MSML 605 - Lecture 5

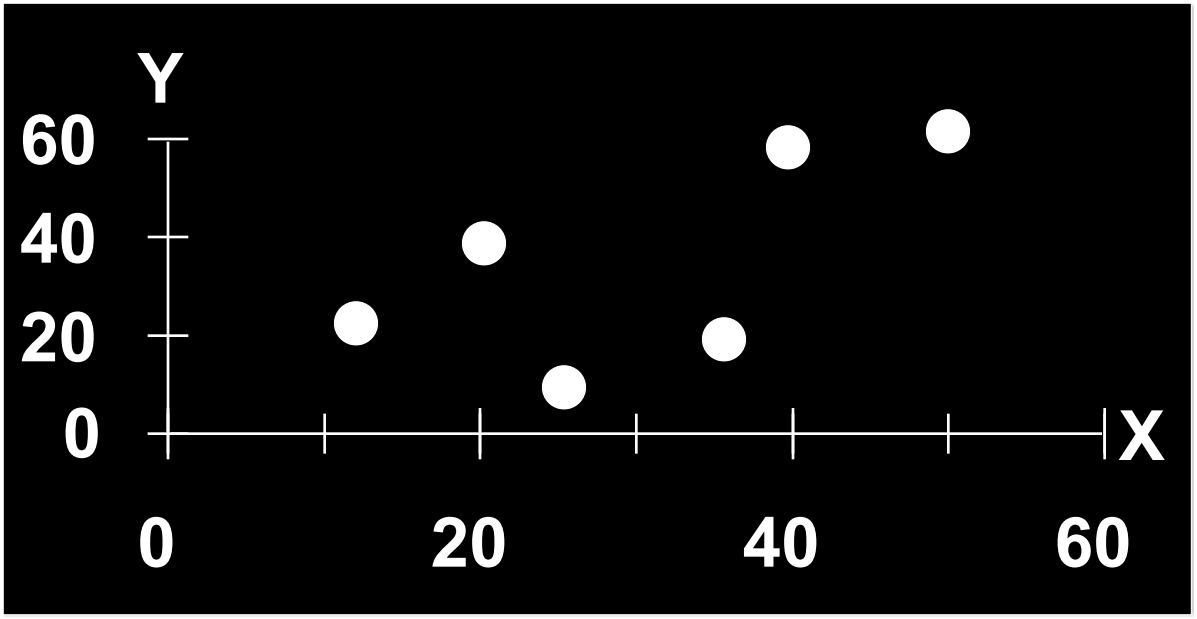
Least Squares Optimization

Example

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

SCATTER PLOT

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



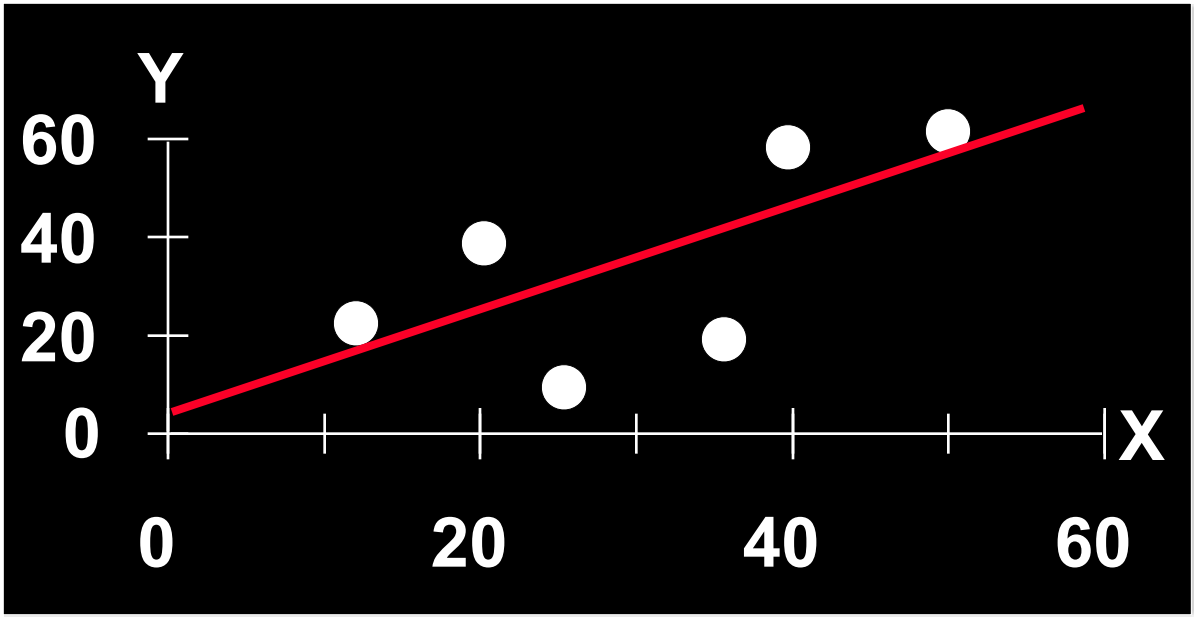
[WF]

QUESTION

How would you draw a line through the points?

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

?????????

