{41} & a_{42} \end{bmatrix}
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} =
\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}
$$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$(1) \cdot w_0 + (450) \cdot w_1 = 2000$$

$$(1) \cdot w_0 + (510) \cdot w_1 = 2100$$

$$(2) \cdot w_0 + (980) \cdot w_1 = 2400$$

$$(3) \cdot w_0 + (1500) \cdot w_1 = 3000$$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$(1) \cdot w_0 + (450) \cdot w_1 = 2000$$
$$(1) \cdot w_0 + (510) \cdot w_1 = 2100$$
$$(2) \cdot w_0 + (980) \cdot w_1 = 2400$$
$$(3) \cdot w_0 + (1500) \cdot w_1 = 3000$$

$$\longrightarrow \begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}$$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$
\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad \longrightarrow \quad
\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix}
\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}
=
\begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}
$$

These are all the vector of x_0

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$
\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad \longrightarrow \quad
\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix}
\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}
=
\begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}
$$

These are all the vector of x_1

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$
\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad \longrightarrow \quad
\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix}
\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}
=
\begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}
$$

This is the matrix for input data $X \in \mathbb{R}^{4\times2}$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$
\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\longrightarrow
\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix}
\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}
=
\begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}
$$

This is the vector of training data or labels y

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$
\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad \longrightarrow \quad
\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix}
\begin{bmatrix} w_0 \\ w_1 \end{bmatrix}
=
\begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}
$$

Finally, the matrix of learnable weights W

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$(1) \cdot w_0 + (450) \cdot w_1 = 2000$$
$$(1) \cdot w_0 + (510) \cdot w_1 = 2100$$
$$(2) \cdot w_0 + (980) \cdot w_1 = 2400$$
$$(3) \cdot w_0 + (1500) \cdot w_1 = 3000$$

$\longrightarrow$

$X \in \mathbb{R}^{4\times2}$ $W \in \mathbb{R}^{2\times1}$ $y \in \mathbb{R}^{4\times1}$

$$\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}$$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

$$X \in \mathbb{R}^{4\times 2} \qquad W \in \mathbb{R}^{2\times 1} \qquad y \in \mathbb{R}^{4\times 1}$$

$$\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad = \quad
X \cdot W = y$$

Linear Algebra

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

$$X \in \mathbb{R}^{4\times 2} \qquad W \in \mathbb{R}^{2\times 1} \qquad y \in \mathbb{R}^{4\times 1}$$

$$(1) \cdot w_0 + (450) \cdot w_1 = 2000$$

$$(1) \cdot w_0 + (510) \cdot w_1 = 2100$$

$$(2) \cdot w_0 + (980) \cdot w_1 = 2400$$

$$(3) \cdot w_0 + (1500) \cdot w_1 = 3000$$

$$= \qquad X \cdot W = y$$

This is a **matrix product!**

Linear Algebra

Linear Independence and Rank of a Matrix

- A set of vectors are **linearly independent** if none of them can be written as a **linear** combination of each other.

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}$$

The vectors $\vec{u}, \vec{v}$ are **not** linearly independent since $\vec{v} = 2 \cdot \vec{u}$

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad \vec{v} = \begin{bmatrix} 7 \\ 9 \\ 13 \end{bmatrix}$$

The vectors $\vec{u}, \vec{v}$ are linearly independent

Linear Algebra

Linear Independence and Rank of a Matrix

- $rank(A)$ is given:
  - Number of linearly independent columns
  - Number of linearly independent rows
- There's a related fact about the upper bound on rank:
- $rank(A) \leq min(m, n)$ - upper bound on rank
  - For an $m \times n$ matrix, you can't have more than m independent rows (there are only m of them) or more than n independent columns.
  - So the rank is bounded by whichever dimension is smaller.
- A matrix has full rank if $rank(A) = min(m, n)$

Linear Algebra

Inverse of Matrices

- For a **square** matrix, the inverse matrix A^{-1} is a matrix such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I$$

A matrix is only invertible if it has **full rank**

- Some properties of inverse:
  - $(A^{-1})^{-1} = A$
  - $(AB)^{-1} = B^{-1}A^{-1}$
  - $(A^T)^{-1} = (A^{-1})^T$

Linear Algebra

Inverse of Matrices

- For a **square** matrix, the inverse matrix A^{-1} is a matrix such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I$$

- If a given matrix A is invertible, then we can find a solution for the equation of the form $Ax = b$ as:

$$Ax = b$$

$$A^{-1}Ax = A^{-1}b$$

$$Ix = A^{-1}b$$

$$x = A^{-1}b$$

Linear Algebra

Inverse of Matrices

- For a **square** matrix, the inverse matrix A^{-1} is a matrix such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I$$

- If a given matrix A is invertible, then we can find a solution for the equation of the form $Ax = b$ as $x = A^{-1}b$

- For the linear regression problem $X \cdot W = y$, a potential solution can be

$$W = X^{-1}y$$

Linear Algebra

Inverse of Matrices

- For a **square** matrix, the inverse matrix A^{-1} is a matrix such that:

$$A \cdot A^{-1} = A^{-1} \cdot A = I$$

- If a given matrix A is invertible, then we can find a solution for the equation of the form $Ax = b$ as $x = A^{-1}b$

- For the linear regression problem $X \cdot W = y$, a potential solution can be

$$W = X^{-1}y$$

BUT

For this to hold true, X must be a **square matrix** and **invertible**, which is rarely the case

Linear Algebra

Determinants

- The determinant $det(A)$ of a matrix A is a **scalar** value of a matrix
- It determines whether a matrix is invertible and how it scales volume
- Some properties:
  - $det(A) \neq 0 \implies A$ is invertible
  - $det(AB) = det(A) \cdot det(B)$
  - $det(A^T) = det(A)$
  - $det(A^{-1}) = \frac{1}{det(A)}$

Linear Algebra
Determinants

- The determinant $det(A)$ of a matrix A is a **scalar** value of a matrix
- It determines whether a matrix is invertible and how it scales volume
- Some properties:
  - $det(A) \neq 0 \implies A$ is invertible
  - $det(AB) = det(A) \cdot det(B)$
  - $det(A^T) = det(A)$
  - $det(A^{-1}) = \frac{1}{det(A)}$

For a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$$det(A) = a \cdot d - b \cdot c$$

