DS 4400

Machine Learning and Data Mining I
Spring 2021

Alina Oprea
Associate Professor
Khoury College of Computer Science
Northeastern University

January 28 2021

Today's Outline

- Announcements
  - First numpy tutorial by Prabal M.
    - Thu, Jan 28, 5-6pm
- Linear algebra review
  - Linear independence
  - Rank of a matrix
- Linear regression
  - MSE as loss function
  - Derivation of optimal solution
  - Correlation coefficient, covariance, and connection to regression

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $x_1,\ldots,x_k$ are linearly independent if $c_1x_1+\ldots+c_kx_k = 0$ implies $c_1=\ldots=c_k=0$ ↑ 0 vector
- DEF: ~~Ot~~herwise they are linearly dependent $x_1,\ldots,x_k \in \mathbb{R}$

$$x_1 = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 0 \\ 3 \\ 3 \end{pmatrix}$$

Find c_1, c_2; $c_1x_1 + c_2x_2 = 0$

$$c_1\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + c_2 \cdot \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \Rightarrow \begin{cases} c_1 + c_2 \cdot 0 = 0 \\ 2c_1 + 3c_2 = 0 \\ c_1 + 3c_2 = 0 \end{cases} \Rightarrow c_1 = 0 \Downarrow c_2 = 0$$

LINEARLY INDEPENDENT

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $x_1,\ldots,x_k$ are linearly independent if $c_1x_1+\ldots+c_kx_k = 0$ implies $c_1=\ldots=c_k=0$

LINEARLY DEPENDENT

- Otherwise they are linearly dependent

Find c_1, c_2, c_3

$$c_1 \cdot x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + c_2 \cdot x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} + c_3 \cdot x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\begin{cases} c_1 + 4c_2 + 2c_3 = 0 & (1) \\ 2c_1 + c_2 - 3c_3 = 0 & (2) \\ 3c_1 + 5c_2 - c_3 = 0 & (3) \end{cases}$$

$(1) + 2 \cdot (3)$

$$\begin{cases} 7c_1 + 14c_2 = 0 \Rightarrow c_1 = -2c_2 \\ -6c_2 + 5c_2 - c_3 = 0 \\ c_2 = -c_3 \Rightarrow c_3 = -c_2 \end{cases}$$

$$\boxed{\begin{aligned} c_2 &= 1 \\ c_1 &= -2 \\ c_3 &= -1 \end{aligned}}$$

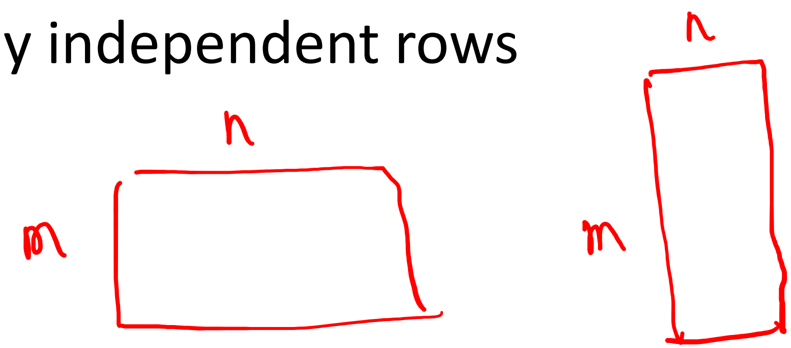
$$\boxed{x_3 = -2x_1 + x_2}$$

Rank of a Matrix

- rank(A) (the rank of a m-by-n matrix A) is

  The maximal number of linearly independent columns

  The maximal number of linearly independent rows



- If A is n by m, then
  - rank(A)<= min(m,n)

- Examples

A square matrix $n \times n$ is invertible if and only if rank$(A) = n$

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

RANK = 2

$$\begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix}$$

RANK = 1

$$\begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

RANK = 2

Linear regression

Supervised Learning: Regression

Training

Data → Pre-processing → Feature extraction → Learning model

Labeled

$x_i, y_i \in R$

Normalization

Feature Selection

Regression

$f(x)$

Testing

New data → Learning model → Predictions

$y' = f(x') \in R$

Unlabeled

x'

$f(x')$

Response variable

Regression

Steps to Learning Process

- Define problem space
- Collect data
- Extract feature
- Pick a model (hypothesis) → CLASS OF MODELS
- Develop a learning algorithm
  - Train and learn model parameters

  OPTIMIZATION "BEST FIT"
- 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:

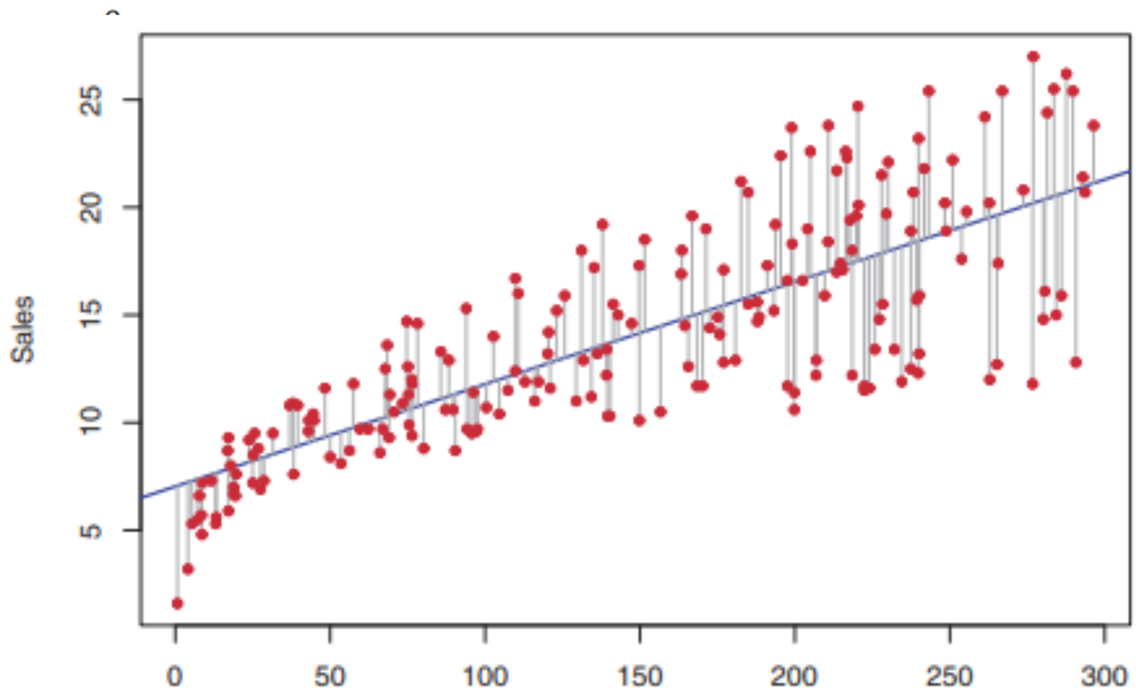
FEATURES

– Data $X = \{\underbrace{x_1, \dots x_N}\}$, where $x_i \in R^d$

– Corresponding labels $Y = \{y_1, \dots y_N\}$, where $y_i \in R$

RESPONSE

y

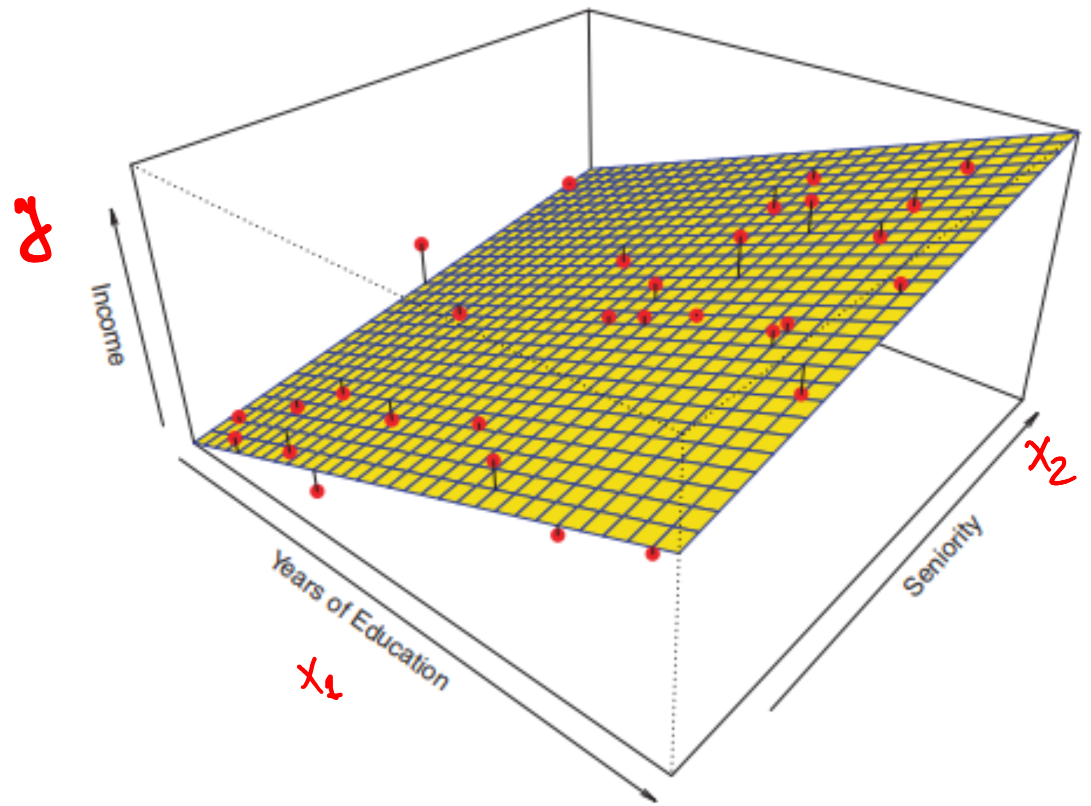


X

SIMPLE LR : 1 FEATURE

d=1

Income Prediction

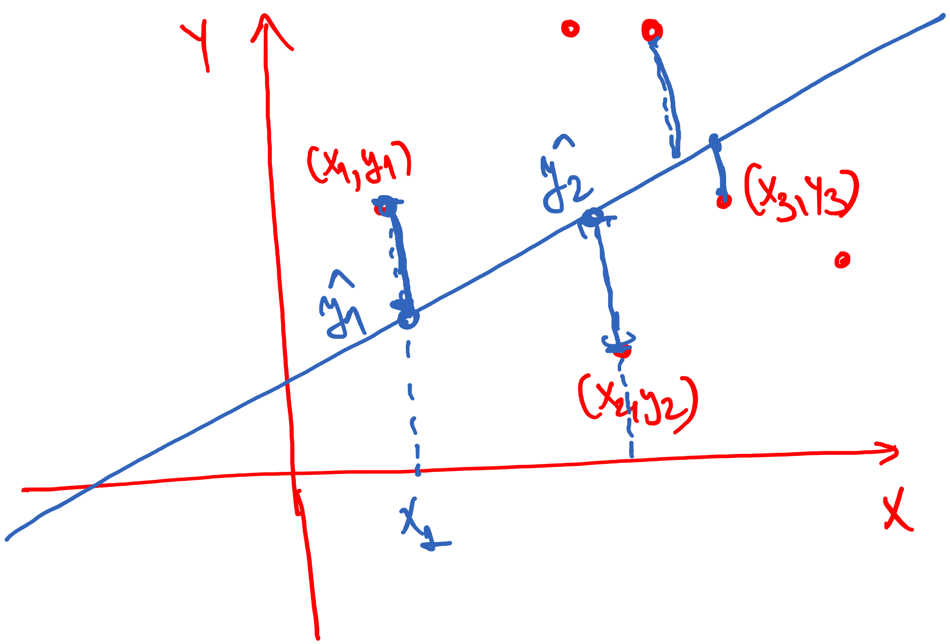


MULIPLE LR
$d > 1$

Hypothesis: Linear Model

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

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



GIVEN: $(x_1, y_1), (x_2, y_2), \ldots$

$\ldots (x_N, y_N)$

$\hat{y}_1 = h_\theta(x_1) = \theta_0 + \theta_1 x_1$

$\hat{y}_2 = \theta_0 + \theta_1 x_2$

$\vdots$

y_i = TRUE RESPONSE

$\hat{y}_i$ = PREDICTED RESPONSE

$y_i - \hat{y}_i$ = RESIDUAL

Least-Squares Linear Regression

CONVEX OBJECTIVE

- Cost Function

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

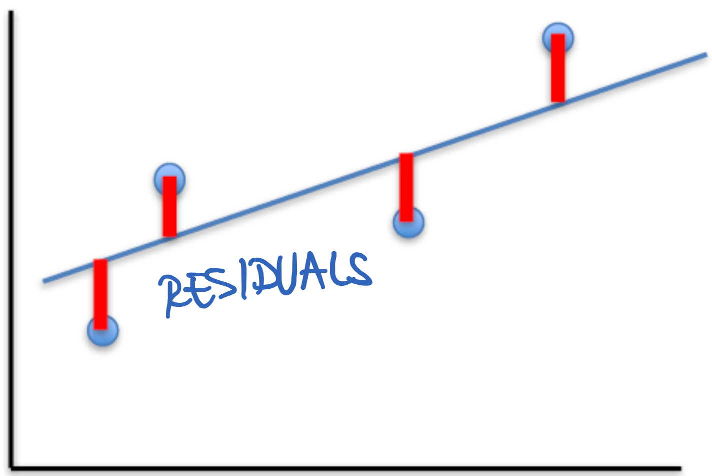