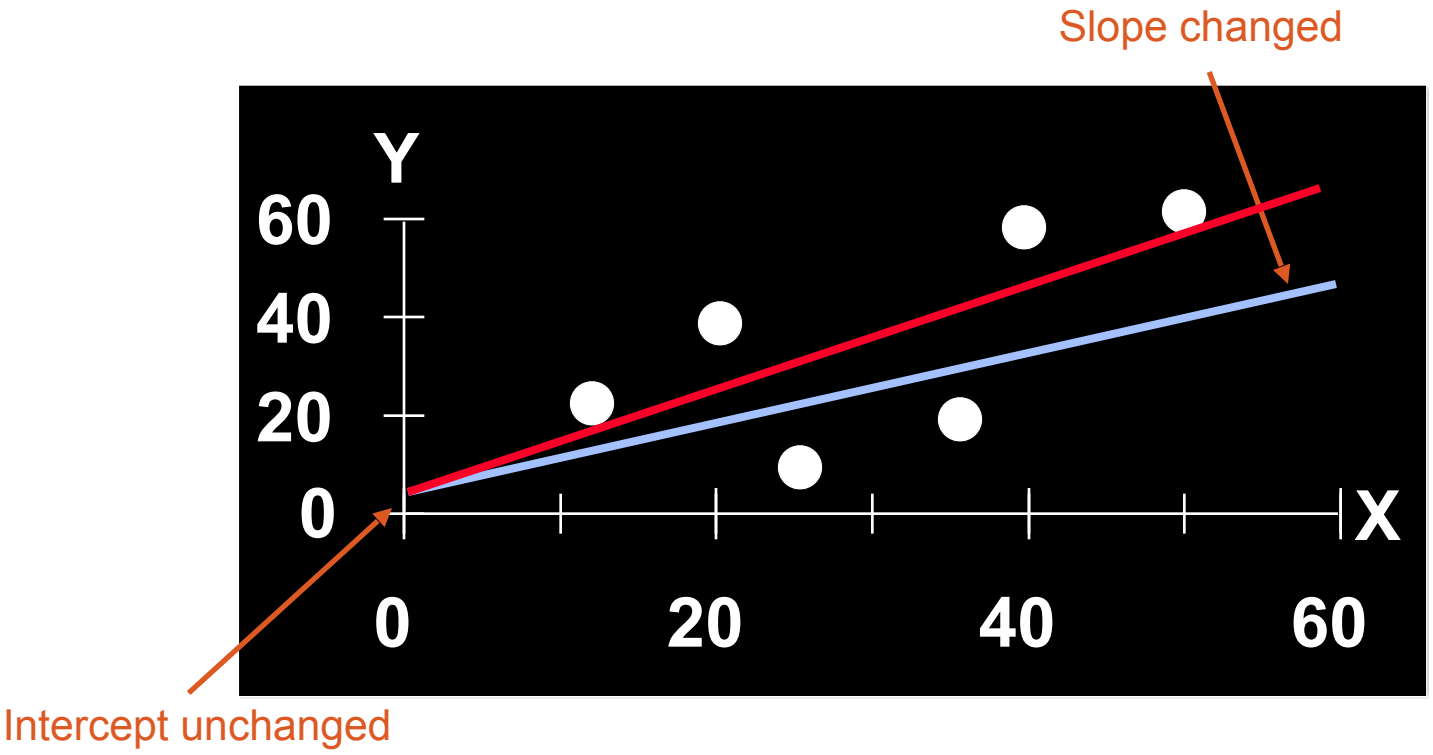


[WF]

QUESTION

How would you draw a line through the points?