Linear Algebra

Eigenvalues and Eigenvectors

- For a **square** matrix A:

$$Av = \lambda v$$

v is a non-zero vector

Linear Algebra

Eigenvalues and Eigenvectors

- For a **square** matrix A:

$$Av = \lambda v$$

v **is a non-zero vector**

λ **is a scalar value**

Linear Algebra

Eigenvalues and Eigenvectors

- For a **square** matrix A:

$$Av = \lambda v$$

v **is a non-zero vector** $\Longrightarrow$ **Eigenvector**

λ **is a scalar value** $\Longrightarrow$ **Eigenvalue**

Interpretation

The Eigenvector v is a vector such that when multiplied by the matrix A,
the vector remains the **same** but is **scaled** by the Eigenvalue λ

Linear Algebra

Eigenvalues and Eigenvectors

- Some properties - $Av = \lambda v$
  - To find Eigenvalues: $det(A - \lambda I) = 0$
  - $det(A) = \Pi\ \lambda_i$ i.e., $det(A)$ is the product of all Eigenvalues of A
  - For **Symmetric** matrices (i.e., $A = A^T$)

$$A = V \Lambda V^T$$

where V is the matrix of all Eigenvectors
and Λ is a diagonal matrix with all Eigenvalues

Linear Algebra

Eigenvalues and Eigenvectors

- Some properties - $Av = \lambda v$
  - To find Eigenvalues: $det(A - \lambda I) = 0$
  - $det(A) = \Pi\ \lambda_i$ i.e., $det(A)$ is the product of all Eigenvalues of A
  - For **Symmetric** matrices (i.e., $A = A^T$)

$$A = V\Lambda V^T$$

This is also called the Eigen**decomposition** of a matrix A since you are decomposing the matrix into two matrices V and Λ

where V is the matrix of all Eigenvectors
and Λ is a diagonal matrix with all Eigenvalues

Today’s Outline

1. Linear Algebra
2. Calculus
3. Simple Linear Regression

Today’s Outline

1. Linear Algebra
2. **Calculus**
3. Simple Linear Regression

Calculus

Derivatives

- The derivative $f'(x) = \frac{df}{dx}$ measures the instantaneous rate of change of the function f at the input x

$$f'(x) = lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

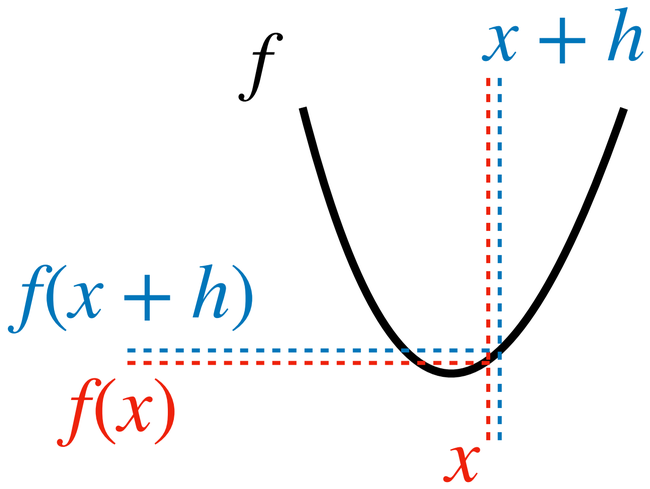
Calculus

Derivatives

- The derivative $f'(x) = \frac{df}{dx}$ measures the instantaneous rate of change of the function f at the input x

$$f'(x) = lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

h **is an infinitesimally small value**



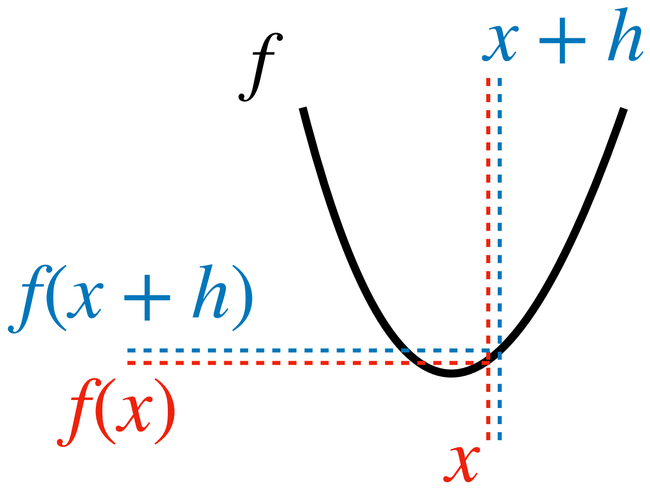
Calculus

Derivatives

- The derivative $f'(x) = \frac{df}{dx}$ measures the instantaneous rate of change of the function f at the input x

$$f'(x) = lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

h is an infinitesimally small value



If $Speed = \frac{Distance}{Time}$, then the **derivative** of speed is **acceleration**, i.e., rate of change of speed

Calculus

Common Derivatives

- $f(x) = x^n$

$$\frac{df}{dx} = n \cdot x^{n-1}$$

- $f(x) = e^x$

$$\frac{df}{dx} = e^x$$

- $f(x) = log(x)$

$$\frac{df}{dx} = \frac{1}{x}$$

- $f(x) = sin(x)$

$$\frac{df}{dx} = cos(x)$$

- $f(x) = cos(x)$

$$\frac{df}{dx} = -\ sin(x)$$

Calculus

Derivative Rules

- Sum Rule

$$(f+g)' = f' + g'$$

- Product Rule

$$(fg)' = fg' + f'\ g$$

- Quotient Rule

$$(\frac{f}{g})' = \frac{f'g - fg'}{g^2}$$

- Chain Rule
  - If $y = f(g(x))$, then

$$\frac{dy}{dx} = f'(g(x)) \cdot g'(x)$$

Calculus

Partial Derivatives

- For a function $f(x_1, x_2, x_3, x_4)$, the partial derivative $\frac{\partial f}{\partial x_i}$ is the derivative with respect to x_i **only** while leaving all other variables as **constant**
- Example: $f(x, y) = x^2y + 3xy^2$

Calculus

Partial Derivatives

- For a function $f(x_1, x_2, x_3, x_4)$, the partial derivative $\frac{\partial f}{\partial x_i}$ is the derivative with respect to x_i **only** while leaving all other variables as **constant**
- Example: $f(x, y) = x^2y + 3xy^2$

$$\frac{\partial f}{\partial x} = 2xy + 3y^2$$

$$\frac{\partial f}{\partial y} = x^2 + 6xy$$

Calculus

Gradient

- The gradient ∇f is a **vector** of all partial derivatives of a function

Given a function $f(x_1, x_2, x_3, x_4)$

$$\nabla f = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \frac{\partial f}{\partial x_4}]$$