$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[\underbrace{h_\theta(x_i) - y_i}_{\text{RESIDUAL}}]^2$$

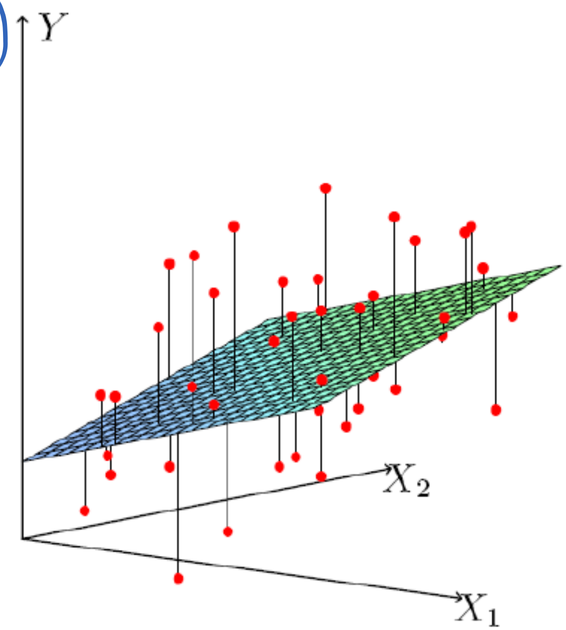
fix

Mean Square Error (MSE)

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

Find θ_0 and θ_1 that min $J(\theta)$





Terminology and Metrics

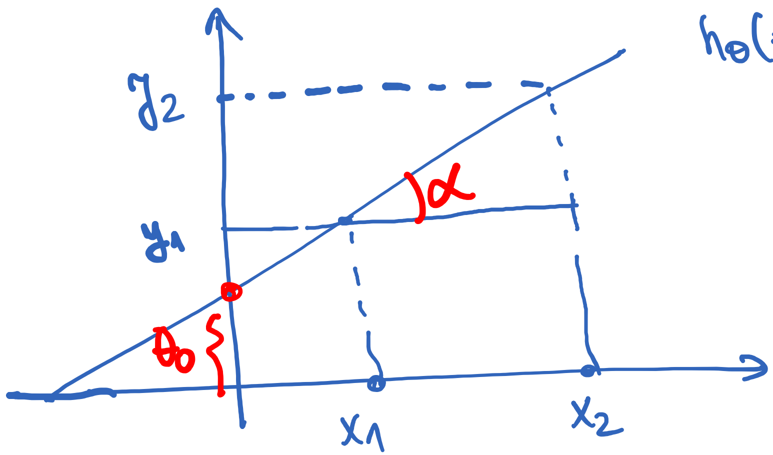
- Residuals
  - Difference between predicted values and actual values
  - Predicted value for example i is: $\widehat{y_i} = h_\theta(x_i)$
  - $R_i = \left|y_i - \widehat{y_i}\right| = |y_i - \underbrace{(\theta_0 + \theta_1 x_i}_{h_\theta(x_i)})|$
- Residual Sum of Squares (RSS)

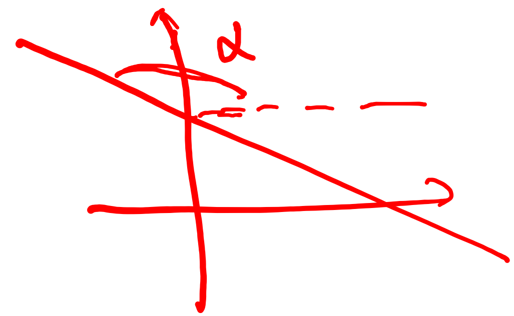
SAME MIN

  - $RSS = \sum {R_i}^2 = \sum\left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$
- Mean Square Error (MSE)
  - $MSE = \frac{1}{N}\sum R_i^2 = \frac{1}{N}\sum\left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$