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

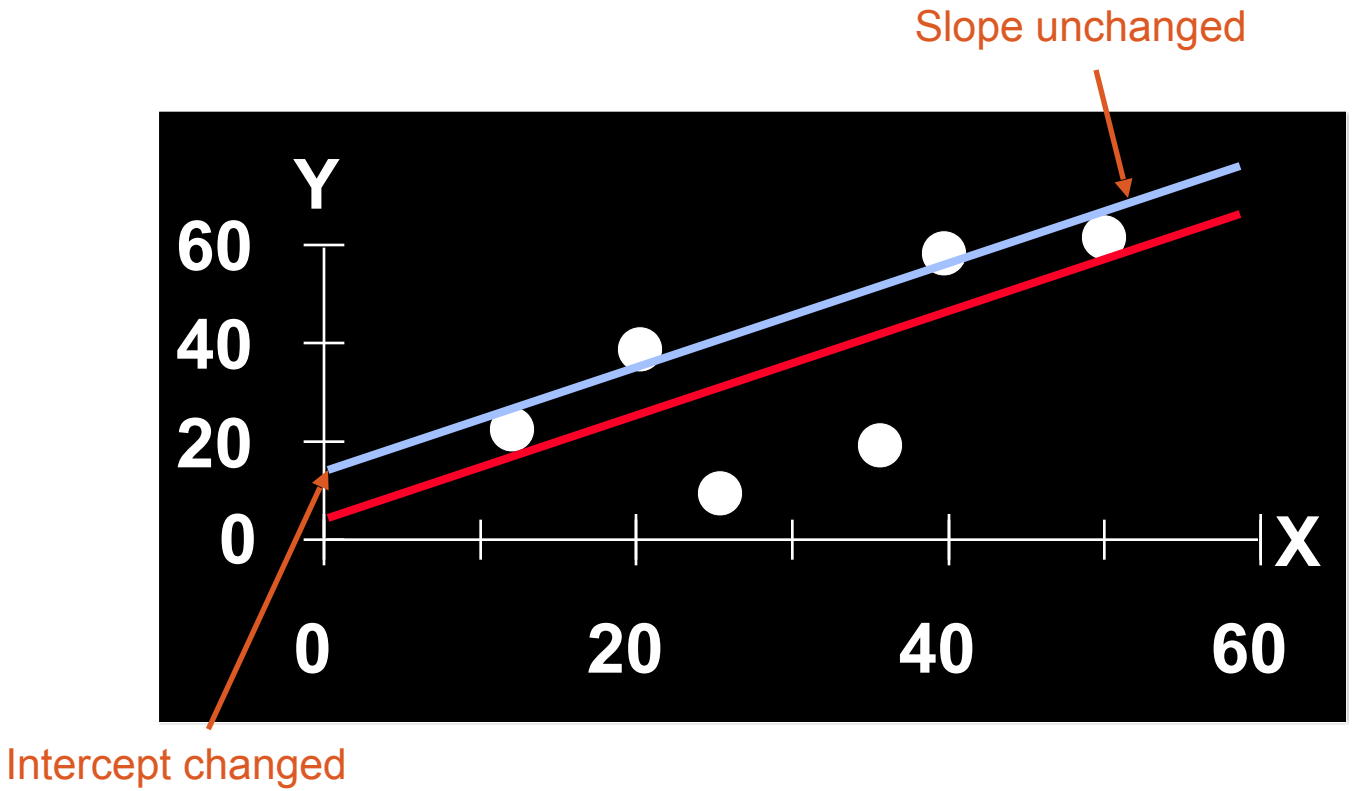


[WF]

QUESTION

How would you draw a line through the points?

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

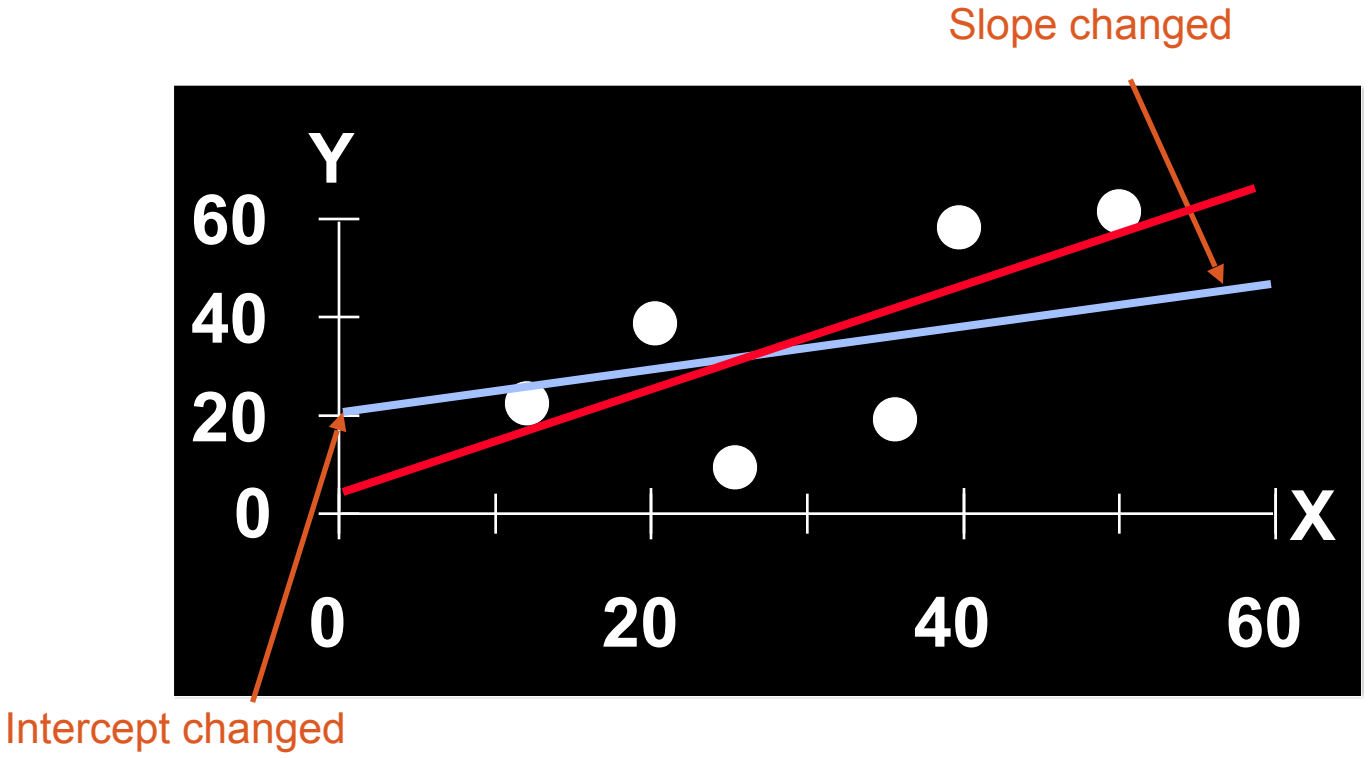


[WF]

QUESTION

How would you draw a line through the points?

