Least Squares Optimization

First some more Linear Algebra

L2 Norm: $||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$

L1 Norm: $||x||_1 = \sqrt{\sum_{i=1}^{n} |x_i|}$

Infinity norm: $||x||_{inf} = max_i |x_i|$

General P Norm: $||x||_p = \left(\sum_{i=1}^{n} x_i^p \right)^{1/p}$

Matrix Norm: $||x||_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}^2} = \sqrt{tr(A^T A)}$

Inner or dot product

$$x \cdot y = ||x||||y||cos\theta$$

- If y is a unit vector then $x \cdot y$ gives the length of x which lies in the direction of y

$$x^T y = x \cdot y = \begin{bmatrix} x_1 \ldots x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i$$

Matrix Rank

- col-rank(A) = maximum number of linearly independent column vectors of A
- row-rank(A) = maximum number of linearly independent row vectors of A. Column rank always equals row rank
- For transformation matrices, the rank tells you the dimensions of the output.

  for instance, if rank of A is 1, then the transformation $p' = Ap$ points onto a line.
- Full rank matrix - if A is m x m and rank is m

Vector spaces

- It is a generalization of the cartesian plane.

  - A cartesian plane is a set of all points (x,y), where x and y are real numbers
- In Computer Science points (x,y) can be thought of as numeric arrays of size 2.
- We can have a set of numeric arrays of size d, denoted as $\mathbb{R}^d$

  $\mathbb{R}$ denotes that each array is composed of real numbers,
  d denotes the number of components in each array
- Each element in $\mathbb{R}^d$ is represented as $[x_1, x_2, \ldots, x_d]$

Vector spaces - Matrices

- An $n \times m$ matrix has n rows and m columns.

- A set of all $n \times m$ matrices, can be denoted as $\mathbb{R}^{n \times m}$ and is a vector space

Basis and Dimensionality

- Point (x, y) on the cartesian plane can be thought of as "x units along x-axis and y units along y-axis"

  $(x, y) = x(1,0) + y(0,1)$

  Any vector, $\mathbf{v}$, in the cartesian plane $\mathbb{R}^2$ can be represented as a linear combination of only two vectors, (1,0) and (0, 1).

- In vector space $\mathbb{R}^d$, consider the vectors $e_1 = [1,0,0,\ldots,0]$, $e_2 = [0,1,0,\ldots,0]$, $e_3 = [0,0,1,\ldots,0]$ and so on.

  e_i has 1 in the i-th position and 0 everywhere else.

Basis and Dimensionality

- Any vector, $\mathbf{v} = [x_1, x_2, \ldots, x_d]$, then $\mathbf{v} = \sum_i x_i e_i$, can be represented as a linear combination of basis vectors, e_i's.

- Vector space of all $n \times m$ images, $\mathbb{R}^{n \times m}$
  Every element in this vector space is an $n \times m$ matrix.

- The matrix e_{ij} has a 1 in the (i, j)-th position and is zero everywhere else.

- Then any vector $\mathbf{v} = \begin{bmatrix} x_{11} & x_{12} & \ldots x_{1m} & \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}$ in $\mathbb{R}^{n \times m}$ can be written as $\sum_{i,j} x_{ij} e_{ij}$

Basis and Dimensionality

- Set of vectors $B_2 = \{(1,0),(0,1)\}$ in $\mathbb{R}^2$, the set

  $B_d = \{e_i; i = 1,\ldots,d\}$ in $\mathbb{R}^d$ and the set

  $B_{n\times m} = \{e_{ij}; i = 1,\ldots,n; j = 1,\ldots,m\}$ in $\mathbb{R}^{n\times m}$ are special.

- Definition: Let V be a vector space, and suppose $U \subset V$. Then a vector $\mathbf{v} \in V$ is said to be in the span of U if it is a linear combination of the vectors in U, that is, if $\mathbf{v} = \sum_{i=1}^{n} \alpha_i u_i$ for some $\mathbf{u}_i \in U$ and some scalars α_i. The span of U is the set of all such vectors $\mathbf{v}$ which can be expressed as linear combinations of vectors in U

  In $\mathbb{R}^2$, the span of the set $B_2 = (0,1),(1,0)$ is all of $\mathbb{R}^2$, since every vector in $\mathbb{R}^2$ can be expressed as a linear combination of vectors in B.

Basis and Dimensionality

- Definition: Let V be a vector space. A set of vectors $U = \mathbf{u}_1, \ldots, \mathbf{u}_n \subset V$ is linearly depended if there exist scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$, not all of them 0, such that $\sum_{i=1}^{n} \alpha_i u_i = \mathbf{u}$. If no such α_i's exist, then the set of vectors is U is linearly independent.

  For example, the set $(1,0), (0,1), (1,-1)$. Then, because $(1,0) - (0,1) - (1,-1) = \mathbf{0}$ This set is linearly dependent.

- Another equivalent definition for a linearly independent set is that no vector in the set is a linear combination of the others.

Basis

- Let V be a vector space. A set of vectors $U \subset V$ is a basis for V if:
    - The span of U is V, that is, every vector in V can be written as a linear combination of vectors from U, and
    - U is linearly independent.

  thus B_2 is a basis for $\mathbb{R}^2$, B_d is the basis for $\mathbb{R}^d$ and $B_{n\times m}$ is a basis for $\mathbb{R}^{n\times m}$

  ** Note that a given vector space can have more than a single basis. For example $B_2' = (0,1),(1,1)$ is also a basis for $\mathbb{R}^2$, however, *all basis sets for a vector space have the same number of elements*

  The number of elements in a basis for a vector space is called the dimensionality of the vector space.

Linear Transformation

- A function $f : U \to V$ is a linear transformation if $f(\alpha \mathbf{u}_1 + \beta \mathbf{u}_2) = \alpha f(\mathbf{u}_1) + \beta f(\mathbf{u}_2)$ for all scalars α, β and for all $\mathbf{u}_1, \mathbf{u}_2 \in U$

  Suppose we have fixed a basis $B_U = \mathrm{b}_1, \ldots, \mathrm{b}_m$ for U, and a basis $B_V = \mathrm{a}_1, \ldots, \mathrm{a}_n$ for V

  $f(b_j) = \sum_{i=1}^{n} M_{ij} a_i$ for some coefficients M_{ij}

  Consider an arbitrary vector u.
  It can be expressed as a linear combination of the basis vectors, so

  $\mathrm{u} = \sum_{j=1}^{m} u_j \mathrm{b}_j$, therefore

  $$f(\mathrm{u}) = f(\sum_{j=1}^{m} u_j \mathrm{b}_j) = \sum_{j=1}^{m} u_j f(\mathrm{b}_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} M_{ij} u_j \mathrm{a}_i$$

  Every linear transformation can be expressed as matrix multiplication.