Interpretation

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





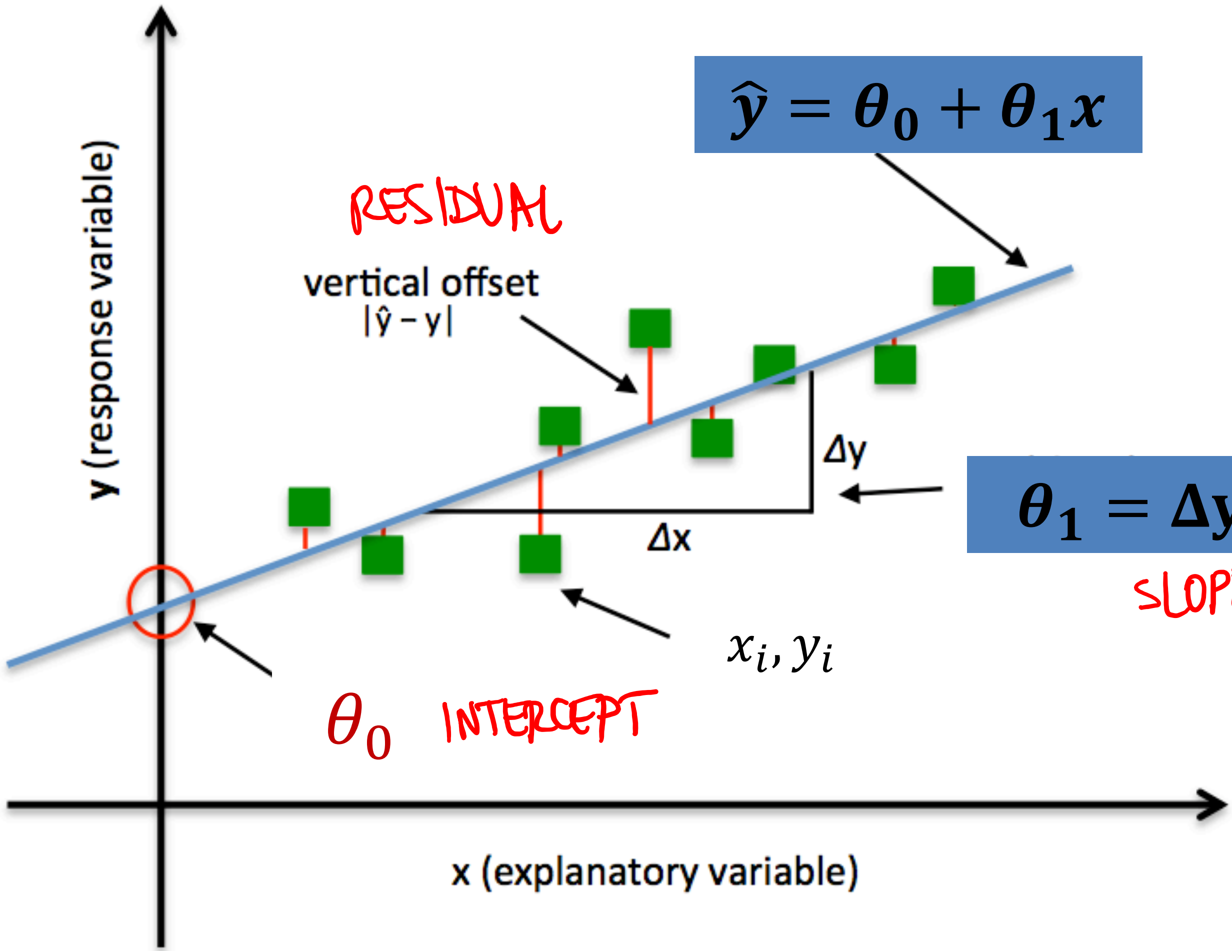
$\theta_1 = \dfrac{y_2 - y_1}{x_2 - x_1}$ SLOPE