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



[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!

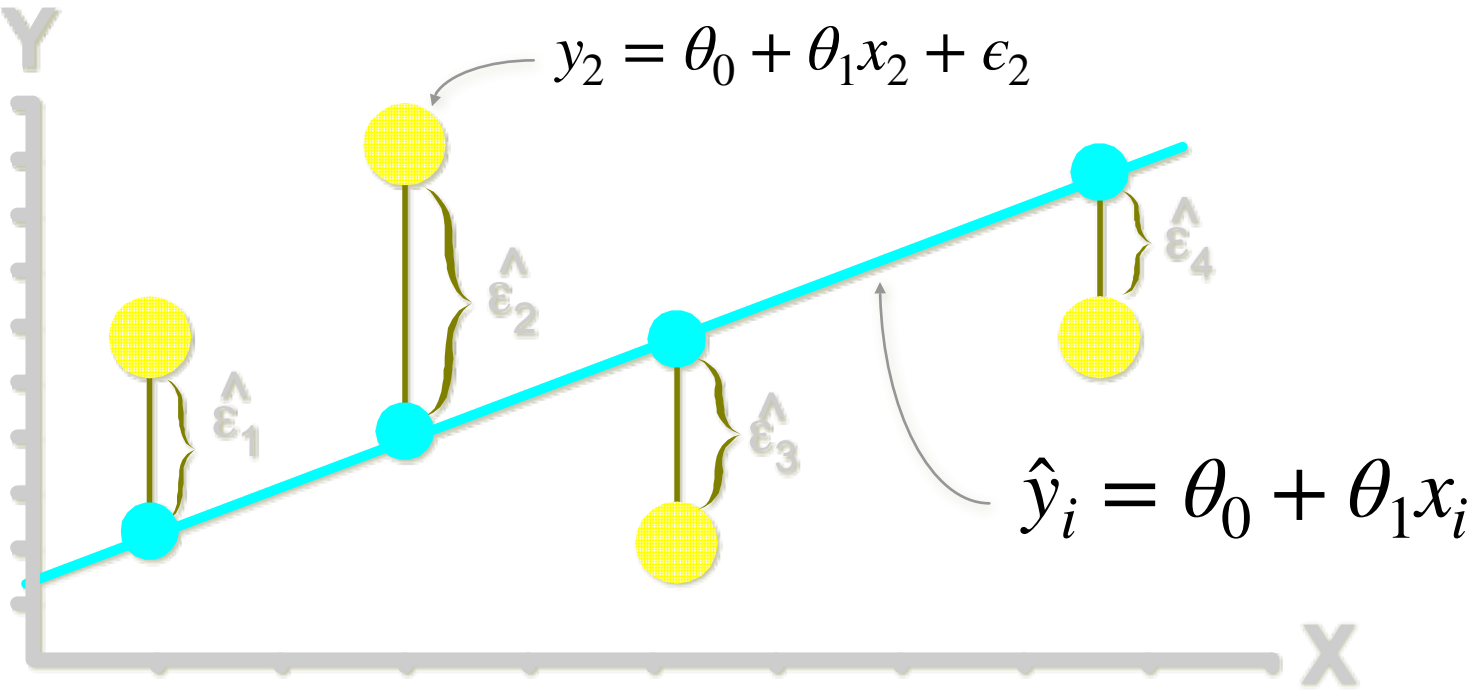
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \epsilon_i^2$$

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]

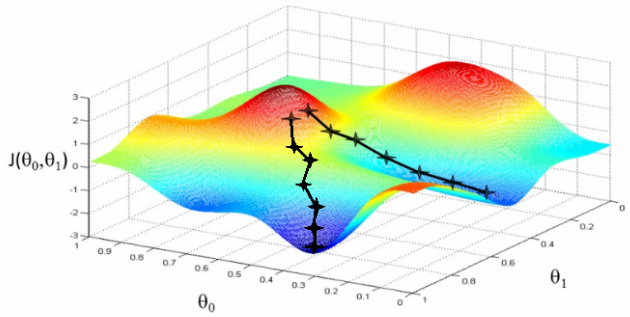
Example

- 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

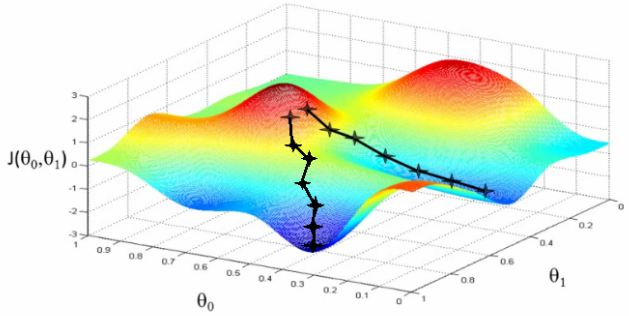


- Multi Linear Regression

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

y	x_1	x_2	x_3
Price (in 1000$)	Area(sq. ft.)	# Bathrooms	# 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

