DS 4400

Machine Learning and Data Mining I

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

September 22 2020

Announcements

- HW 1

  PIAZZA: PDF) LATEX ; GRADESCOPE - PDF

  CODE - GOOGE FORM

  ZIP FILE

  - Is due on Monday, Sept. 28
- Python tutorials
  - Panda data frames tutorial by Alex Wang
    - Wed, Sept. 23, 5-6pm
  - Same Zoom link as office hours
  - Recording of first tutorial is available on Canvas under "Lecture Recording"

CLASS TIMES: TUE. 11:45 AM - 1:25 PM

THU: 2:50 PM - 4:30 PM

Outline

MODULE 2

- GRADIENT DESCENT
- REGULARIZATION

- Linear regression
- Simple linear regression
  - MSE as loss function
  - Derivation of optimal solution
  - Correlation coefficient, covariance, and connection to regression
  - Example of linear regression fit
  - Lab in Python
- Multiple linear regression

Linear regression

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → Learning model

Data	Pre-processing	Feature extraction	Learning model
Labeled	Normalization	Feature Selection	Regression
$x_i, y_i \in R$			$f(x)$

Testing

New data → Learning model → Predictions $y' = f(x') \in R$

New data	Learning model	Predictions
Unlabeled		Response variable
x'	$f(x)$	Regression

Steps to Learning Process

- Define problem space
- Collect data
- Extract feature

NUMERICAL

- Pick a model (hypothesis)
- Develop a learning algorithm
  - Train and learn model parameters

OPTIMIZATION

FIT MODEL TO DATA

- Make predictions on new data
  - Testing phase
- In practice, usually re-train when new data is available and use feedback from deployment

Linear regression

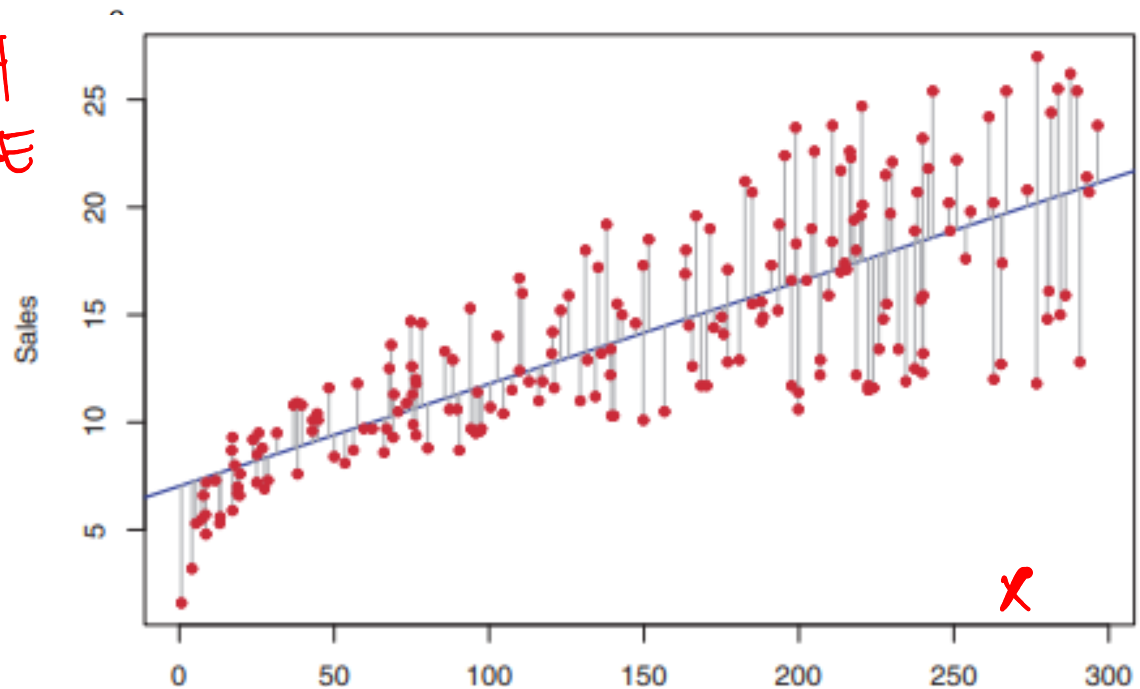
- One of the most widely used techniques
- Fundamental to many complex models
  - Generalized Linear Models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to solve in closed form
- Efficient practical algorithm (gradient descent)

Linear regression

Given:

- Data $X = \{x_1, \ldots x_N\}$, where $x_i \in R^d$
- Corresponding labels $Y = \{y_1, \ldots y_N\}$, where $y_i \in R$