$\tan(\alpha) = \theta_1$

$\theta_1 = 1$, $h_\theta(x) = \theta_0 + X \Rightarrow \alpha = 45^\circ$

$\theta_1 = -1$, $h_\theta(x) = \theta_0 - X \Rightarrow \alpha = 135^\circ$

Interpretation



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

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

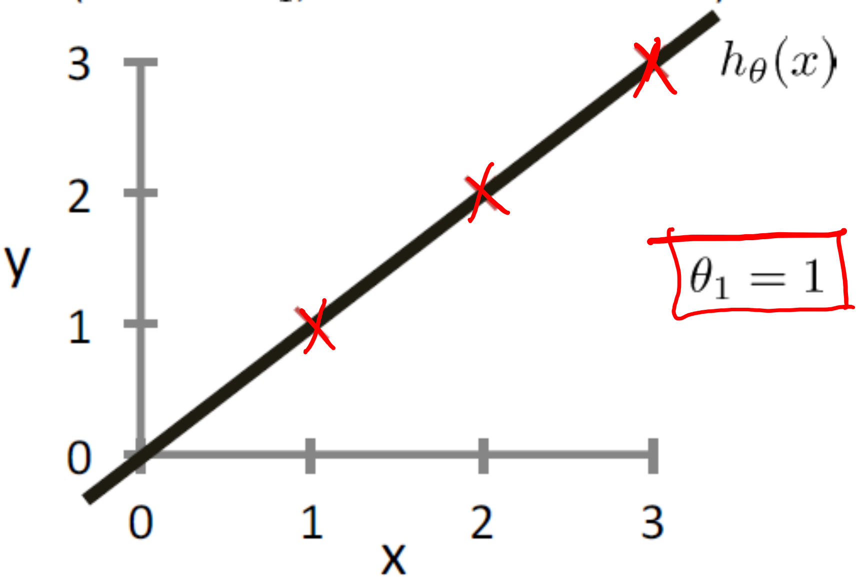
Intuition on MSE

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

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

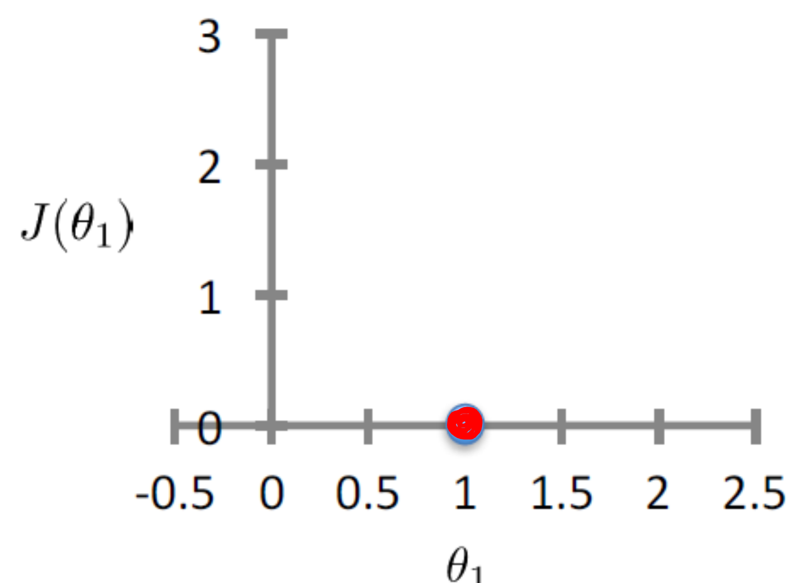
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Fix $\boxed{\theta_0 = 0}$ $h_\theta(x) = \theta_1 x$

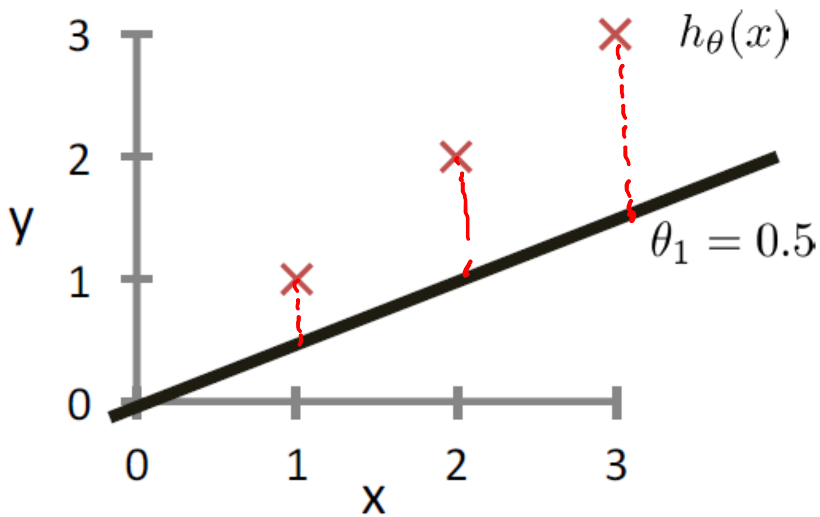
Intuition on MSE

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

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

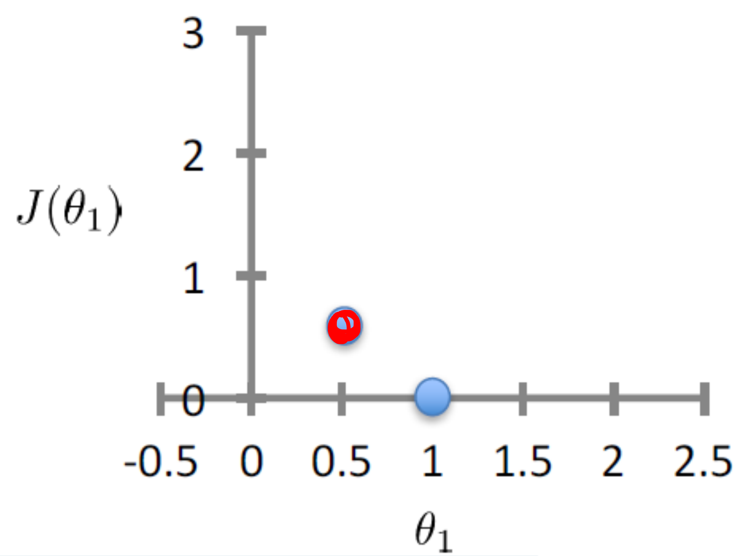
$h_\theta(x)$

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



$J(\theta_1)$

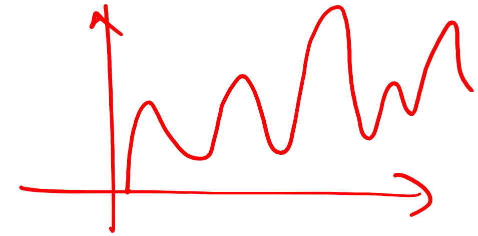
(function of the parameter θ_1)