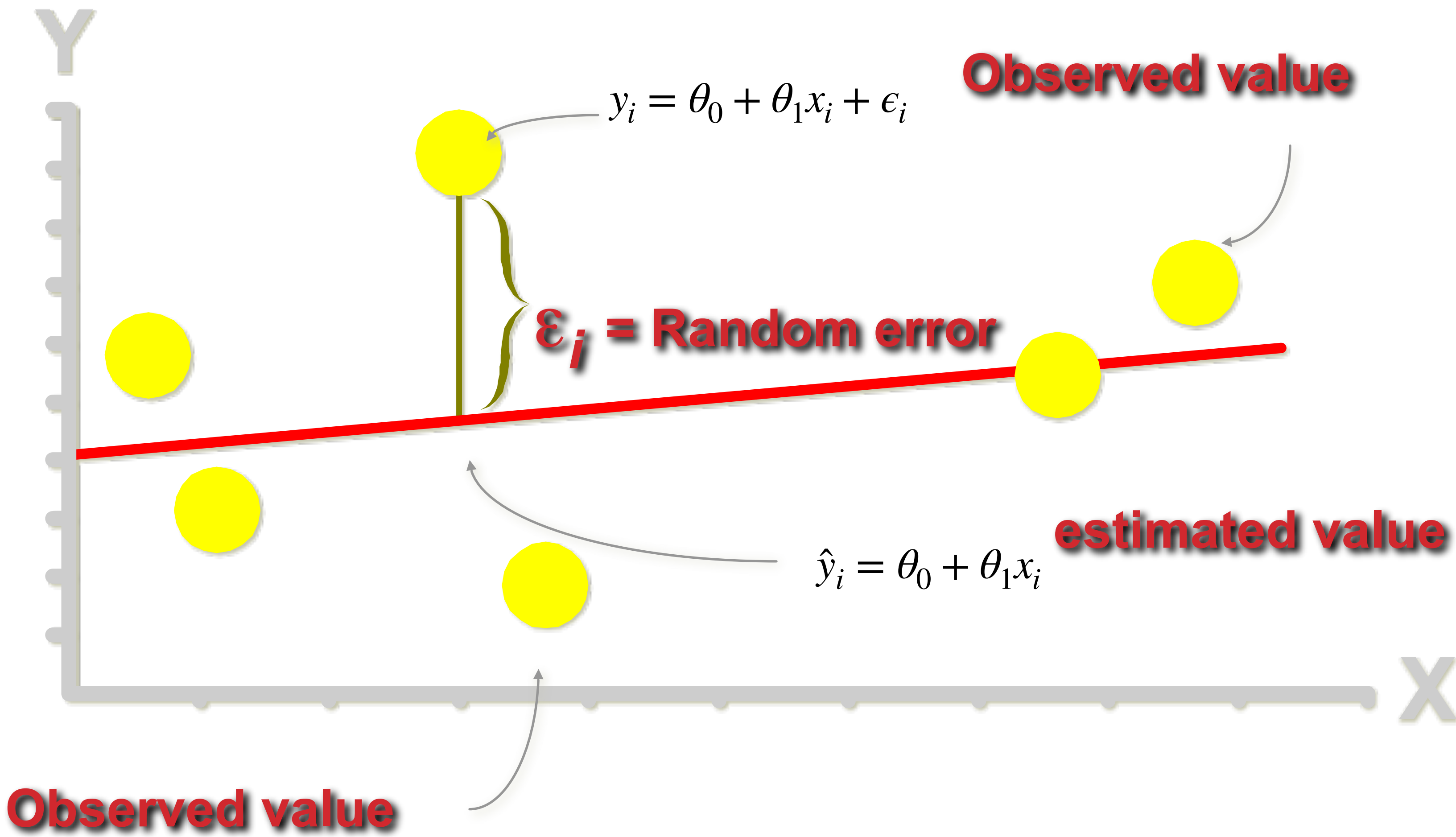
Example

- Single Variable Linear Regression

estimate $\hat{y}_i = \theta_0 + \theta_1 x_i$

x Area(sq. ft.)	y Price (in 1000$)
1600	220
1400	180
2100	350
…	…
….	….
2400	500

LINEAR REGRESSION



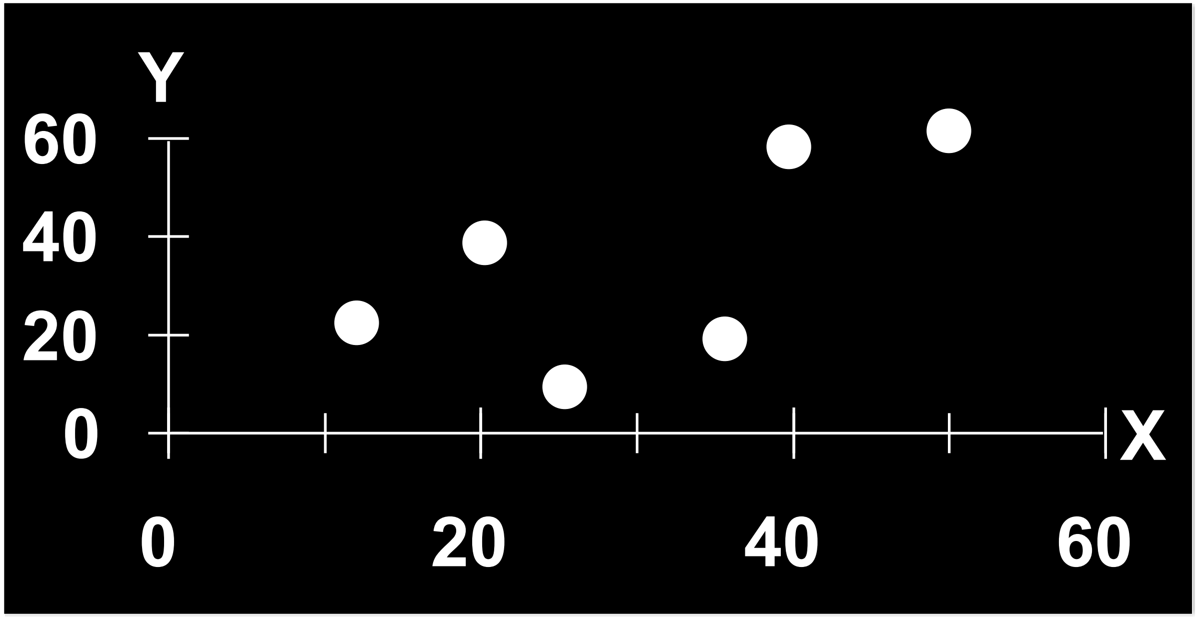
[WF]

?