TRAINING DATA



Y RESPONSE

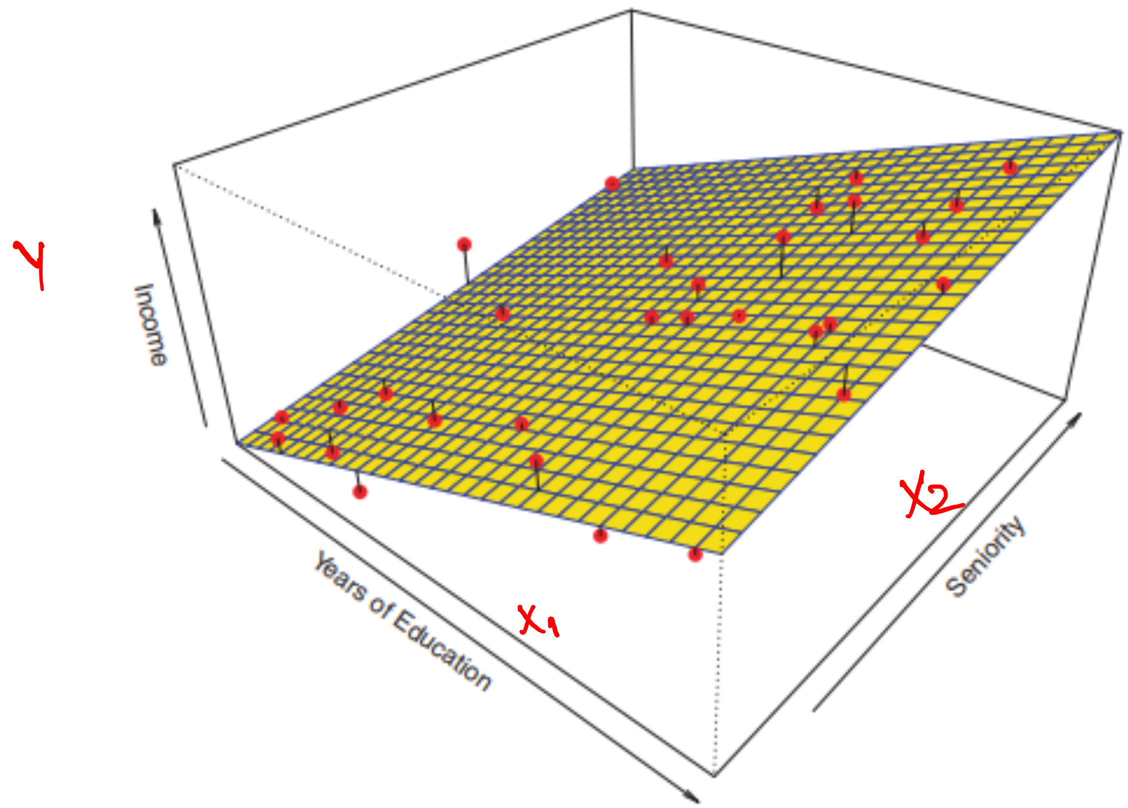
X PREDICTOR

SIMPLE LINEAR REGR

1 PREDICTOR

ORDINARY LEAST SQUARES (OLS)

Income Prediction



MULTIPLE LINEAR REGRESSION

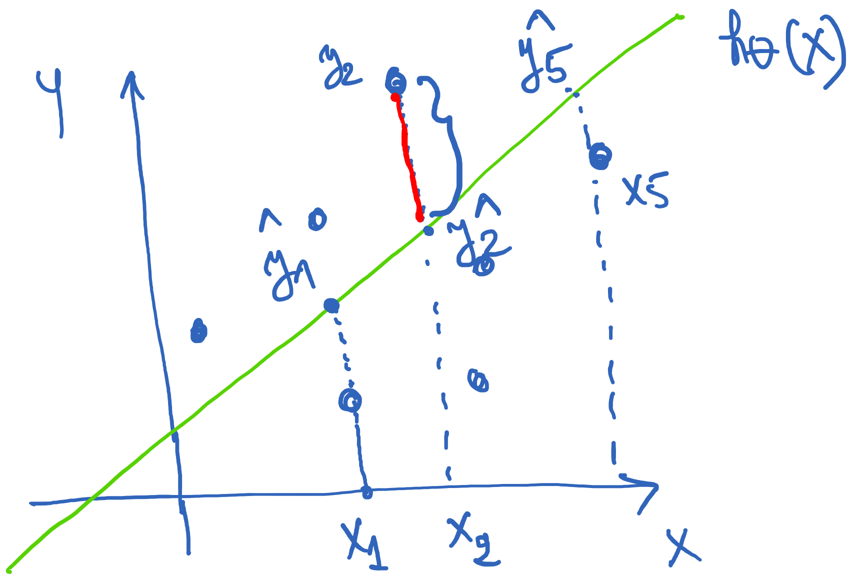
Hypothesis: Linear Model

Hypothesis $\boxed{h_\theta(x) = \theta_0 + \theta_1 x}$

(x_i, y_i)

TRAINING DATA

Simple linear regression: line with 2 parameters: θ_0, θ_1



$\boxed{\hat{y}_1} = h_\theta(x_1)$ PREDICTED VALUE

$\boxed{y_1}$ = TRUE VALUE

$\hat{y}_1 - y_1$: ERRORS

$\hat{y}_2 - y_2$ RESIDUALS

Least-Squares Linear Regression

- $\boxed{\text{Cost Function}}$

LOSS FUNCTION

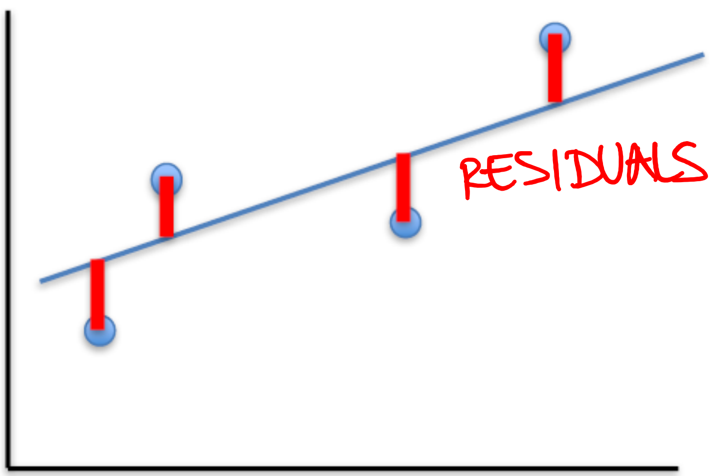
TRAINING DATA POINTS

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\underbrace{[h_\theta(x_i) - y_i]}_{\text{RESIDUAL}}^2$$