Based on example by Andrew Ng

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

Intuition on MSE



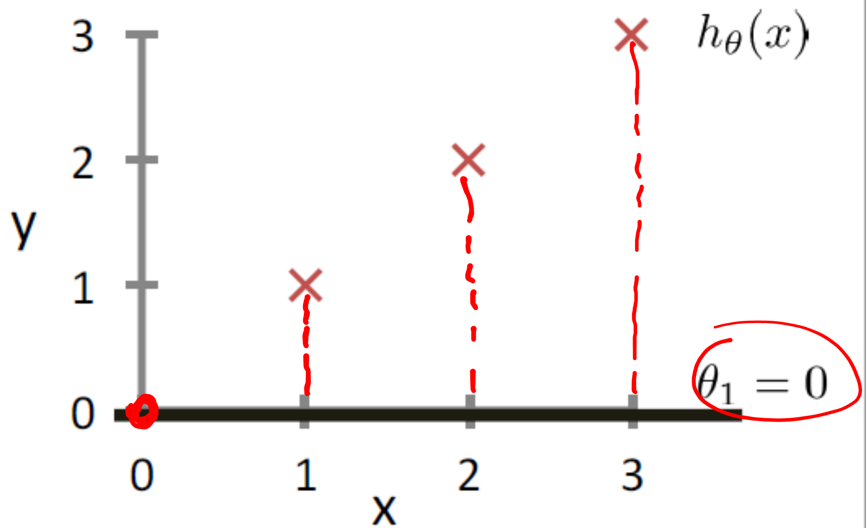
$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

$J(\theta_0, \theta_1)$

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

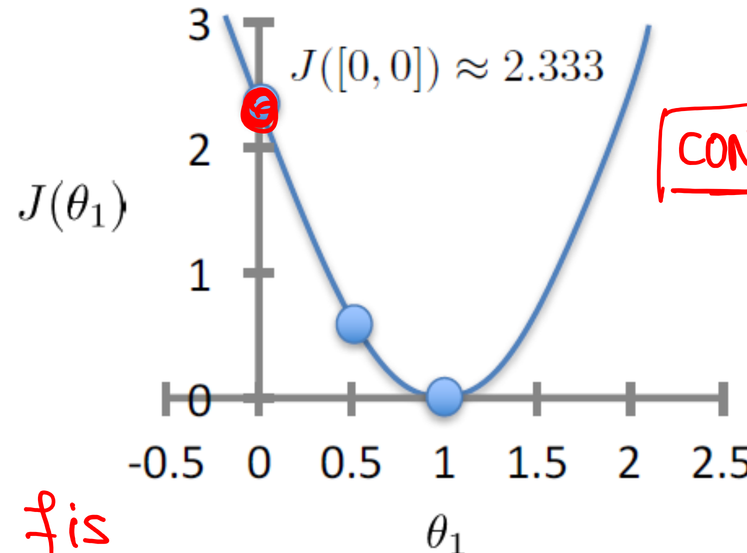
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)

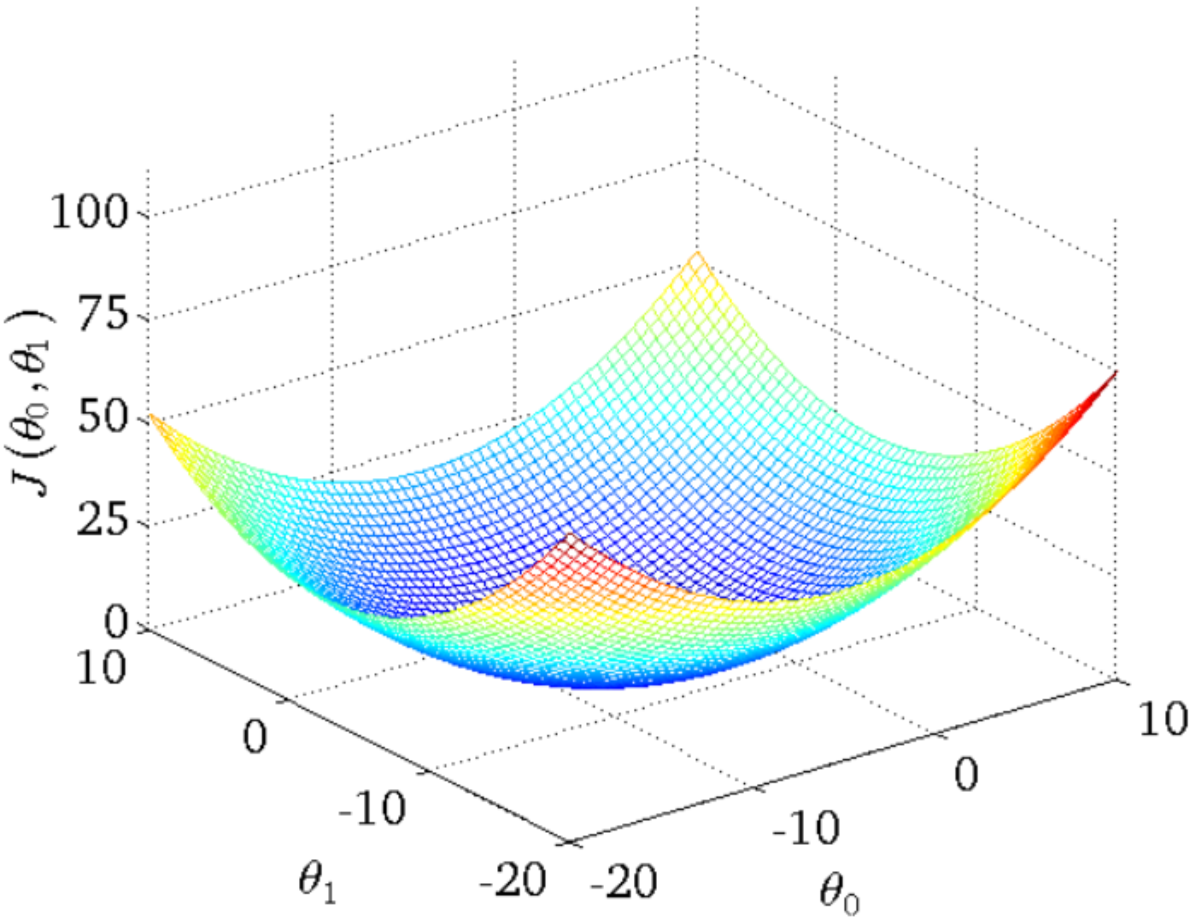


$\boxed{\text{CONVEX}}$

f is CONEX if $f''(x) \geq 0$

Based on example by Andrew Ng

MSE function



Convex function, unique minimum

Solution for simple linear regression

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

Find θ_0 and θ_1 to min $J(\theta)$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