ESTIMATING PARAMETERS: LEAST SQUARES METHOD

SCATTER PLOT

Plot all (X_i, Y_i) pairs, and plot your learned model



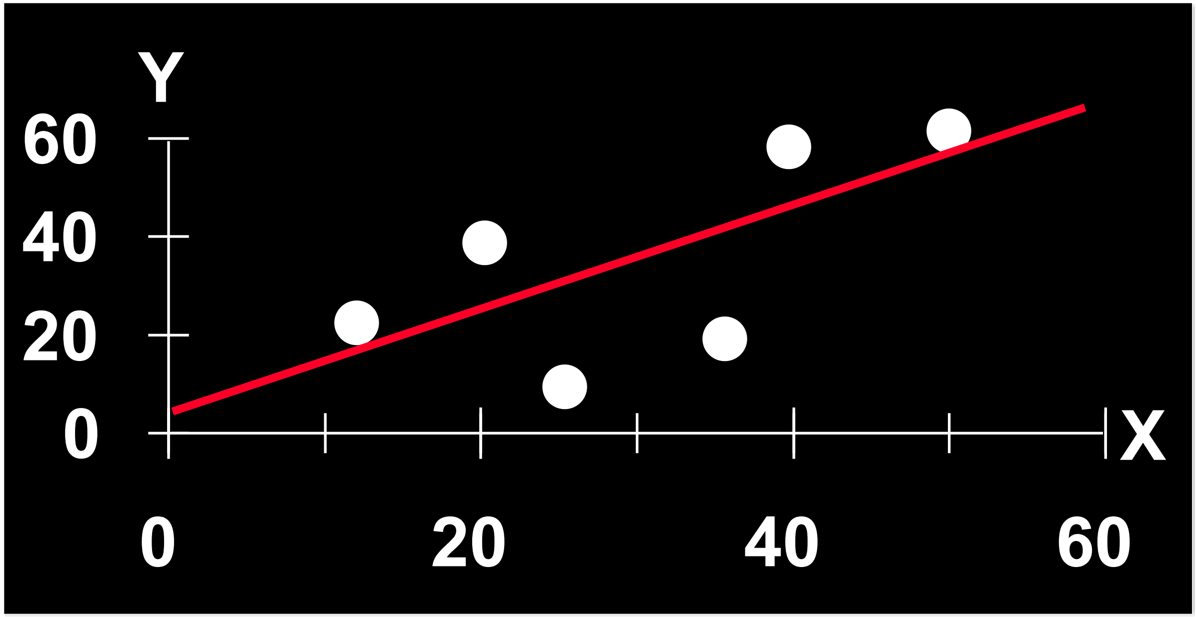
[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line "fits the best" …?

?????????

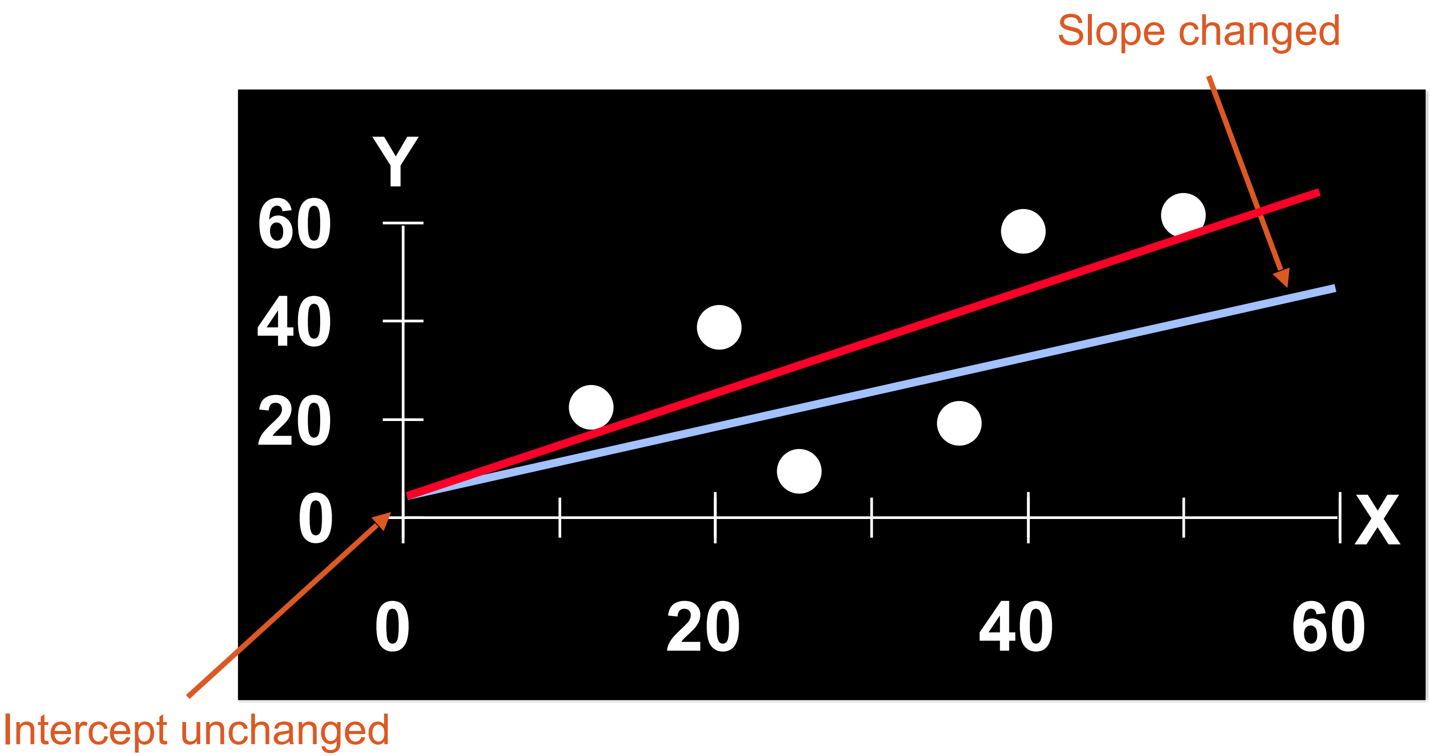


[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line "fits the best" ?????????