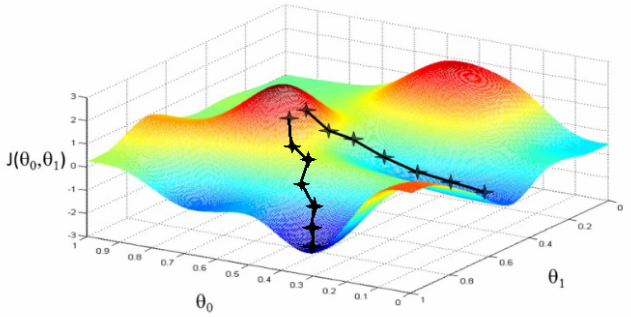
$$\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_n x_{in}$$

	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

$$y_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_n x_{in}$$

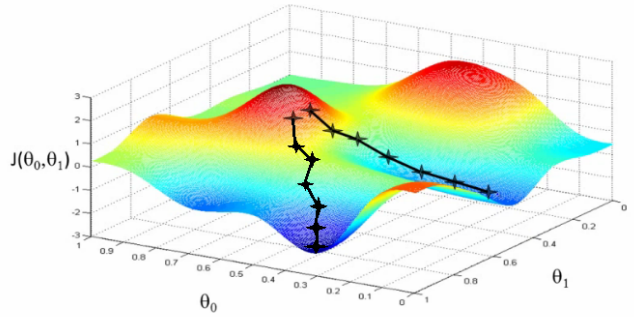
	y	x_0	x_1	x_2	x_3
	Price (in 1000$)		Area(sq. ft.)	# Bathrooms	# Bedrooms
	220	1	1600	2.5	3
y_i	180	1	1400	1.5	3
	350	1	2100	3.5	4
	…	…	…	…	…
	….	….	….	…	…
	500	1	2400	4	5

x_i

1	x_{i0}
1400	x_{i1}
1.5	x_{i2}
3	x_{i3}

Multivariate Regression Model

- Model:

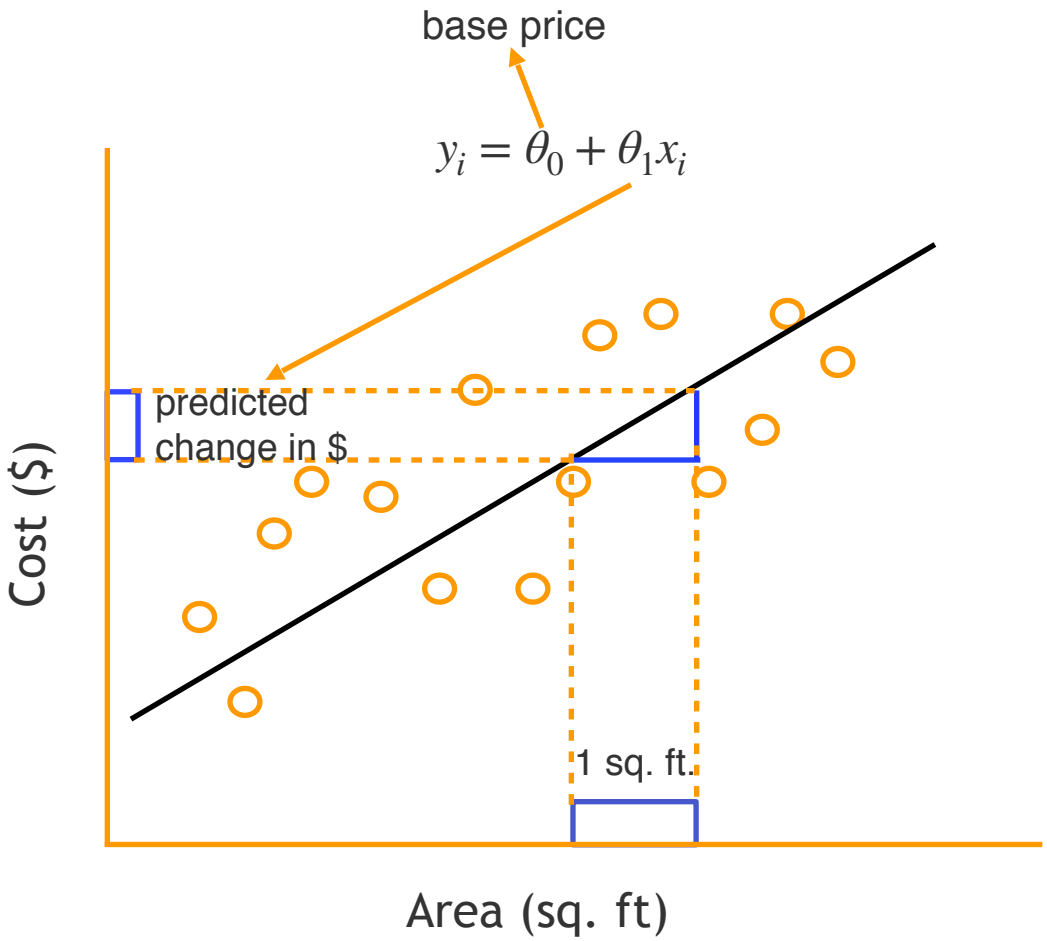


$$\hat{y}_i = \theta_0 x_{i0} + \theta_1 x_{i1} + \theta_2 x_{i2} + \ldots + \theta_n x_{in}$$