$$\begin{cases} \frac{\partial J(\theta)}{\partial \theta_0} = \frac{1}{N}\sum_{i=1}^{N} 2(\theta_0 + \theta_1 x_i - y_i) = 0 & (1) \\ \frac{\partial J(\theta)}{\partial \theta_1} = \frac{1}{N}\sum_{i=1}^{N} 2(\theta_0 + \theta_1 x_i - y_i) \cdot x_i = 0 & (2) \end{cases}$$

(1): $\theta_0 + \theta_1 \bar{x} - \bar{y} = 0 \Rightarrow \boxed{\theta_0 = \bar{y} - \theta_1 \bar{x}}$

(2): $\theta_0 \cdot \bar{x} + \theta_1 \frac{\sum x_i^2}{N} - \frac{\sum x_i y_i}{N} = 0$

$\bar{x}\bar{y} - \theta_1 \bar{x}^2 + \theta_1 \frac{\sum x_i^2}{N} - \frac{\sum x_i y_i}{N} = 0$

$$N\bar{x}\bar{y} - N\theta_1\bar{x}^2 + \theta_1 \sum x_i^2 - \sum x_i y_i = 0$$

$$\theta_1 = \frac{\boxed{\sum x_i y_i - N\bar{x}\bar{y}}}{\boxed{\sum x_i^2 - N\bar{x}^2}}$$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$ SAMPLE MEAN

SAMPLE VARIANCE

$\mathrm{Var}(x)\,(N-1)$

$$\mathrm{Var}(x) = \frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

CLAIM: $$\theta_1 = \frac{\boxed{\sum (x_i - \bar{x})(y_i - \bar{y})}}{\boxed{\sum (x_i - \bar{x})^2}}$$

$$\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \underbrace{\left(\sum x_i\right)}_{N\bar{x}}\bar{y} - \bar{x}\underbrace{\sum y_i}_{N\bar{y}} + N\bar{x}\bar{y}$$

$$= \sum x_i y_i - N\bar{x}\bar{y}$$

$$\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\underbrace{\sum x_i}_{N\bar{x}} + N\bar{x}^2 = \sum x_i^2 - N\bar{x}^2$$

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

Solution for simple linear regression

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

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

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

- Solution of min loss

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

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

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

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
- $Cov(X, Y) = E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]$
- Properties

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

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

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

Covariance

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

$$= E[XY] - E[X \cdot E(Y)] - E(E(X)Y) + E(X)E(Y)$$

$$= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)$$

$$= E(XY) - E(X)E(Y)$$

If X and Y are indep.

$P(X = x \wedge Y = y) = P(X = x) \cdot P(Y = y)$ $\rightsquigarrow$ $E(XY) = E(X)E(Y)$

$\forall x, y$

$\Rightarrow Cov(X,Y) = 0$

Estimating mean, variance, and covariance

$x_1, \dots, x_N$ DATA (SAMPLES OF X) ; $y_1, \dots, y_N$

SAMPLES OF Y

SAMPLE MEAN: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

SAMPLE VAR: $\mathrm{Var}(\bar{x}) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

SAMPLE COVAR: $\mathrm{cov}(\bar{x}, \bar{y}) = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$