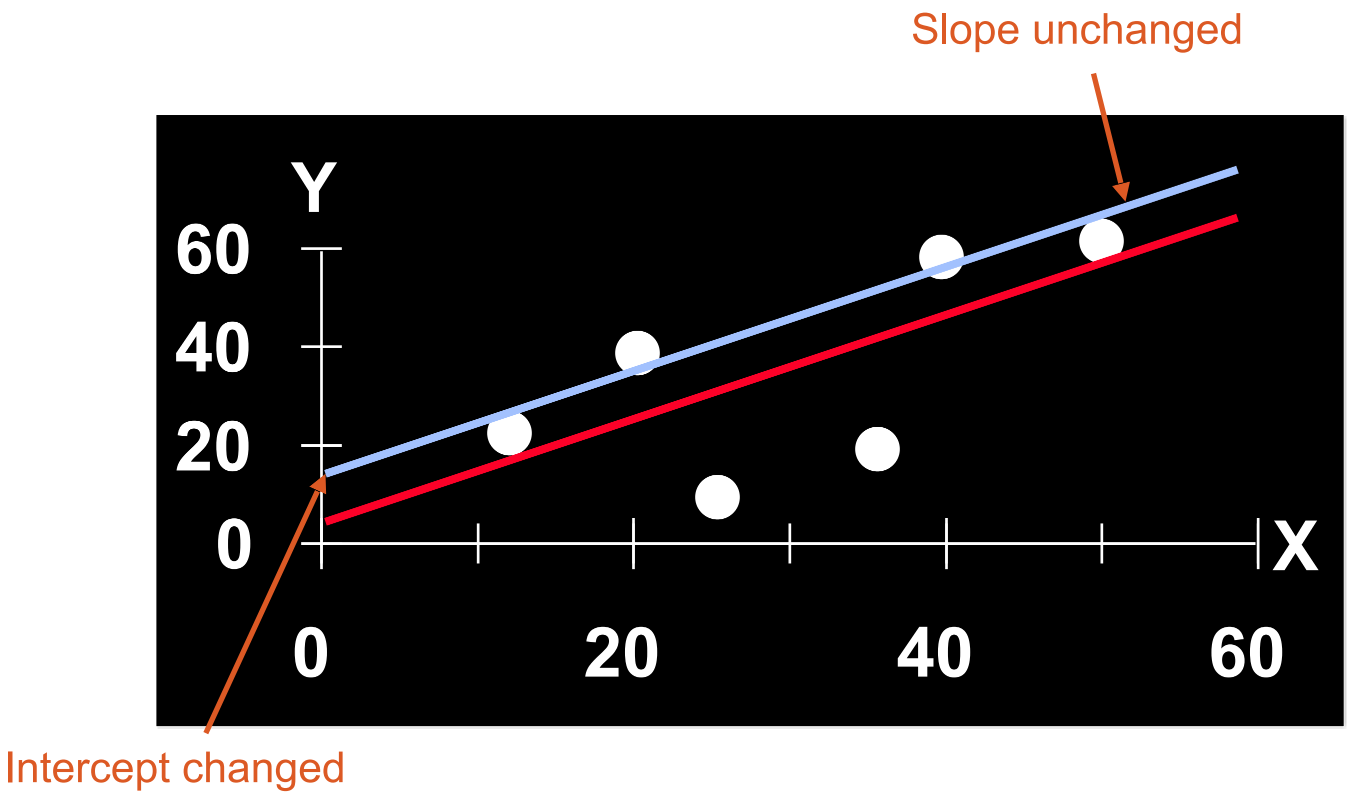


[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line “fits the best” ?????????



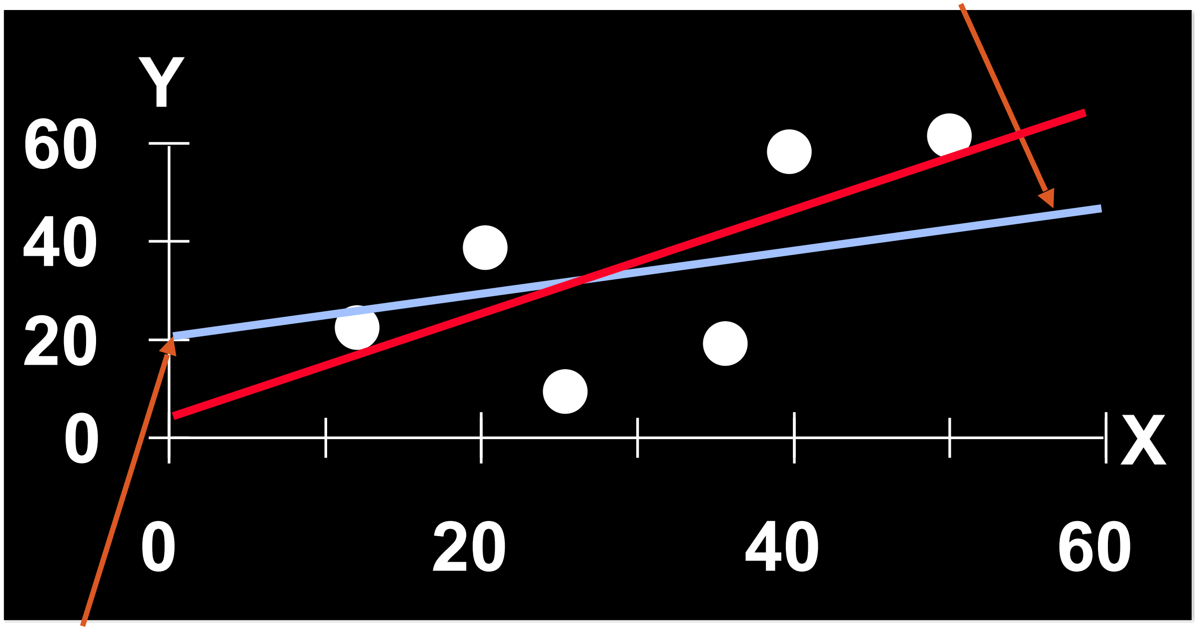
[WF]

QUESTION

How would you draw a line through the points?

How do you determine which line "fits the best" ?????????

Slope changed



Intercept changed

[WF]

LEAST SQUARES

Best fit: difference between the true (observed) Y-values and the estimated Y-values is minimized:

- Positive errors offset negative errors …
- … square the error!