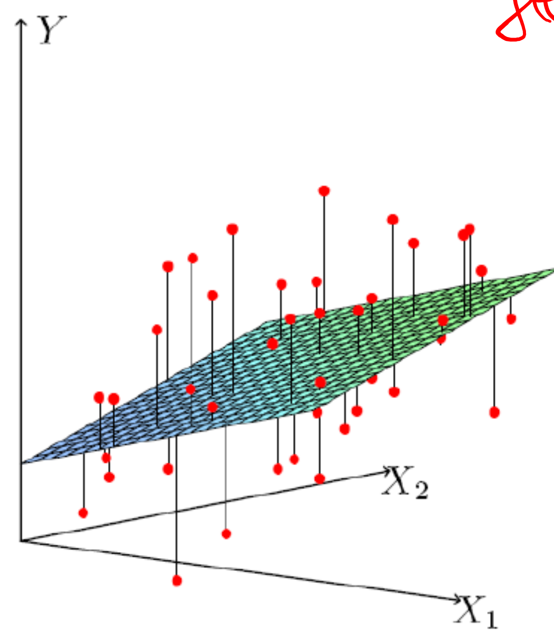
Mean Square Error (MSE)

- Fit by solving $\boxed{\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta})}$

$J(\theta)$: CONVEX



SIMPLE LR

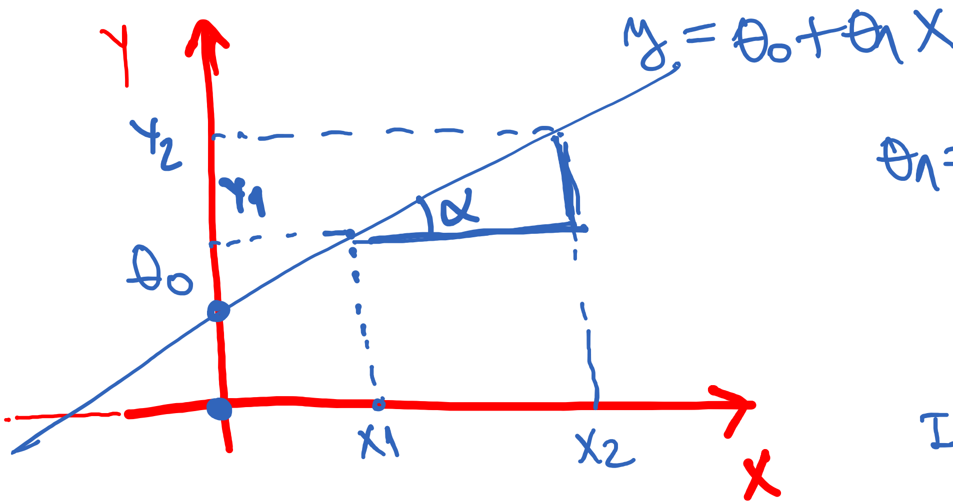


MULTIPLE LR

Terminology and Metrics

- Residuals
  - Difference between predicted values and actual values
  - Predicted value for example i is: $\widehat{y_i} = h_\theta(x_i)$
  - $R_i = \left| y_i - \widehat{y_i} \right| = |y_i - (\theta_0 + \theta_1 x_i)|$
- Residual Sum of Squares (RSS)
  - $RSS = \sum {R_i}^2 = \sum [y_i - (\theta_0 + \theta_1 x_i)]^2$
- Mean Square Error (MSE)
  - $MSE = \frac{1}{N} \sum R_i^2 = \frac{1}{N} \sum [y_i - (\theta_0 + \theta_1 x_i)]^2$

Interpretation



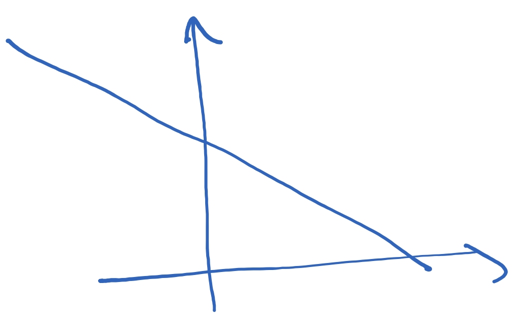
$$\theta_1 = \frac{y_2 - y_1}{x_2 - x_1} \quad \text{SLOPE}$$