- Properties
  - Gradient points in the direction of steepest ascent
  - Negative gradient points in the direction of steepest descent
  - At a minimum or maximum point, $\nabla f = 0$

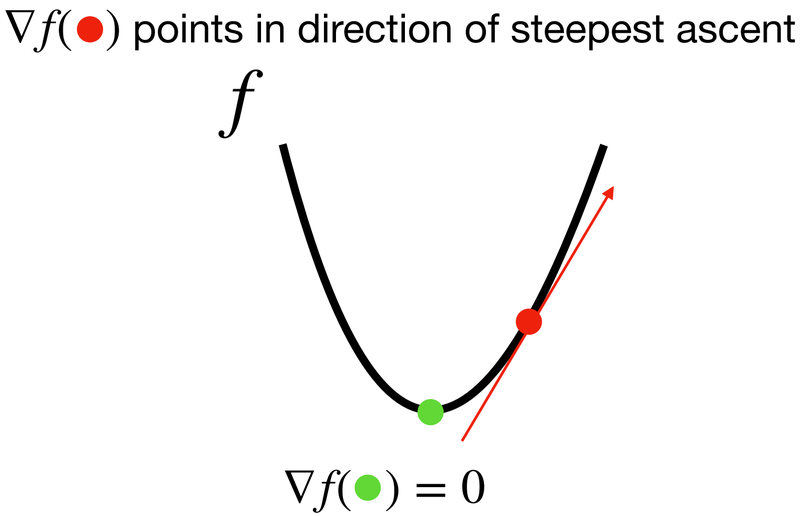
Calculus

Gradient

- The gradient ∇f is a **vector** of all partial derivatives of a function

Given a function $f(x_1, x_2, x_3, x_4)$

$$\nabla f = [\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3}, \frac{\partial f}{\partial x_4}]$$



- Properties
  - Gradient points in the direction of steepest ascent
  - Negative gradient points in the direction of steepest descent
  - At a **minimum** or **maximum** point, $\nabla f = 0$

Calculus

Gradient

- If you have multiple functions
  $y_1 = f_1(x_1, x_2, x_3)$
  $y_2 = f_2(x_1, x_2, x_3)$
  $y_3 = f_3(x_1, x_2, x_3)$
- Then the **Jacobian** is defined as

$$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \frac{\partial y_1}{\partial x_3} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_2}{\partial x_3} \\ \frac{\partial y_3}{\partial x_1} & \frac{\partial y_3}{\partial x_2} & \frac{\partial y_3}{\partial x_3} \end{bmatrix}$$

Calculus

Some Properties of Vector and Matrix Gradients

- If $\vec{x} \in \mathbb{R}^{d \times 1}$ and $\vec{v} \in \mathbb{R}^{d \times 1}$ are two vectors,

$$\frac{\partial v^T x}{\partial x} = v^T$$

- If $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^{d \times 1}$

$$\frac{\partial Ax}{\partial x} = A$$

- If $A \in \mathbb{R}^{d \times d}$ (**square**) and $x \in \mathbb{R}^{d \times 1}$

$$\frac{\partial x^T Ax}{\partial x} = (A + A^T)x$$

If A is also symmetric, then

$$\frac{\partial x^T Ax}{\partial x} = 2Ax$$

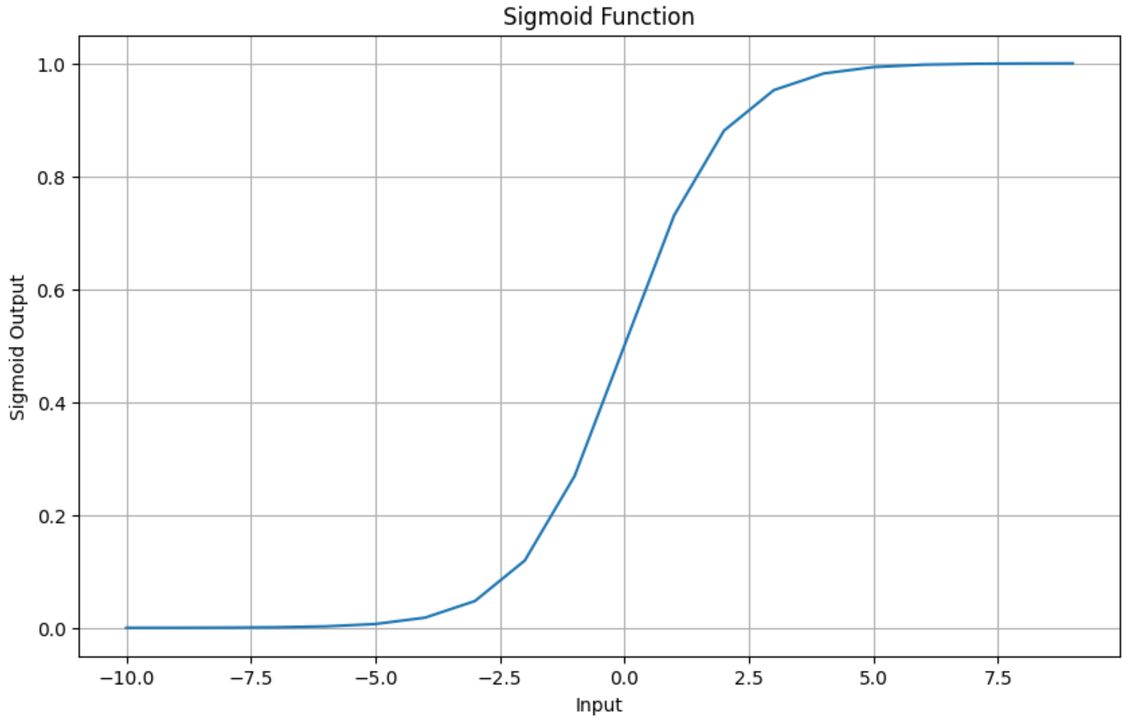
- If $x \in \mathbb{R}^{d \times 1}$

$$\frac{\partial \|x\|^2}{\partial x} = 2x^T$$

Calculus

Exercise

- Sigmoid: $\sigma(x) = \dfrac{1}{1 + e^{-x}}$
- Derivative $\sigma'(x) =$ **?**



Calculus

Exercise

- Sigmoid: $\sigma(x) = \frac{1}{1 + e^{-x}}$
- Derivative :

$$\text{Let } f(x) = 1 + e^{-x} \text{ and } g(x) = \frac{1}{x} = x^{-1}$$

$$\sigma(x) = g(f(x))$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\text{Use chain rule } \frac{dy}{dx} = f'(g(x)) \cdot g'(x)$$