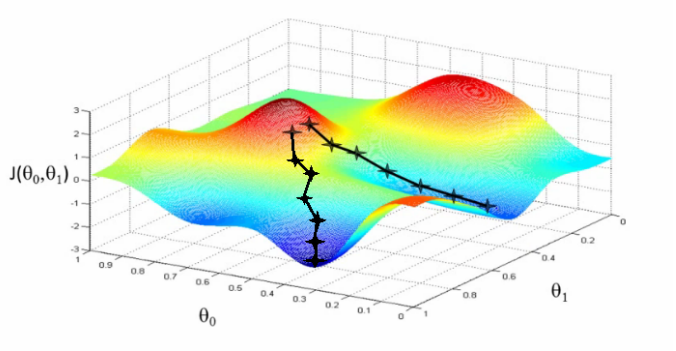
$$\hat{y}_i = \sum_{j=0}^{n} \theta_{ij} x_{ij}$$

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 n = x_n

Single Variable Linear Regression

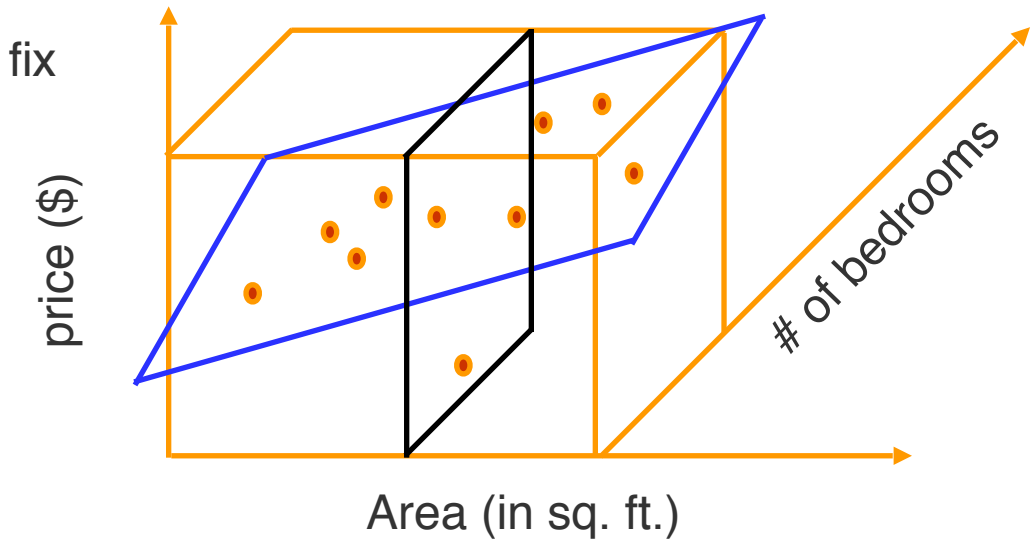


Interpreting Coefficients



- Two Linear Features

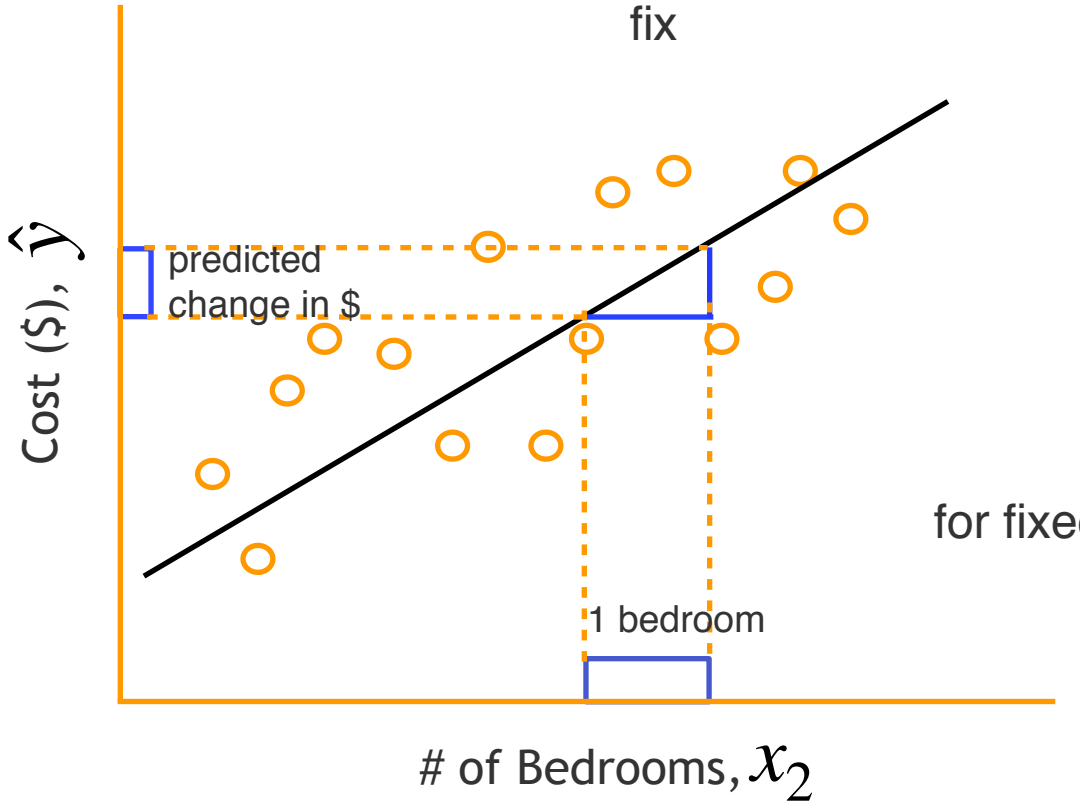
$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x_1 + \hat{\theta}_2 x_2$$