$$\tan(\alpha) = \theta_1$$

If $\theta_1 = 1 \Rightarrow \alpha = 45^\circ$

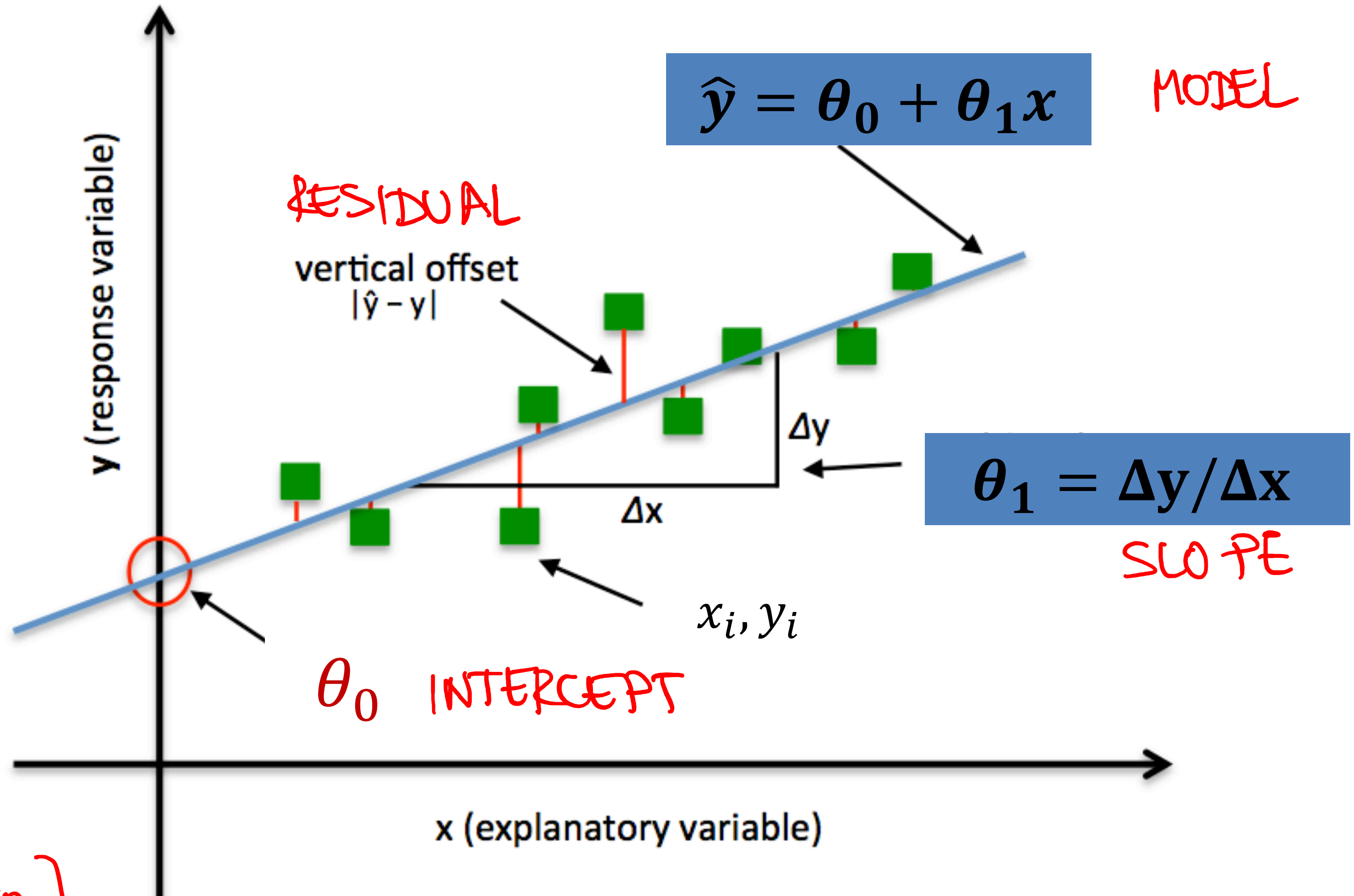
$\theta_1 = -1 \Rightarrow \alpha = 135^\circ$

$x = 0 \Rightarrow y = \theta_0$ INTERCEPT



$\theta_1 = -1$

Interpretation



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

$$h_\theta(x) = \theta_0 + \theta_1 x$$ MODEL

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2$$ ERROR METRIC

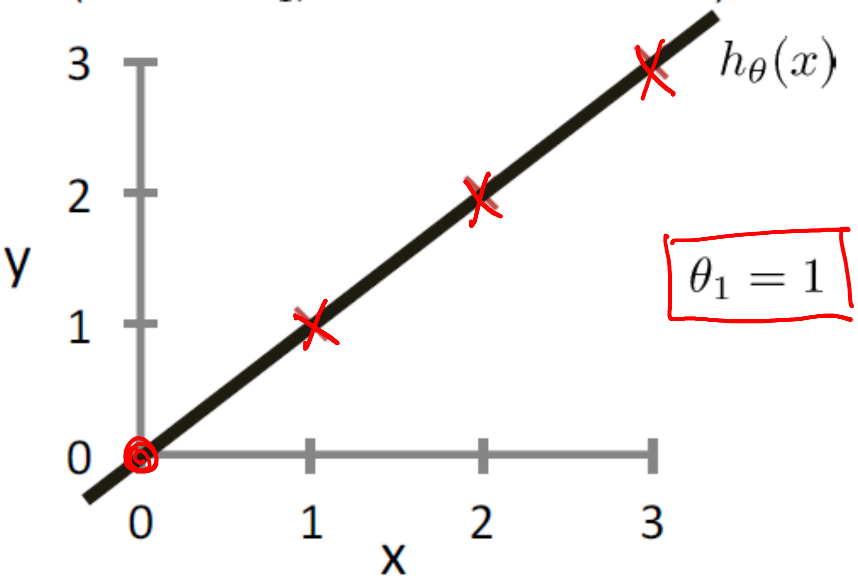
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