Calculus

Exercise

- Sigmoid: $\sigma(x) = \frac{1}{1 + e^{-x}}$
- Derivative :

$$\text{Let } f(x) = 1 + e^{-x} \text{ and } g(x) = \frac{1}{x} = x^{-1}$$

$$\sigma(x) = g(f(x))$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\sigma'(x) = -1 \cdot (1 + e^{-x})^{-2} \cdot -e^{-x}$$

Calculus

Exercise

- Sigmoid: $\sigma(x) = \frac{1}{1 + e^{-x}}$
- Derivative :

Let $f(x) = 1 + e^{-x}$ and $g(x) = \frac{1}{x} = x^{-1}$

$$\sigma(x) = g(f(x))$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\sigma'(x) = -1 \cdot (1 + e^{-x})^{-2} \cdot -e^{-x}$$

Calculus

Exercise

- Sigmoid: $\sigma(x) = \frac{1}{1 + e^{-x}}$
- Derivative :

$$\text{Let } f(x) = 1 + e^{-x} \text{ and } g(x) = \frac{1}{x} = x^{-1}$$

$$\sigma(x) = g(f(x))$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\sigma'(x) = -1 \cdot (1 + e^{-x})^{-2} \cdot -e^{-x}$$

$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

Calculus

Exercise

- Sigmoid: $\sigma(x) = \frac{1}{1 + e^{-x}}$
- Derivative :

$$\text{Let } f(x) = 1 + e^{-x} \text{ and } g(x) = \frac{1}{x} = x^{-1}$$

$$\sigma(x) = g(f(x))$$

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\sigma'(x) = -1 \cdot (1 + e^{-x})^{-2} \cdot -e^{-x}$$

$$\sigma'(x) = \frac{e^{-x}}{(1 + e^{-x})^2}$$

$$\sigma'(x) = \sigma(x) \cdot \frac{e^{-x}}{1 + e^{-x}}$$

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x))$$

Today’s Outline

1. Linear Algebra
2. Calculus
3. Simple Linear Regression

Today’s Outline

1. Linear Algebra
2. Calculus
3. **Simple Linear Regression**

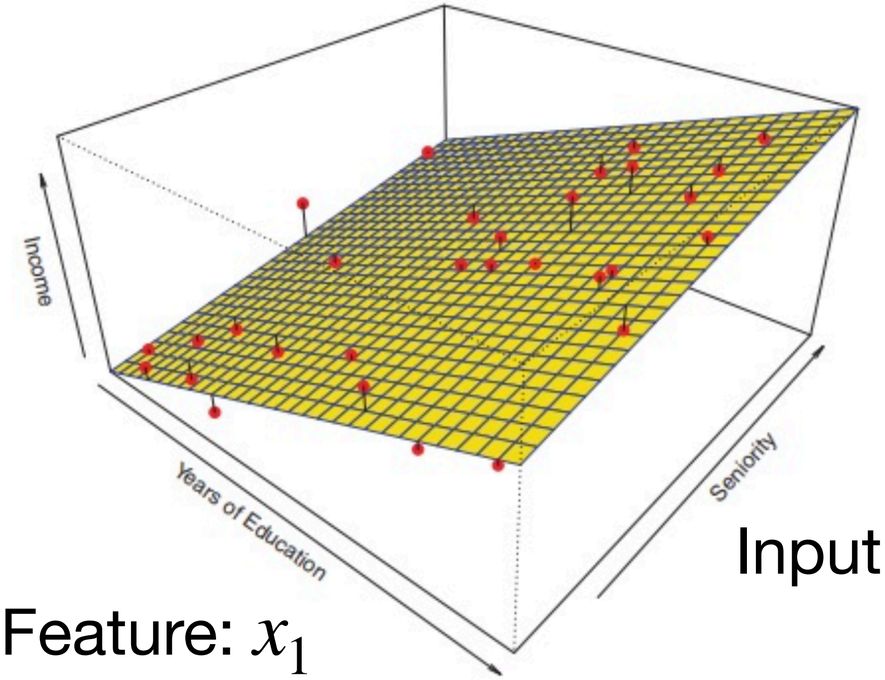
Linear Regression

- The oldest statistical learning method (Legendre and Gauss 1805)
- One of the most widely used techniques
- Fundamental to many complex models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to find optimal solution in closed form
- Efficient to find a solution using practical algorithms like gradient descent