Single Variable Linear Regression

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

fix



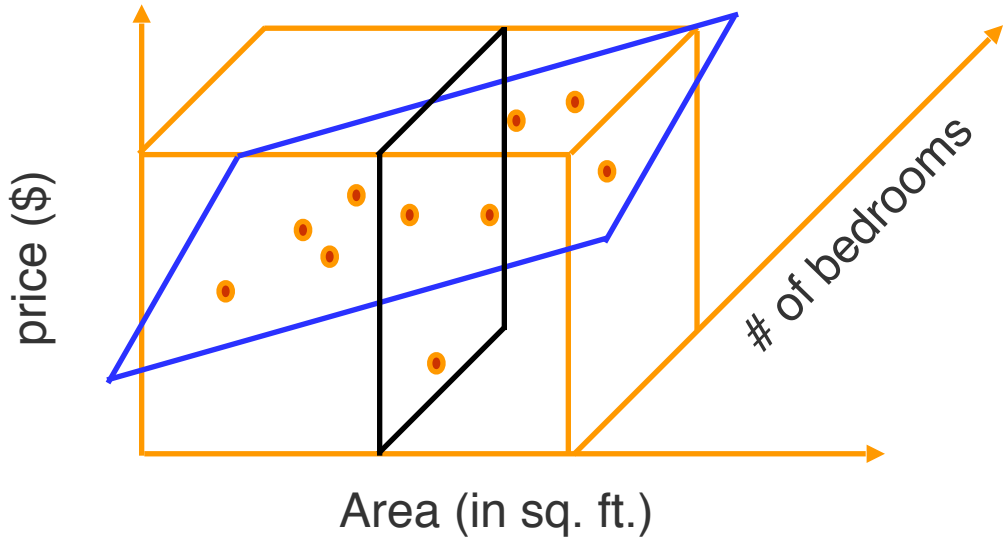
for fixed area (sq. ft.), x_1

Interpreting Coefficients

- Multiple Features

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

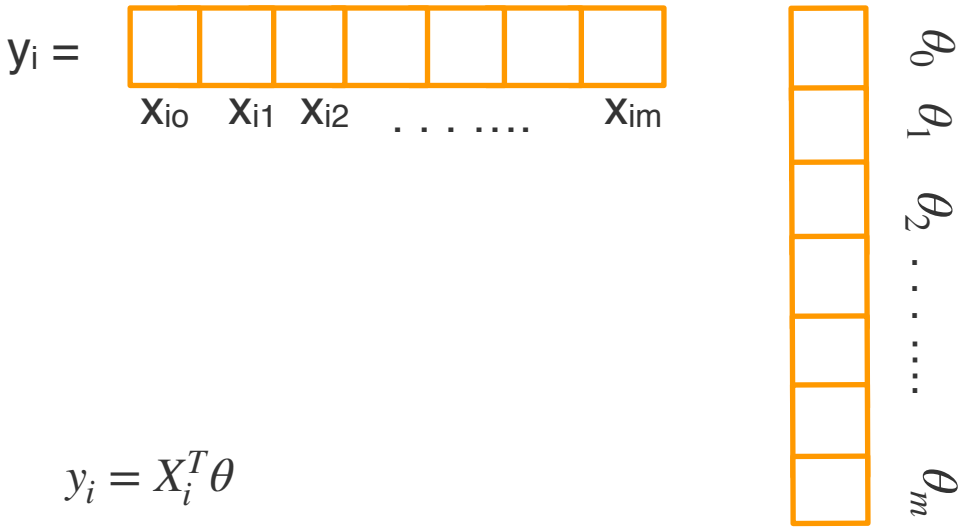
fix fix fix fix



One Observation Model

- Matrix Notation
  For observation i

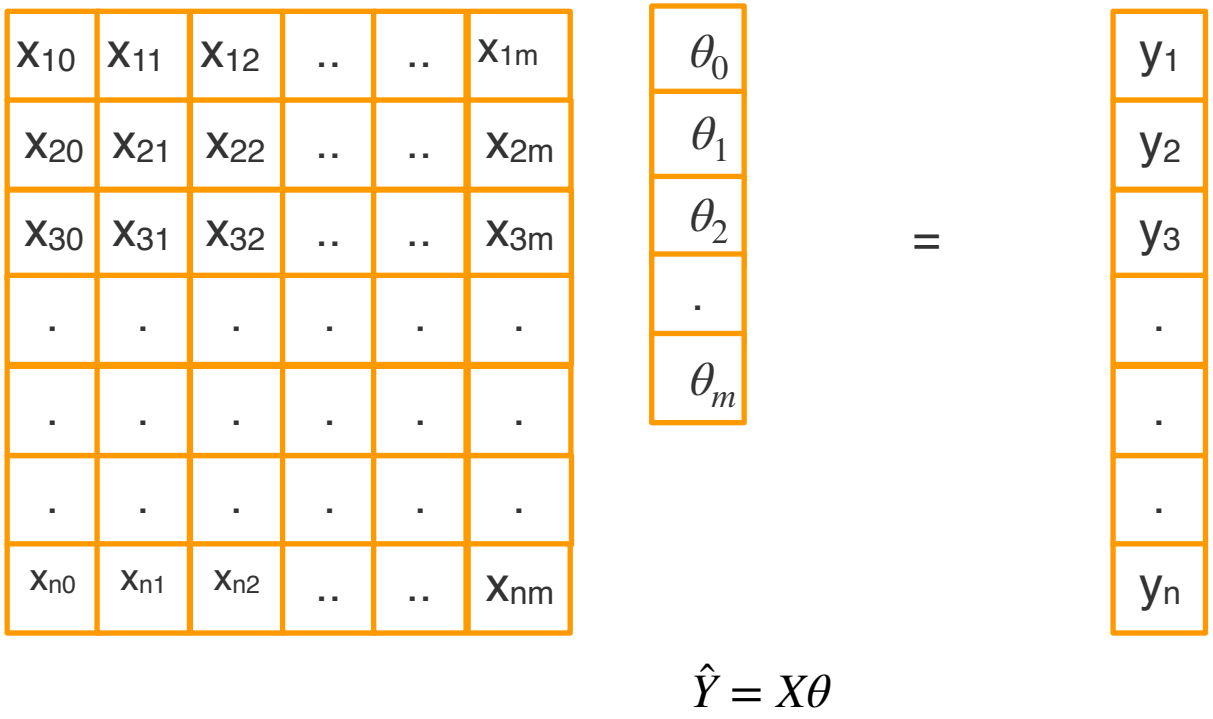
$$\hat{y}_i = \sum_{j=0}^{m} \theta_{ij} x_{ij}$$



$$y_i = X_i^T \theta$$

All Observation Model

- Matrix Notation
  For all observations



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

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:

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

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

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:

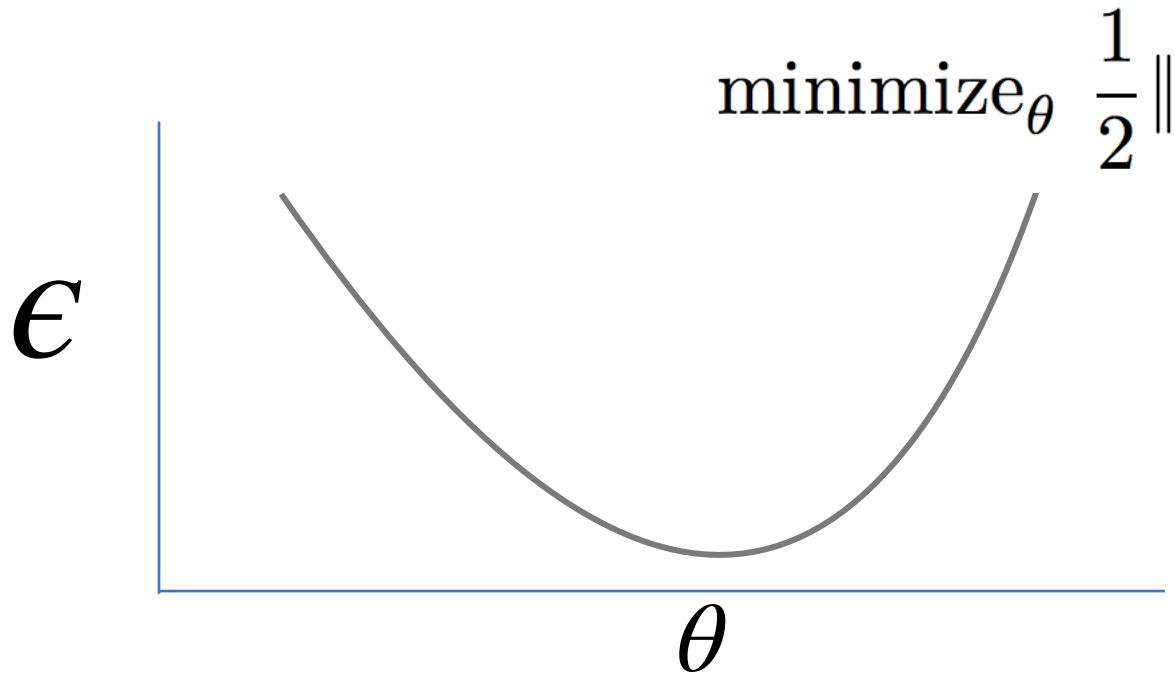
$$\text{minimize}_\theta \ \frac{1}{2}\|X\theta - y\|_2^2$$

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

*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!

- Chain rule:

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

- Gradient of :

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

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

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

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

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