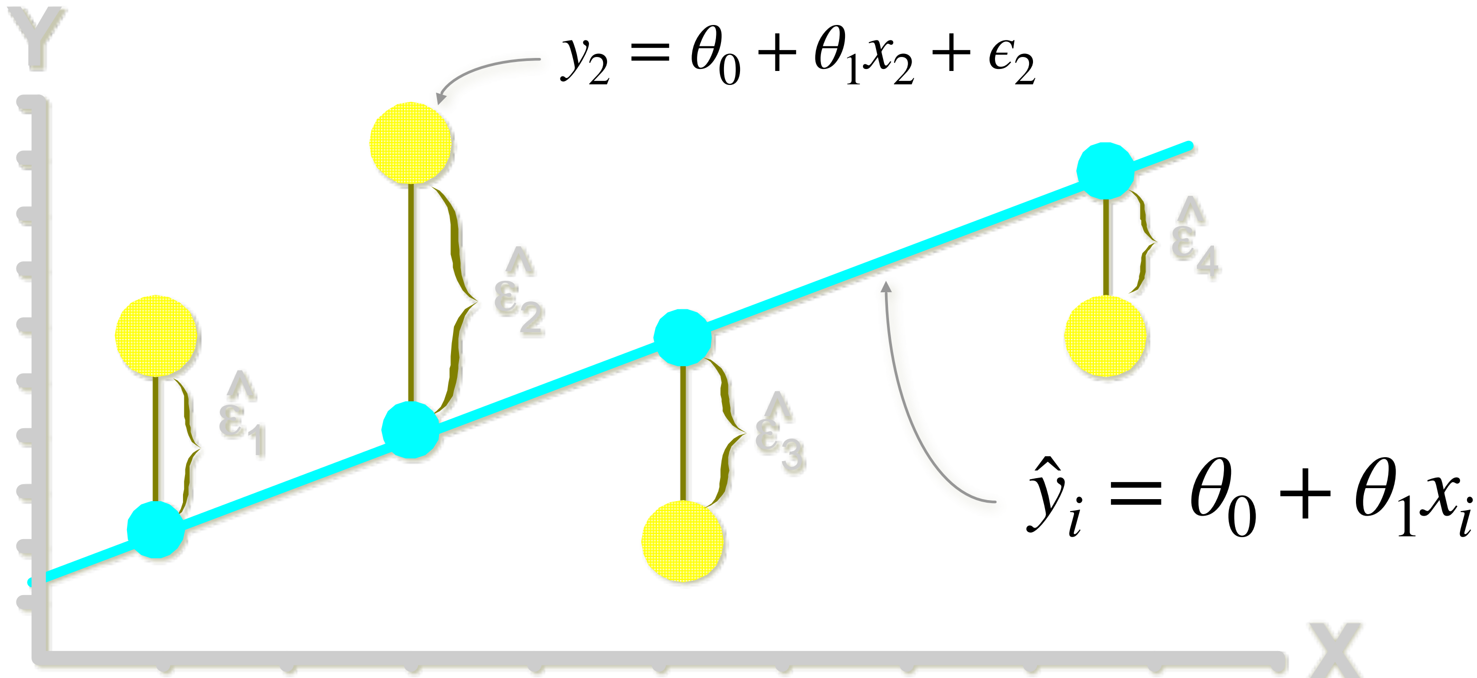
$$\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

Least squares minimizes the sum of the squared errors

[WF]

LEAST SQUARES, GRAPHICALLY

LS Minimizes $\sum_{i=1}^{n} \epsilon_i^2 = \epsilon_1^2 + \epsilon_2^2 + \ldots + \epsilon_n^2$



[WF]

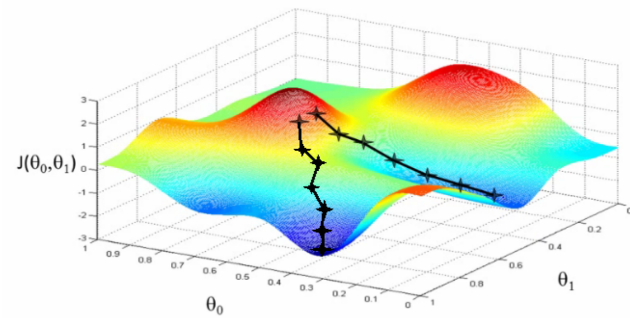
EXAMPLE

Single Variable Linear Regression

estimate $\hat{y}_i = \theta_0 + \theta_1 x_i$

x	y
Area(sq. ft.)	Price (in 1000$)
1600	220
1400	180
2100	350
…	…
….	….
2400	500

MULTIVARIATE REGRESSION

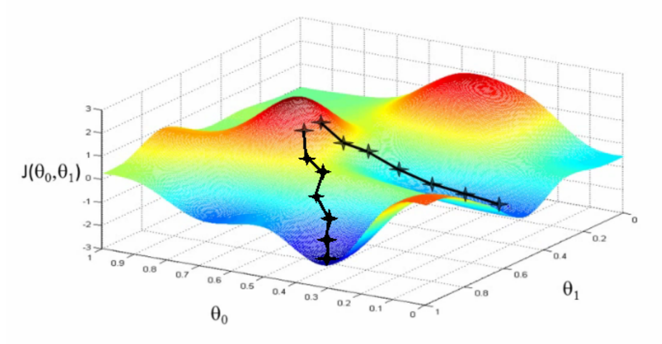


Multi Linear Regression

$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_m x_{im}$$

y Price (in 1000$)	x_1 Area(sq. ft.)	x_2 # Bathrooms	x_3 # Bedrooms
220	1600	2.5	3
180	1400	1.5	3
350	2100	3.5	4
…	…	…	…
….	….	…	…
500	2400	4	5

MULTIVARIATE REGRESSION



Multi Linear Regression

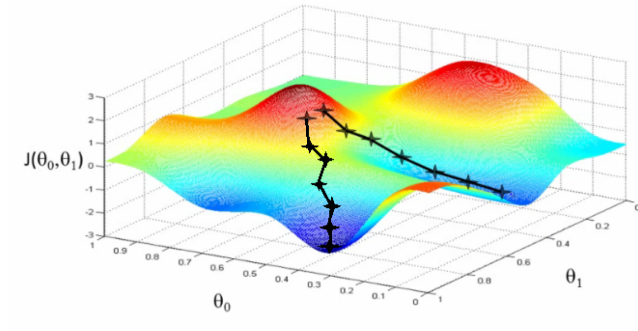
$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_m x_{im}$$