Linear Regression

Example Task - Income Prediction

True label: y

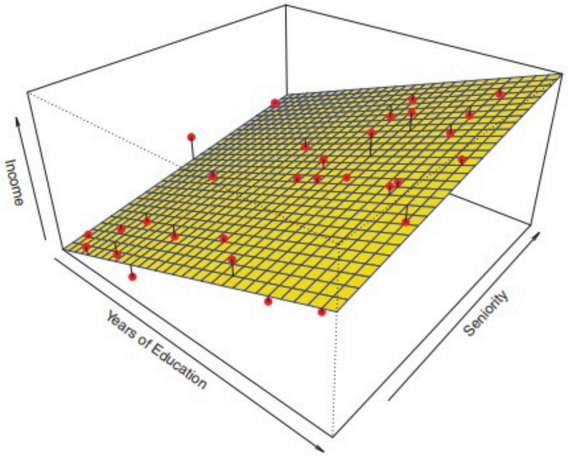


Input Feature: x_1

Input Feature: x_0

Linear Regression

Example Task - Income Prediction

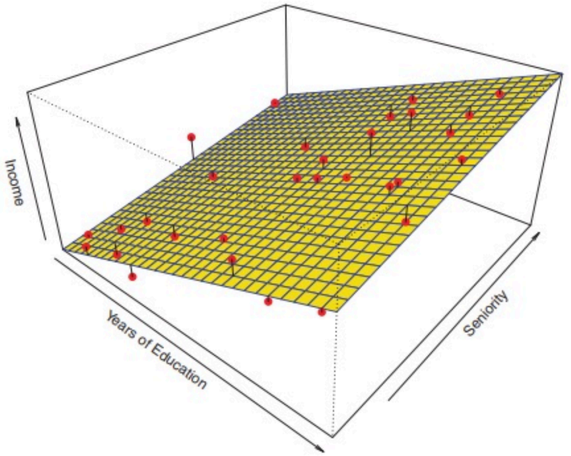


- Linear Model

$$f_\theta(x) = \theta_0 + \theta_1 x_0 + \theta_2 x_1$$

Linear Regression

Example Task - Income Prediction

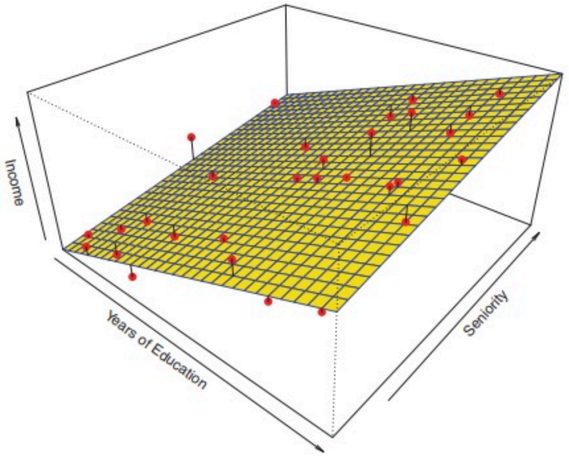


- Linear Model

$$f_\theta(x) = \theta_0 + \theta_1 x_0 + \theta_2 x_1$$

Learnable parameters

Linear Regression



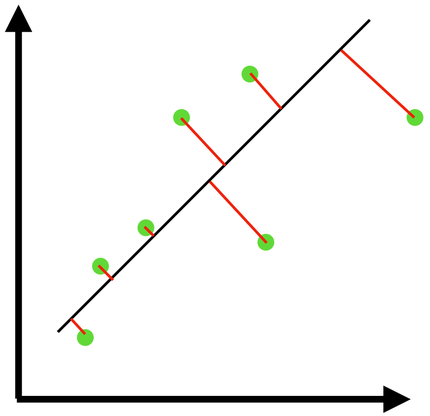
- Linear Model

$$f_\theta(x) = \theta_0 + \theta_1 x_0 + \theta_2 x_1$$

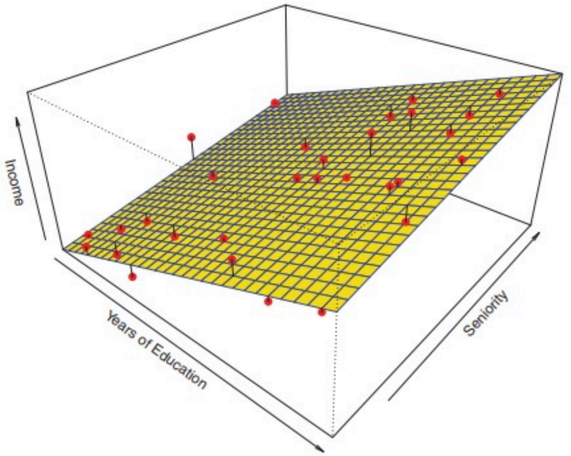
- Loss Functions (also called Cost Functions)

The red lines are called residuals

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2$$ - Mean Squared Error



Linear Regression



- Linear Model

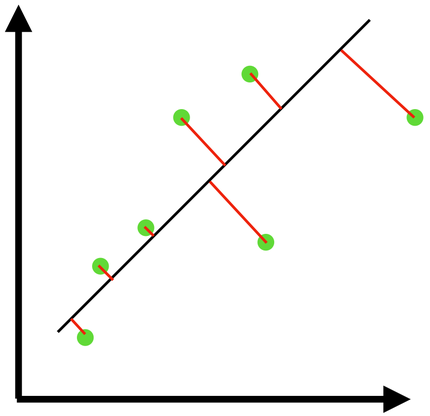
$$f_\theta(x) = \theta_0 + \theta_1 x_0 + \theta_2 x_1$$

- Loss Functions (also called Cost Functions)

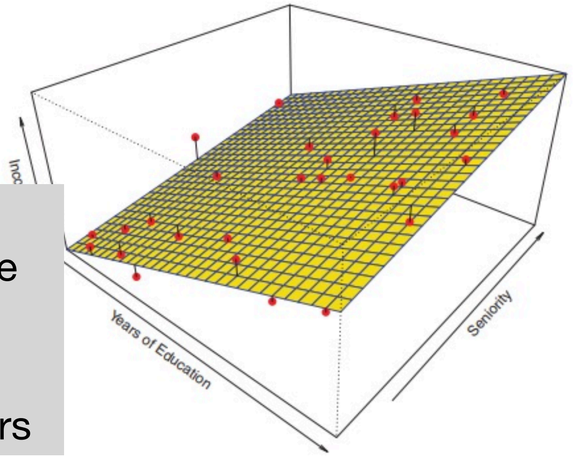
The red lines are called residuals

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2$$ - Mean Squared Error

$$L(\theta) = \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2$$ - Residual Sum of Squares



Linear Regression



Side Note

$f_\theta(x)$ and $\hat{y}$ are generally used interchangeably and are used to denote the predicted value

Similarly, w and θ are used interchangeably to denote learnable parameters

- Linear Model

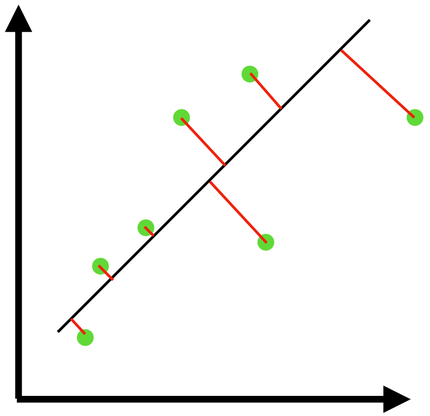
$$f_\theta(x) = \theta_0 + \theta_1 x_0 + \theta_2 x_1$$

- Loss Functions (also called Cost Functions)