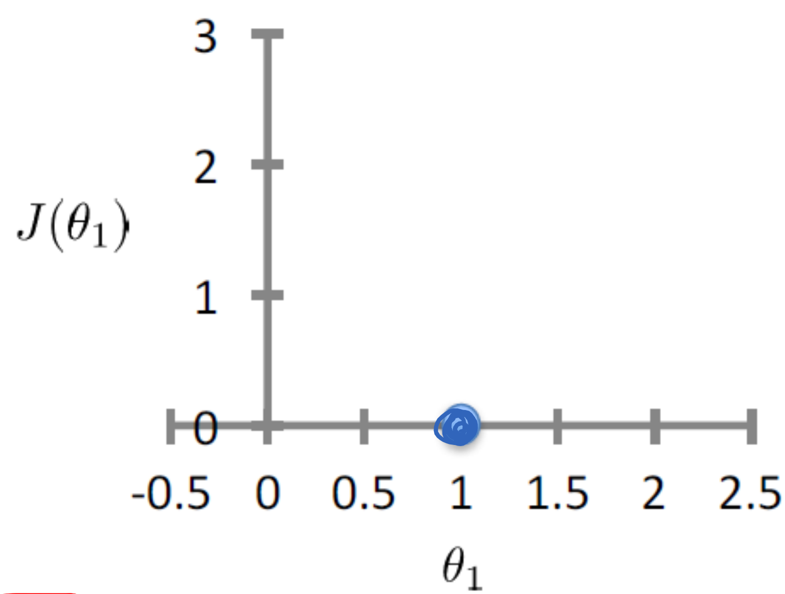
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



$\boxed{\text{Fix } \theta_0 = 0}$

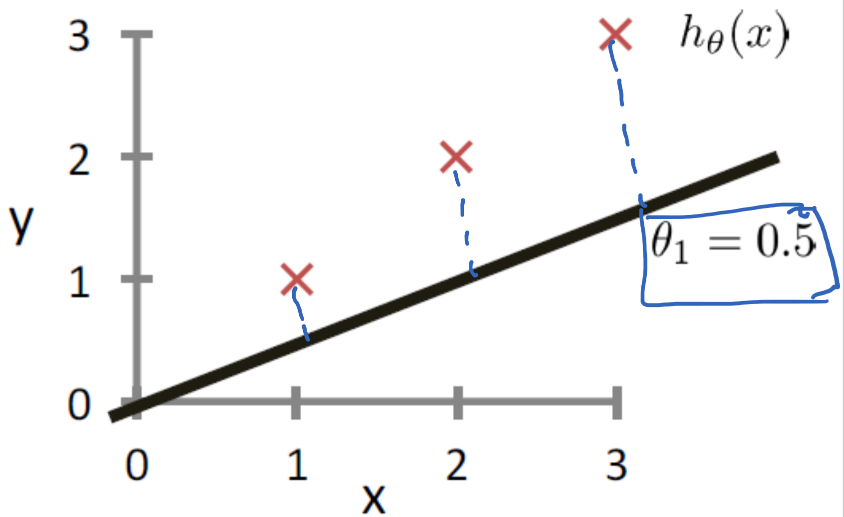
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

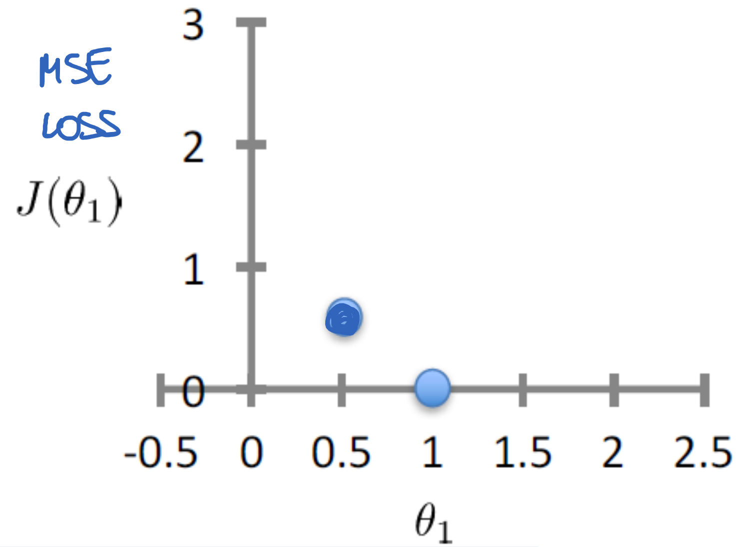
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

(function of the parameter θ_1)



Based on example by Andrew Ng

$$J([0, 0.5]) = \frac{1}{\cancel{2} \times 3}\left[(0.5-1)^2 + (1-2)^2 + (1.5-3)^2\right] \approx 0.58$$

MSE

Intuition on MSE

INPUT: (x_i, y_i)

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2 \Rightarrow$$

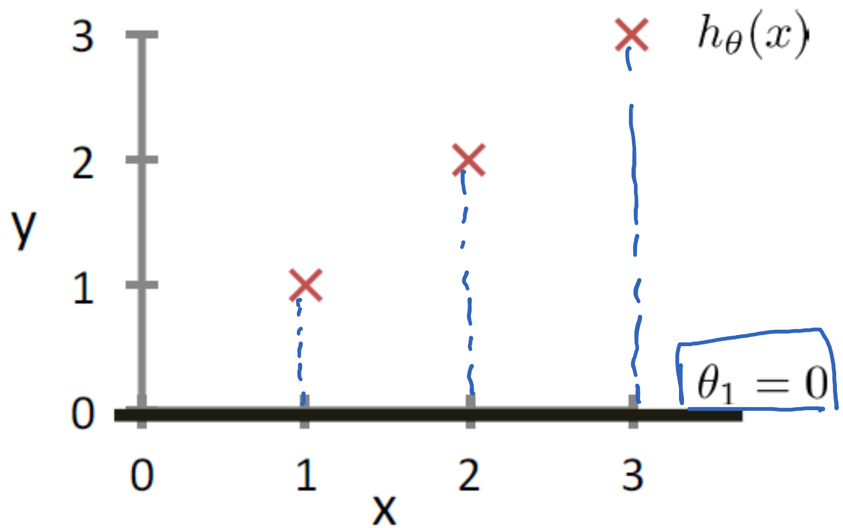
OUTPUT

θ_0, θ_1 : $\min J(\theta)$

For insight on J(), let's assume $x \in \mathbb{R}$ so $\boldsymbol{\theta} = [\theta_0, \theta_1]$