	Price (in 1000$)	Area(sq. ft.)	# Bathrooms	# Bedrooms
	220	1600	2.5	3
y_i	180	1400	1.5	3
	350	2100	3.5	4
	…	…	…	…
	….	….	…	…
	500	2400	4	5

x_i

1400	x_{i1}
1.5	x_{i2}
3	x_{i3}

MULTIVARIATE REGRESSION



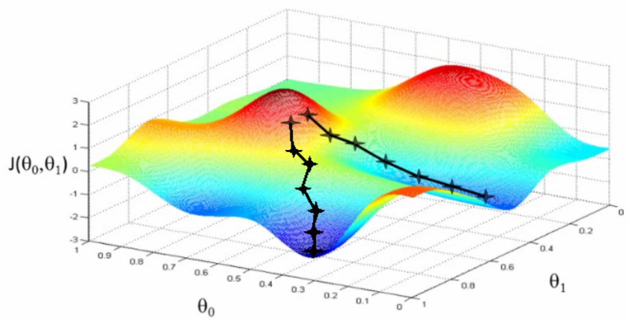
Multi Linear Regression

$$y_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_m x_{im}$$

y Price (in 1000$)	x_0	x_1 Area(sq. ft.)	x_2 # Bathrooms	x_3 # Bedrooms
220	1	1600	2.5	3
y_i $\boxed{180}$	$\boxed{1}$	1400	1.5	3
350	1	2100	3.5	4
…	…	…	…	…
….	….	….	…	…
500	1	2400	4	5

$$x_i = \begin{array}{cc} \boxed{1} & x_{i0} \\ \boxed{1400} & x_{i1} \\ \boxed{1.5} & x_{i2} \\ \boxed{3} & x_{i3} \end{array}$$

MULTIVARIATE REGRESSION MODEL



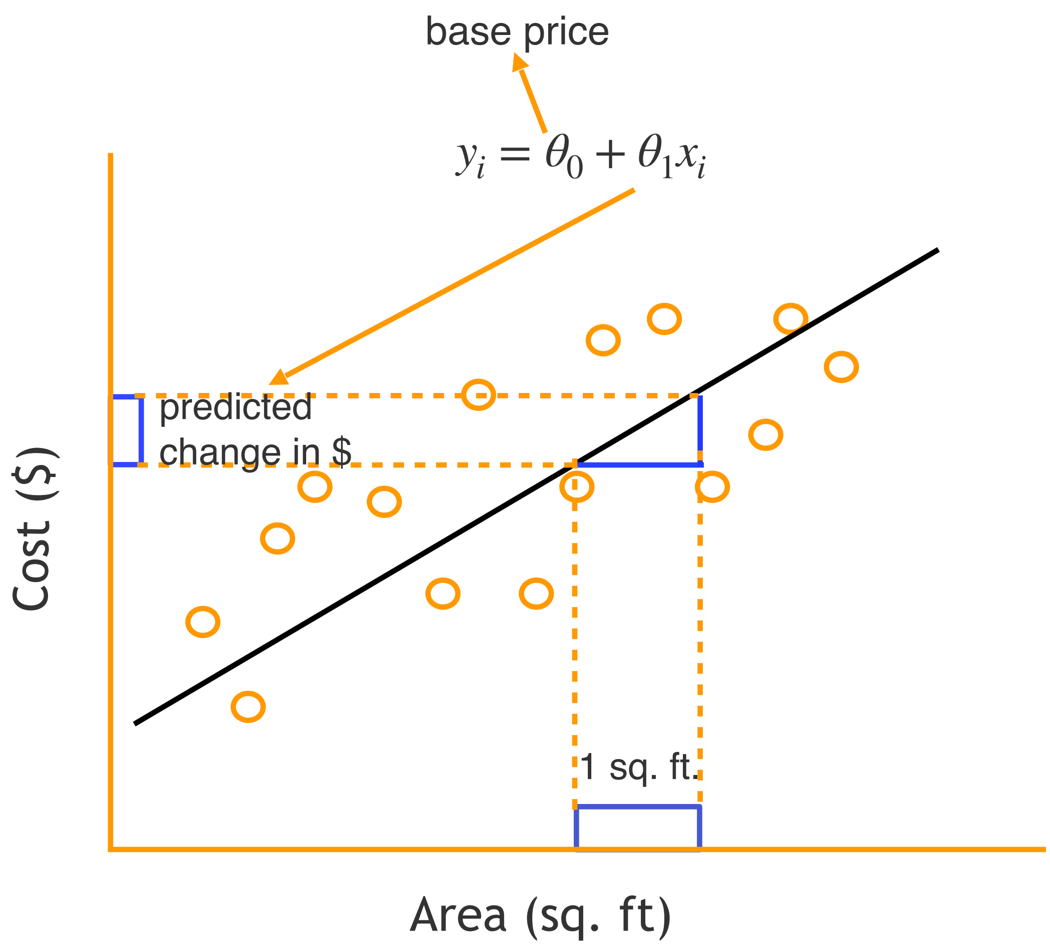
Model:

$$\hat{y}_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_m x_{im}$$

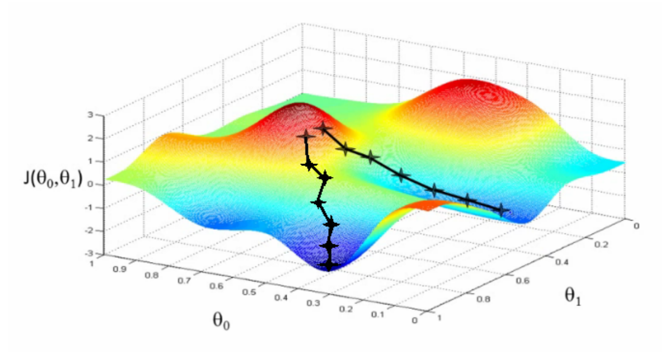
$$\hat{y}_i = \sum_{j=0}^{m} \theta_{ij} x_{ij}$$

feature 1 = x_0 …. (constant, 1)
feature 2 = x_1 …. (area, sq. ft.)
feature 3 = x_2 …. (# of bedrooms)
feature 4 = x_3 …. (# of bathrooms)
….
….
feature m = x_m