The red lines are called residuals

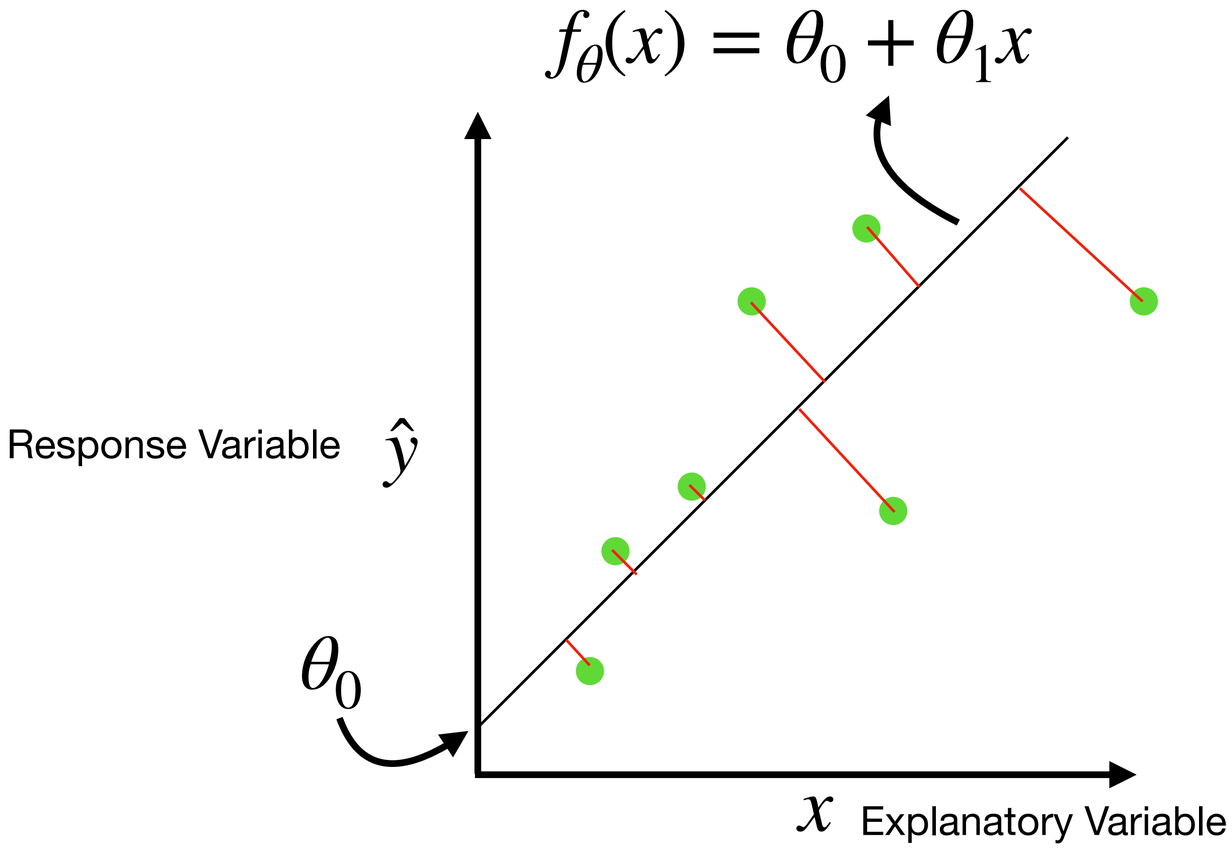
$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2 \text{ - Mean Squared Error}$$

$$L(\theta) = \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2 \text{ - Residual Sum of Squares}$$



Linear Regression

- Linear Model



Linear Regression

- Linear Model

$$f_\theta(x) = \theta_0 + \theta_1 x$$

- How do we find the solution to this? How do we find the optimal θ?
  - We optimize θ to minimize the loss function

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} [f_\theta(x_i) - y_i]^2$$

$$L(\theta) = \frac{1}{m} \sum_{i=1}^{m} [\theta_0 + \theta_1 \cdot x - y_i]^2$$

Linear Regression

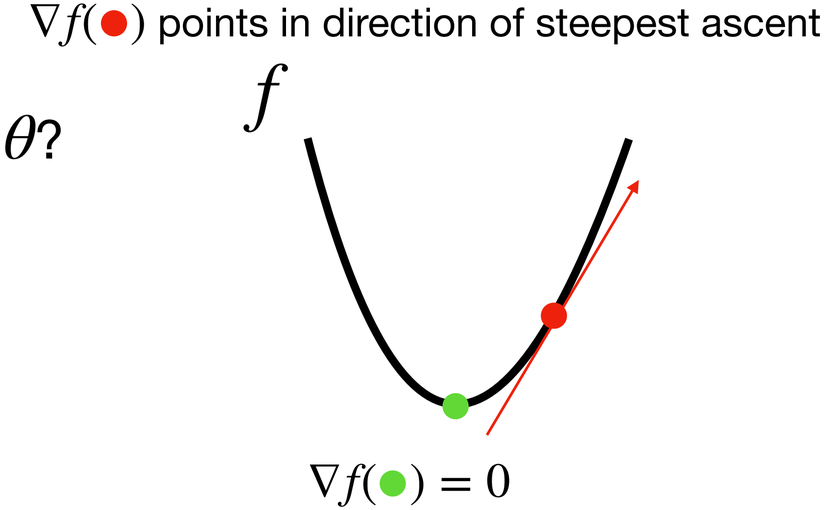
- Linear Model

$$f_\theta(x) = \theta_0 + \theta_1 x$$

- How do we find the solution to this? How do we find the optimal θ?
  - We optimize θ to minimize the loss function

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}[f_\theta(x_i) - y_i]^2$$

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}[\theta_0 + \theta_1 \cdot x - y_i]^2$$



Linear Regression

- How do we find the solution to this? How do we find the optimal θ?
  - We optimize θ to minimize the loss function

$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}[f_\theta(x_i) - y_i]^2$$

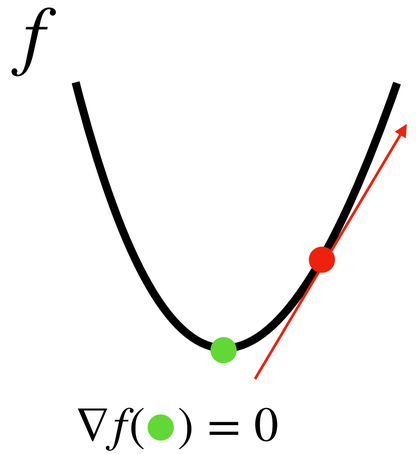
$$L(\theta) = \frac{1}{m}\sum_{i=1}^{m}[\theta_0 + \theta_1 \cdot x_i - y_i]^2$$

Find the point where $\nabla L(\theta) = 0$

$$\frac{\partial L(\theta)}{\partial \theta_0} = \frac{2}{m}\sum_{i=1}^{m}(\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$\nabla f(\bullet)$ points in direction of steepest ascent



Linear Regression

Finding θ_0

$$\frac{\partial L(\theta)}{\partial \theta_0} = \frac{2}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i) = 0$$

Linear Regression

Finding θ_0

$$\frac{\partial L(\theta)}{\partial \theta_0} = \frac{2}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i) = 0$$