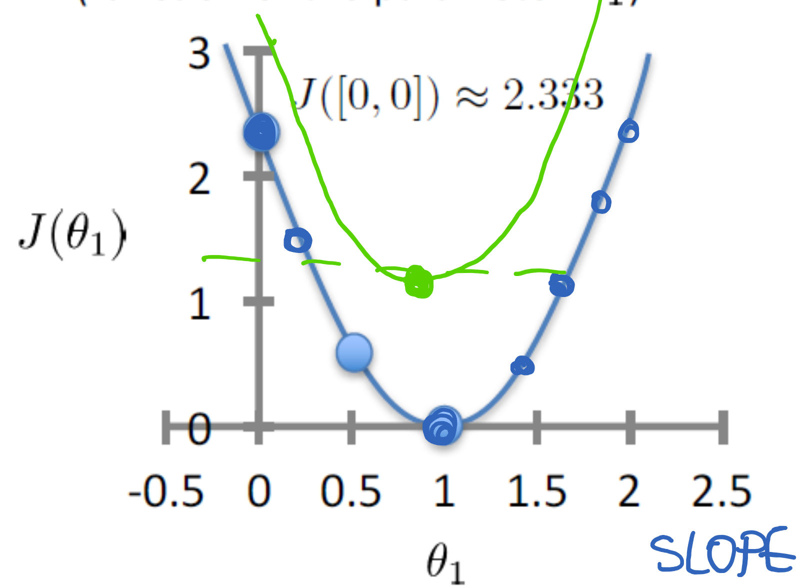
$h_\theta(x)$

(for fixed θ_1, this is a function of x)



$J(\theta_1)$

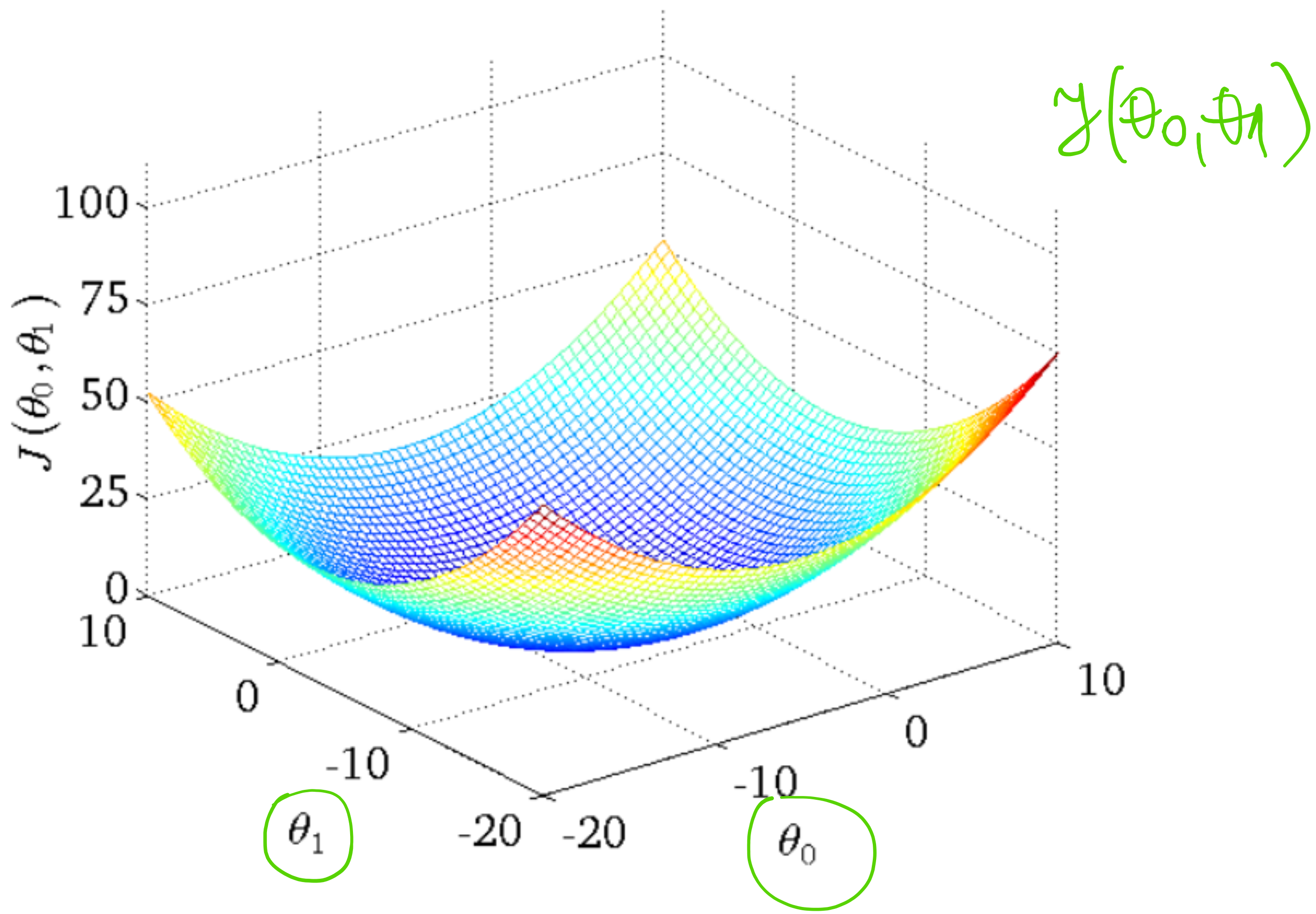
(function of the parameter θ_1)