SINGLE VARIABLE LINEAR REGRESSION

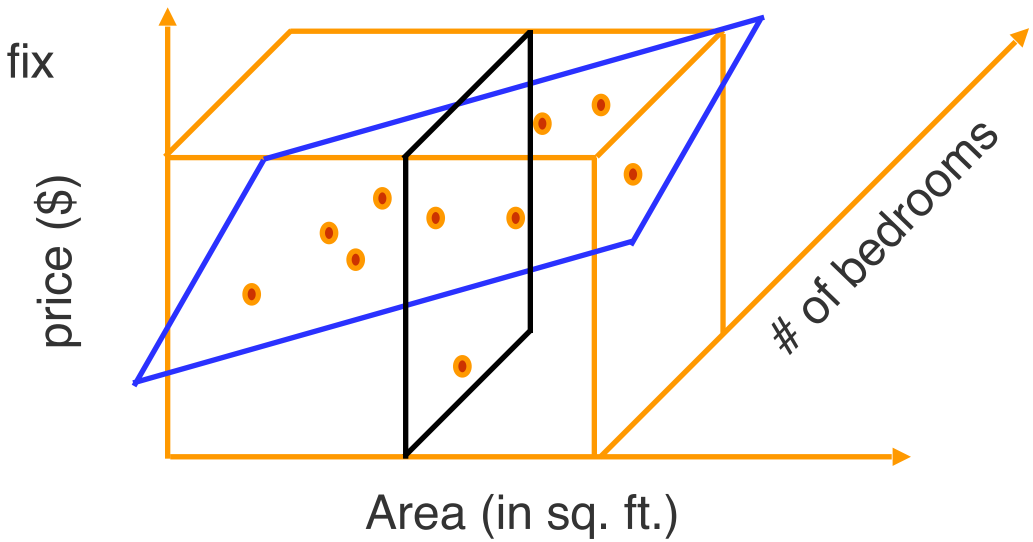


INTERPRETING COEFFICIENTS



$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2$$

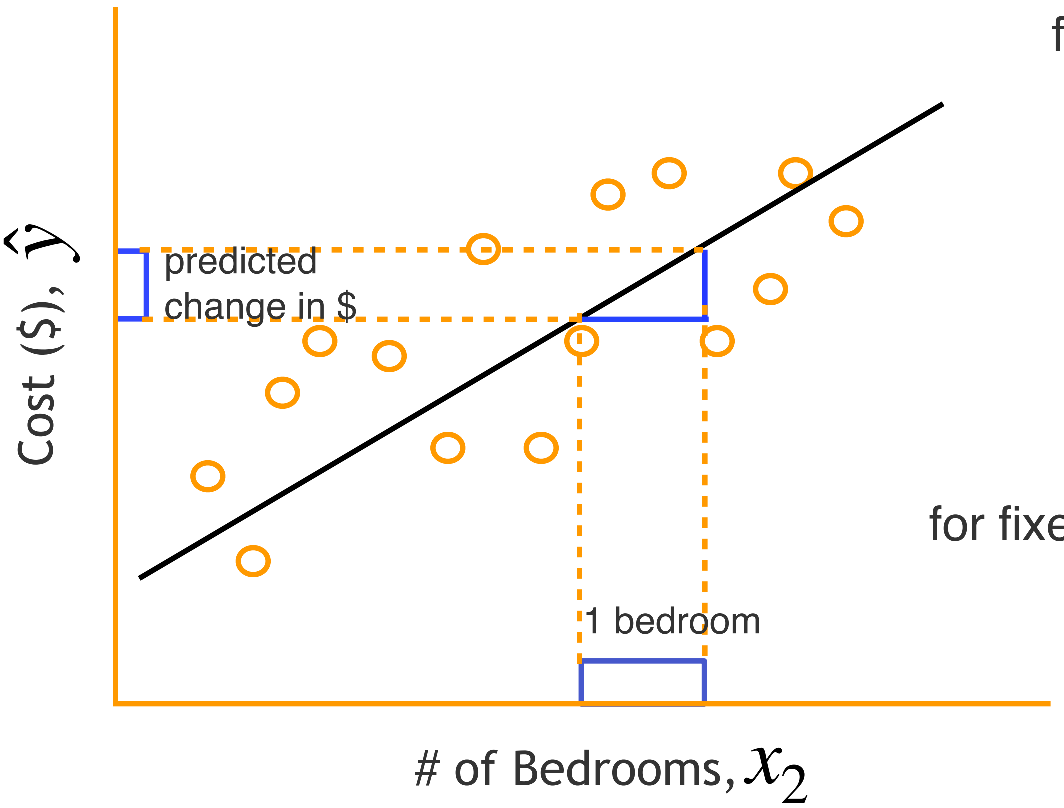
Two Linear Features



SINGLE VARIABLE LINEAR REGRESSION

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2$$

fix



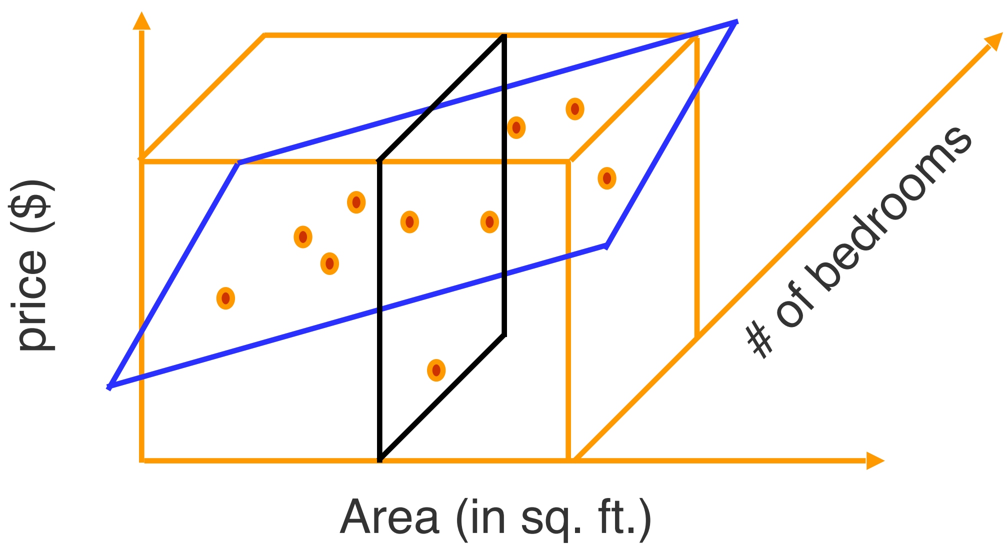
for fixed area (sq. ft.), x_1

INTERPRETING COEFFICIENTS

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2 + \ldots + \hat{\theta}_j x_j + \ldots + \hat{\theta}_m x_m$$