We can ignore $\frac{1}{m}$ since its simply a scaling factor here

$$\sum_i y_i = \sum_i \theta_0 + \theta_1 \sum_i x_i$$

Linear Regression

Finding θ_0

$$\frac{\partial L(\theta)}{\partial \theta_0} = \frac{2}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i) = 0$$

We can ignore $\frac{1}{m}$ since its simply a scaling factor here

$$\sum_i y_i = \sum_i \theta_0 + \theta_1 \sum_i x_i$$

Since $\sum_i^m \theta_0 = m\theta_0$

$$\sum_i y_i = m\theta_0 + \theta_1 \sum_i x_i$$

Linear Regression

Finding θ_0

$$\frac{\partial L(\theta)}{\partial \theta_0} = \frac{2}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x_i - y_i) = 0$$

We can ignore $\frac{1}{m}$ since its simply a scaling factor here

$$\sum_i y_i = \sum_i \theta_0 + \theta_1 \sum_i x_i$$

Since $\sum_i^m \theta_0 = m\theta_0$

$$\sum_i y_i = m\theta_0 + \theta_1 \sum_i x_i$$

Divide both sides my m

$$\bar{y} = \theta_0 + \theta_1 \bar{x}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m} \sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m} \sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m} \sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m} \sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

$$\sum_i x_i y_i - \bar{y} \sum_i x_i = \theta_1 (\bar{x} \sum_i x_i + \sum_i x_i^2)$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

$$\sum_i x_i y_i - \bar{y} \sum_i x_i = \theta_1 (\bar{x} \sum_i x_i + \sum_i x_i^2)$$

Since $\sum_i^m \theta_1 = m\theta_1$

$$\sum_i x_i y_i - m\bar{x}\bar{y} = \theta_1 (\sum_i x_i^2 - m\bar{x}^2)$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

$$\sum_i x_i y_i - \bar{y} \sum_i x_i = \theta_1 (\bar{x} \sum_i x_i + \sum_i x_i^2)$$

Since $\sum_i^m \theta_1 = m\theta_1$

$$\sum_i x_i y_i - m\bar{x}\bar{y} = \theta_1 (\sum_i x_i^2 - m\bar{x}^2)$$

$$\theta_1 = \frac{\sum_i x_i y_i - m\bar{x}\bar{y}}{(\sum_i x_i^2 - m\bar{x}^2)}$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

$$\sum_i x_i y_i - \bar{y} \sum_i x_i = \theta_1 (\bar{x} \sum_i x_i + \sum_i x_i^2)$$

Since $\sum_i^m \theta_1 = m\theta_1$

$$\sum_i x_i y_i - m\bar{x}\bar{y} = \theta_1 (\sum_i x_i^2 - m\bar{x}^2)$$

$$\theta_1 = \frac{\sum_i x_i y_i - m\bar{x}\bar{y}}{(\sum_i x_i^2 - m\bar{x}^2)}$$

This can be rewritten as

$$\theta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Linear Regression

Finding θ_1

$$\frac{\partial L(\theta)}{\partial \theta_1} = \frac{2}{m}\sum_{i=1}^{m} x_i \cdot (\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\sum_i x_i y_i = \theta_0 \sum_i x_i + \theta_1 \sum_i x_i^2$$

Substitue $\theta_0 = \bar{y} - \theta_1 \bar{x}$

$$\sum_i x_i y_i = (\bar{y} - \theta_1 \bar{x}) \sum_i x_i + \theta_1 \sum_i x_i^2$$

$$\sum_i x_i y_i - \bar{y} \sum_i x_i = \theta_1 (\bar{x} \sum_i x_i + \sum_i x_i^2)$$

Since $\sum_i^m \theta_1 = m\theta_1$

$$\sum_i x_i y_i - m\bar{x}\bar{y} = \theta_1 (\sum_i x_i^2 - m\bar{x}^2)$$

$$\theta_1 = \frac{\sum_i x_i y_i - m\bar{x}\bar{y}}{(\sum_i x_i^2 - m\bar{x}^2)}$$

This can be rewritten as

$$\theta_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Which is

$$\theta_1 = \frac{Cov(x, y)}{Var(x)}$$

Linear Regression

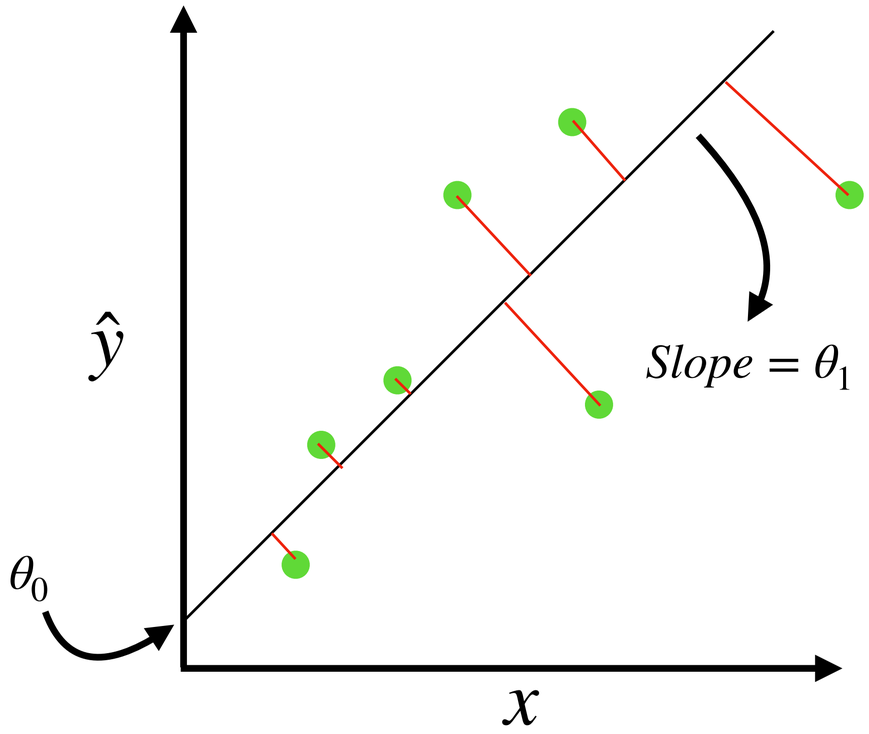
$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_1 = \frac{Cov(x, y)}{Var(x)}$$

$$f_\theta(x) = \theta_0 + \theta_1 x$$

The slope $\theta_1 = \frac{Cov(x, y)}{Var(x)}$ makes sense:

- If x and y covary strongly (move together), the slope is steeper
- If x has high variance (spread out), the slope is gentler
- The **sign** of covariance determines if the line goes up or down



Linear Regression

Solutions in Matrix Form

- Let's look at the matrix formulation of the same problem

$$L(\theta) = \frac{1}{m}\sum_i (y_i - \hat{y}_i)^2$$

But in matrix form, $f_\theta(x) = \hat{Y} = X\theta$, where $X \in \mathbb{R}^{m \times d}$ has m rows of data and d columns of features and $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \in \mathbb{R}^{d \times 1}$

$$L(\theta) = \frac{1}{m}\sum (Y - X\theta)^2$$

(think back to system of equations for why this is true)

Quick Recap

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

y	x_0	x_1
Price	# Rooms	Sq. Ft.
2000	1	450
2100	1	510
2400	2	980
3000	3	1500

$$(1) \cdot w_0 + (450) \cdot w_1 = 2000$$
$$(1) \cdot w_0 + (510) \cdot w_1 = 2100$$
$$(2) \cdot w_0 + (980) \cdot w_1 = 2400$$
$$(3) \cdot w_0 + (1500) \cdot w_1 = 3000$$

$\longrightarrow$

$X \in \mathbb{R}^{4\times2}$ $W \in \mathbb{R}^{2\times1}$ $y \in \mathbb{R}^{4\times1}$

$$\begin{bmatrix} 1 & 450 \\ 1 & 510 \\ 2 & 980 \\ 3 & 1500 \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} = \begin{bmatrix} 2000 \\ 2100 \\ 2400 \\ 3000 \end{bmatrix}$$

Quick Recap

Systems of Linear Equations - Linear Regression Example

- Consider the equation $y = w_0 x_0 + w_1 x_1$

$$X \in \mathbb{R}^{4\times 2} \qquad W \in \mathbb{R}^{2\times 1} \qquad y \in \mathbb{R}^{4\times 1}$$

$$\begin{aligned}
(1) \cdot w_0 + (450) \cdot w_1 &= 2000 \\
(1) \cdot w_0 + (510) \cdot w_1 &= 2100 \\
(2) \cdot w_0 + (980) \cdot w_1 &= 2400 \\
(3) \cdot w_0 + (1500) \cdot w_1 &= 3000
\end{aligned}
\quad = \quad
X \cdot W = y$$

Linear Regression

Expanded Loss Function in Matrix Form

$$L(\theta) = \frac{1}{m} \sum (Y - X\theta)^2$$

Linear Regression

Expanded Loss Function in Matrix Form

$$L(\theta) = \frac{1}{m} \sum (Y - X\theta)^2$$

$$L(\theta) = (Y - X\theta)^T (Y - X\theta)$$

(why is this true?)

Linear Regression

Expanded Loss Function in Matrix Form

$$L(\theta) = \frac{1}{m} \sum (Y - X\theta)^2$$

$$L(\theta) = (Y - X\theta)^T (Y - X\theta)$$

(why is this true?)

$$L(\theta) = (Y^T - \theta^T X^T)(Y - X\theta)$$

(because $(AB)^T = B^T A^T$)

Linear Regression

Expanded Loss Function in Matrix Form

$$L(\theta) = \frac{1}{m}\sum (Y - X\theta)^2$$

$$L(\theta) = (Y - X\theta)^T(Y - X\theta)$$

(why is this true?)

$$L(\theta) = (Y^T - \theta^T X^T)(Y - X\theta)$$

(because $(AB)^T = B^T A^T$)

$$L(\theta) = Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta$$

(the two terms in the centre are equivalent, why?)

Linear Regression

Expanded Loss Function in Matrix Form

$$L(\theta) = \frac{1}{m}\sum (Y - X\theta)^2$$

$$L(\theta) = (Y - X\theta)^T(Y - X\theta)$$

(why is this true?)

$$L(\theta) = (Y^T - \theta^T X^T)(Y - X\theta)$$

(because $(AB)^T = B^T A^T$)

$$L(\theta) = Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta$$

(the two terms in the centre are equivalent, why?)

$$L(\theta) = Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta$$

Linear Regression

Derivative

$$L(\theta) = Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta$$

Linear Regression

Derivative

$$L(\theta) = Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta$$

For any vector A

$$\nabla_A (B^T A) = B$$

For any symmetric matrix B

$$\nabla_A (A^T B A) = 2BA$$

Linear Regression

Derivative

$$L(\theta) = Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta$$

For any vector A

$$\nabla_A (B^T A) = B$$

For any symmetric matrix B

$$\nabla_A (A^T B A) = 2BA$$

So we have

$$\nabla L(\theta) = -2X^T Y + 2X^T X\theta$$

Linear Regression

Derivative

$$L(\theta) = Y^T Y - 2\theta^T X^T Y + \theta^T X^T X\theta$$

For any vector A

$$\nabla_A (B^T A) = B$$

For any symmetric matrix B

$$\nabla_A (A^T B A) = 2BA$$

So we have

$$\nabla L(\theta) = -2X^T Y + 2X^T X\theta$$

Linear Regression

Solution

We want to find the minimum so set gradient to zero

$$\nabla L(\theta) = -2X^T Y + 2X^T X\theta = 0$$

Linear Regression

Solution

We want to find the minimum so set gradient to zero

$$\nabla L(\theta) = -2X^T Y + 2X^T X\theta = 0$$

$$2X^T X\theta = 2X^T Y$$

$$X^T X\theta = X^T Y$$

Linear Regression

Solution

We want to find the minimum so set gradient to zero

$$\nabla L(\theta) = -2X^T Y + 2X^T X\theta = 0$$

$$2X^T X\theta = 2X^T Y$$

$$X^T X\theta = X^T Y$$

If $X^T X$ is invertible, then

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

Linear Regression

Regression vs Correlation

- **Correlation**
  - Find a numerical value expressing the relationship between variables
  - Measures linear dependence
- **Regression**
  - Estimate values of response variable on the basis of the values of predictor variable
  - The slope of linear regression is related to correlation coefficient
  - Regression scales to more than 2 variables, but correlation does not

Practical Example

https://zohairshafi.github.io/pages/lectures/Lecture_2_Notebook.ipynb

Conclusion

- We went through some more linear algebra and calculus
- We defined linear regression
  - We derived the optimal solution for linear regression and mean squared error loss
  - We derived the optimal solution for linear regression and mean squared loss in the matrix form
  - We saw some practical examples of linear regression in a Jupyter notebook
- Next Class:
  - Practical issues, feature normalization and gradient descent