CONVEX

Based on example
by Andrew Ng

MSE function



Convex function, unique minimum

Solution for simple linear regression

M points

- Dataset $\boxed{x_i \in R, y_i \in R}, h_\theta(x) = \theta_0 + \theta_1 x$
- $J(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\theta_0 + \theta_1 x_i - y_i)^2$ MSE / Loss

Find θ_0 and θ_1

for which $\min J(\theta_0, \theta_1)$

Relationship between Two Random Variables

- Model X (feature / predictor) and Y (response) as two random variables
- Fit of simple linear regression depends on dependence between X and Y
- Covariance
  - Measures the strength of relationship between two random variables
- Pearson correlation
  - Normalized between [-1,1]
  - Proportional to covariance

Covariance

- X and Y are random variables
- $Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$
- Properties

(1) $Cov(X,X) = E\big[(X - E(X))(X - E(X))\big] = E\big[(X - E(X))^2\big] = Var(X)$

(2) $Cov(X,Y) = Cov(Y,X)$

(3) $Cov(aX,Y) = a \cdot Cov(X,Y)$

Covariance

- X and Y are random variables
- $Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$ DEF

$$Cov(X,Y) = E[XY - XE[Y] - YE(X) + E(X)E(Y)]$$
$$= E[XY] - E(X)E(Y) - E(X)E(Y) + E(X)E(Y)$$
$$= E[XY] - E(X)E(Y)$$

If X and Y are indep. $\Rightarrow E[XY] = E[X]E[Y]$

$$\Downarrow$$

$$\boxed{Cov(X,Y) = 0}$$

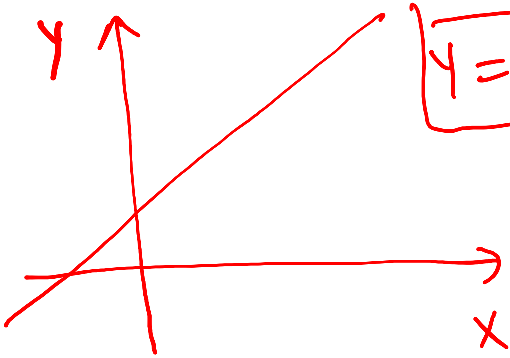
A, B ; $\underbrace{P(A \cap B) = P(A)P(B)}_{\text{DEF}}$

Pearson Correlation

$$\rho = \mathrm{Corr}(X,Y) = \boxed{\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}} \in [-1,1]$$

Standard deviation

$$\boxed{\sigma_X = \sqrt{Var(X)}}$$



$$\boxed{Y = aX + b} \Rightarrow \rho = 1$$

$a \neq 0$

$$\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, aX+b) = \mathrm{Cov}(X, aX) = a \cdot \mathrm{Cov}(X,X)$$
$$= a \cdot Var(X)$$

$\sigma_Y =$

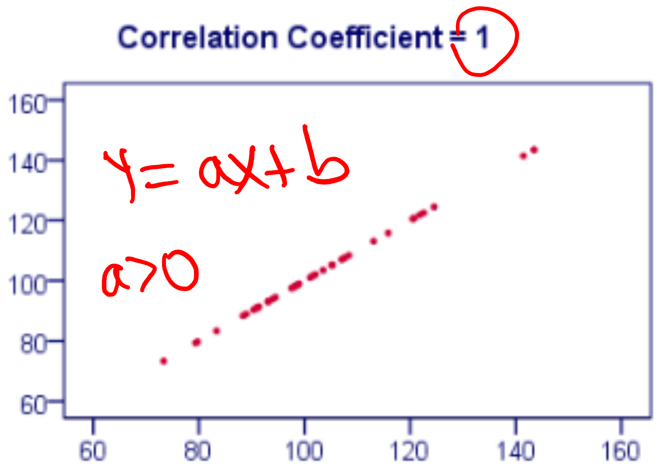
$$Var(Y) = a^2 Var(X) \Rightarrow \sigma_Y = a\sigma_X$$

$$\boxed{\rho = \frac{a \cdot Var(X)}{\sigma_X \cdot a \cdot \sigma_X} = 1}$$

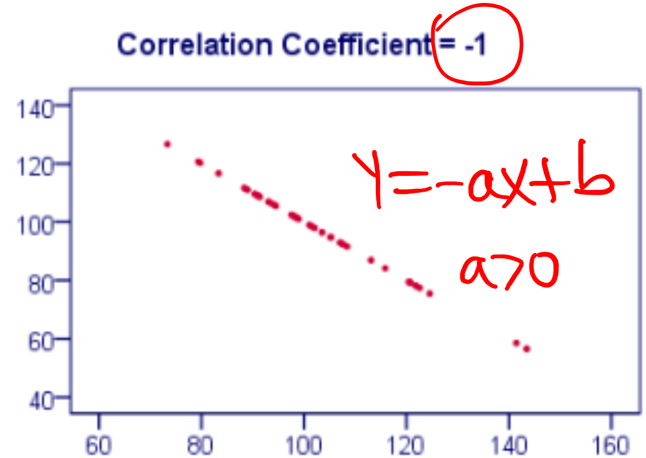
Pearson Correlation

$$\boxed{\rho = \text{Corr}(X,Y) = \frac{\text{Cov(X, Y)}}{\sigma_X \sigma_Y} \in [-1,1]}$$