fix fix fix fix

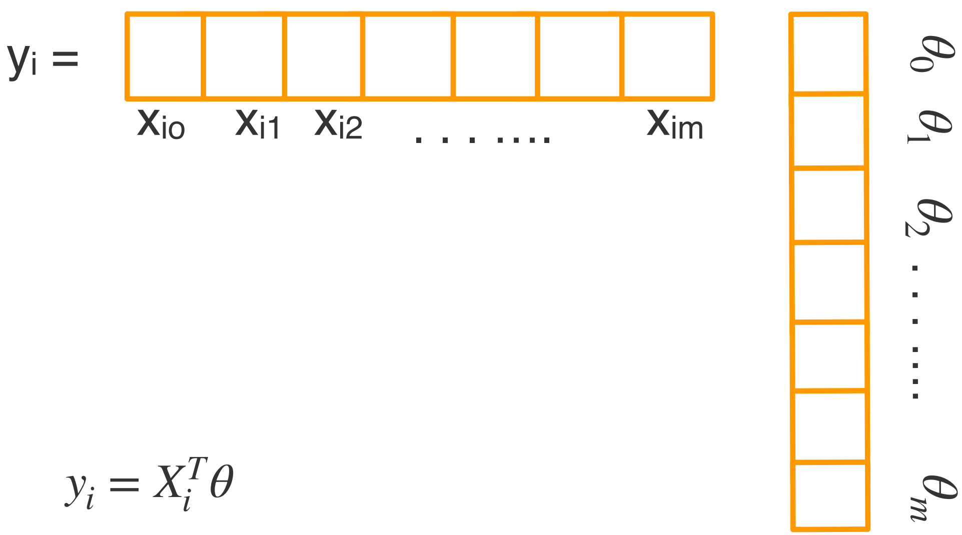
Multiple Features



ONE OBSERVATION MODEL

Matrix Notation
For observation i

$$\hat{y}_i = \sum_{j=0}^{m} \theta_{ij} x_{ij}$$



$$y_i = X_i^T \theta$$

ALL OBSERVATION MODEL

Matrix Notation For all observations

$$
\begin{bmatrix}
x_{10} & x_{11} & x_{12} & .. & .. & x_{1m} \\
x_{20} & x_{21} & x_{22} & .. & .. & x_{2m} \\
x_{30} & x_{31} & x_{32} & .. & .. & x_{3m} \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
x_{n0} & x_{n1} & x_{n2} & .. & .. & x_{nm}
\end{bmatrix}
\begin{bmatrix}
\theta_0 \\ \theta_1 \\ \theta_2 \\ . \\ \theta_m
\end{bmatrix}
=
\begin{bmatrix}
y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n
\end{bmatrix}
$$

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

LEAST SQUARES OPTIMIZATION

Each row is a feature vector paired with a label for a single input

Rewrite inputs:

n labeled inputs

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \cdots \\ (x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times m}, y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$$

m features

Rewrite optimization problem:

$$\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$$

*Recall $||z||_2^2 = z^T z = \sum_i z_i^2$

LEAST SQUARES OPTIMIZATION

Rewrite inputs:

Each row is a feature vector paired with a label for a single input

n labeled inputs

$$X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \cdots \\ (x^{(n)})^T \end{bmatrix} \in \mathbb{R}^{n \times m}, y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$$

m features

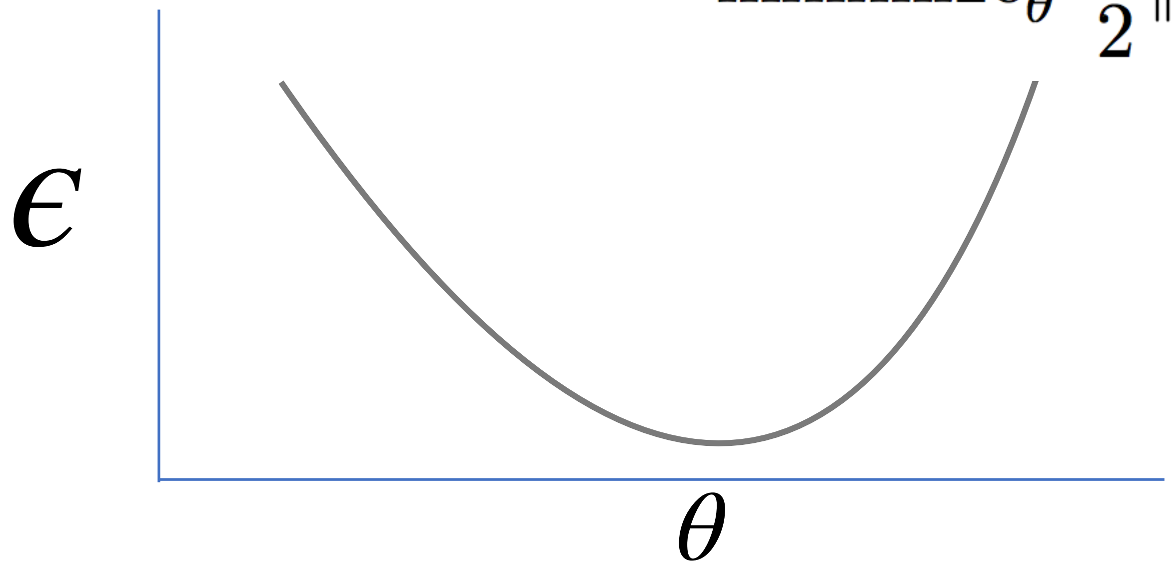
Rewrite optimization problem: $\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$

$$\implies \text{minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

*Recall $||z||_2^2 = z^T z = \sum z_i^2$

ERROR FUNCTION

$$\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$$



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

GRADIENTS

Minimizing a multivariate function involves finding a point where the gradient is zero:

$$\nabla_\theta f(\theta) = 0 \text{ (the vector of zeros)}$$

Points where the gradient is zero are local minima

- If the function is convex, also a global minimum

Let's solve the least squares problem!

We'll use the multivariate generalizations of some concepts from MATH141/142 …

- Chain rule: $\nabla_\theta f(X\theta) = X^T \nabla_{X\theta} f(X\theta)$
- Gradient of squared ℓ² norm: $\nabla_\theta ||\theta - z||_2 = 2(\theta - z)$

LEAST SQUARES

Recall the least squares optimization problem:

$$\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$$

What is the gradient of the optimization objective ????????

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 =$$

Chain rule:

$$\nabla_\theta f(X\theta) = X^T \nabla_{X\theta} f(X\theta)$$

$$X^T \nabla_{X\theta} \frac{1}{2}\|X\theta - y\|_2^2 =$$

Gradient of norm:

$$\nabla_\theta \|\theta - z\|_2^2 = 2(\theta - z)$$

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 = X^T(X\theta - y)$$

LEAST SQUARES

Recall: points where the gradient equals zero are minima.

$$\nabla_\theta \frac{1}{2}\|X\theta - y\|_2^2 = X^T(X\theta - y)$$

So where do we go from here??????????

$$X^T(X\theta - y) = 0$$

Solve for model parameters θ

$$X^T X\theta - X^T y = 0 \Rightarrow X^T X\theta = X^T y$$

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

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