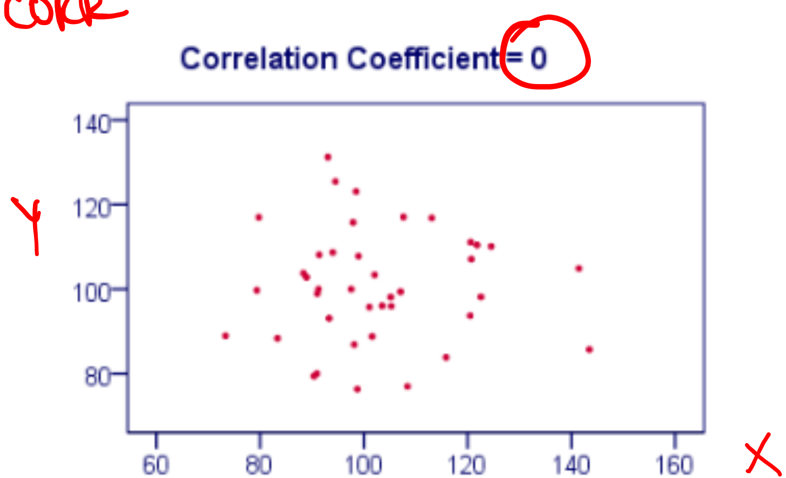
Standard deviation
$$\sigma_X = \sqrt{Var(X)}$$



POSITIVE CORR

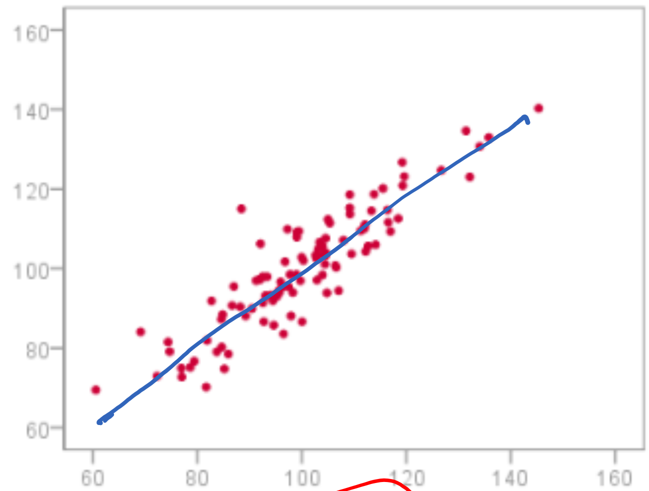


NEGATIVE CORR



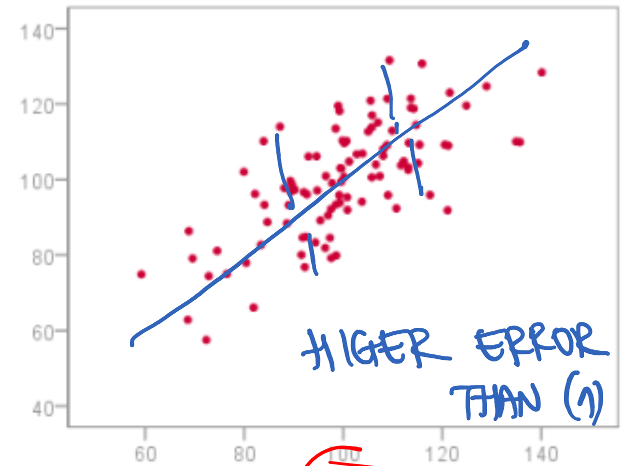
Positive/Negative Correlation

Positive Correlation



r = 0.9

(1)

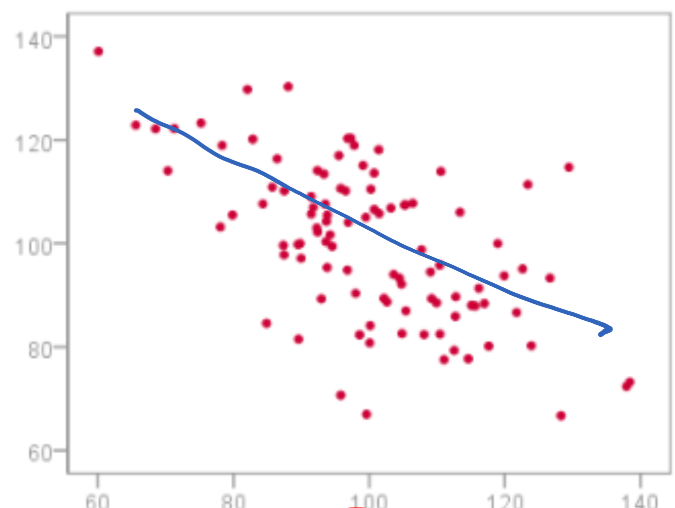


HIGER ERROR THAN (1)

r = 0.7

(2)

Negative Correlation



r = -0.6

HIGHER ERROR THAN (1) & (2)

How Well Does the Model Fit?

- Correlation between feature and response
  - Pearson's correlation coefficient

$$\rho = Corr(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} = \boxed{\frac{\mathrm{Cov(X, Y)}}{\sigma_X \sigma_Y}}$$

σ_X^2

$$\theta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

SLOPE

If $\sigma_X = \sigma_Y \Rightarrow \boxed{\rho = \theta_1}$

Regression vs Correlation

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