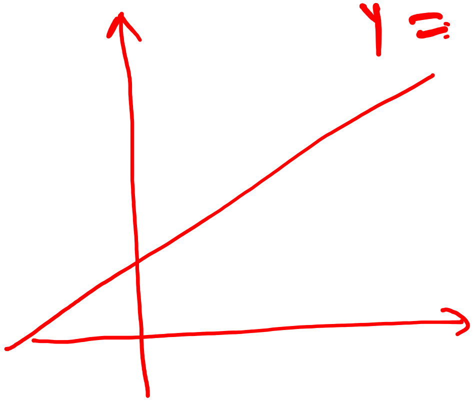
Pearson Correlation

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

Standard deviation

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

$Y = \theta_0 + X \cdot \theta_1$



$$\operatorname{Cov}(X,Y) = \operatorname{Cov}(X, \theta_0 + \theta_1 X)$$

$$= \operatorname{Cov}(X, \theta_1 X) = \theta_1 \cdot \operatorname{Var}(X)$$

$$\rho = \frac{\theta_1 \cdot \operatorname{Var}(X)}{\sigma_X^2} \neq \theta_1$$

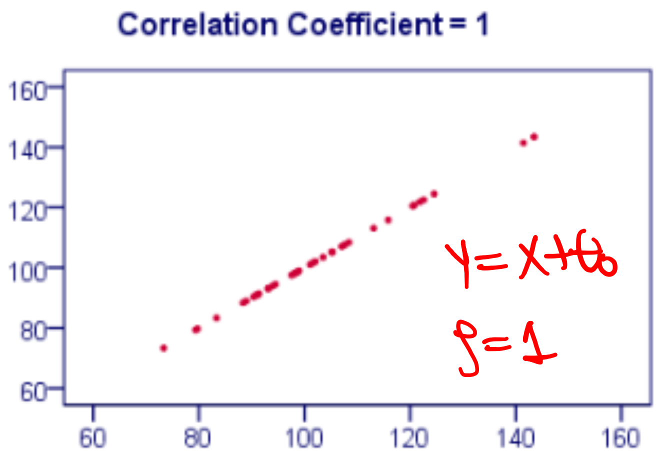
SLOPE

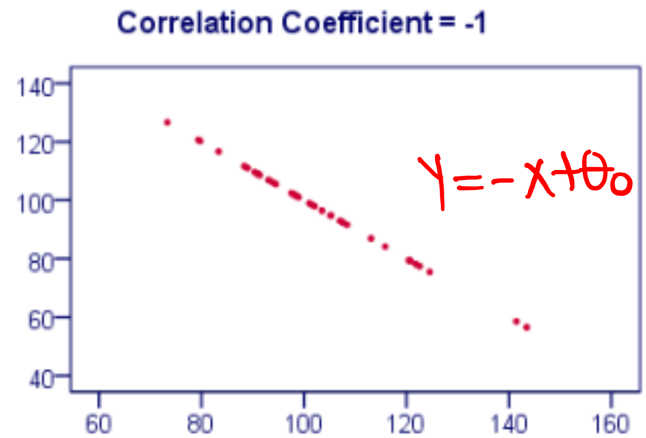
Pearson Correlation

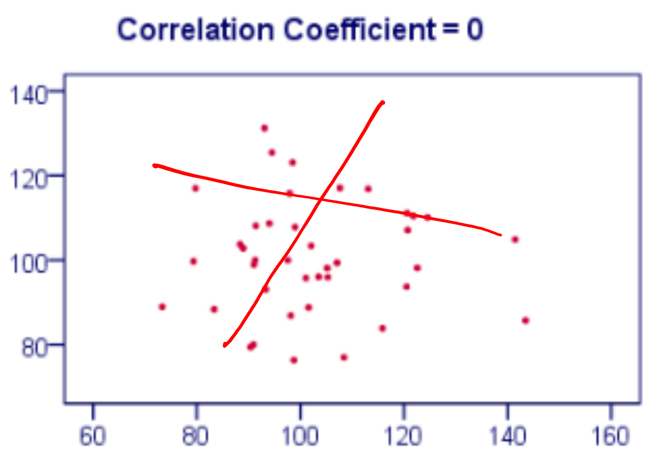
$$\rho = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov(X,Y)}}{\sigma_X \sigma_Y} \in [-1,1]$$

Standard deviation

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

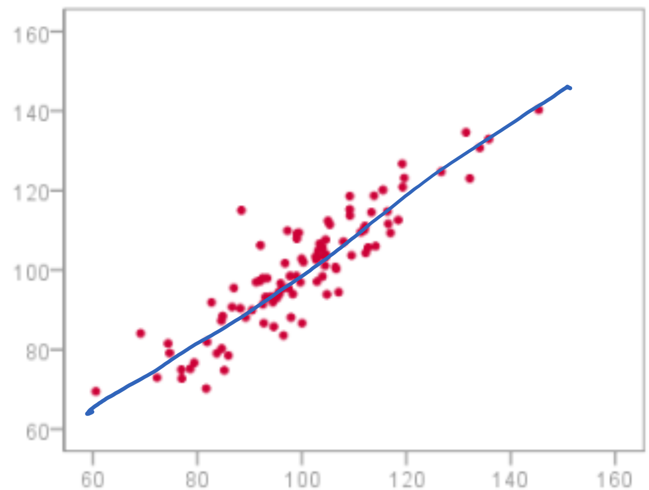




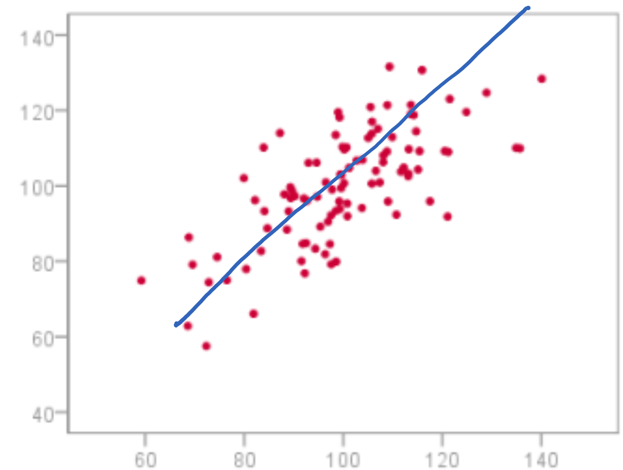


Positive/Negative Correlation

Positive Correlation

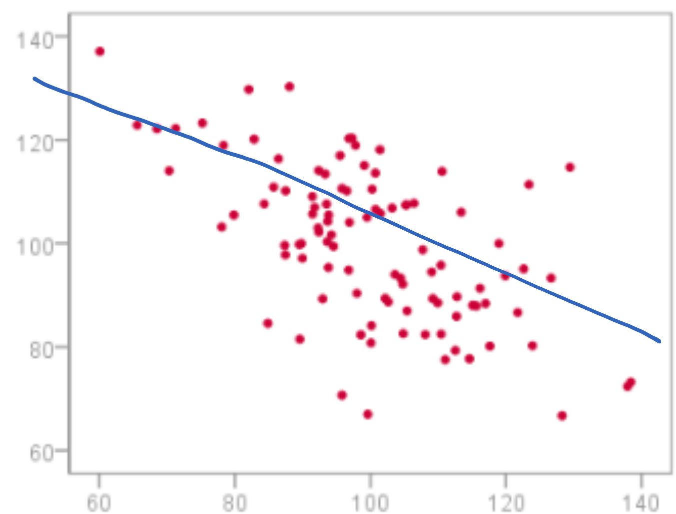


r = 0.9



r = 0.7

Negative Correlation



r = -0.6

30

How Well Does the Model Fit?

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

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

If $\sigma_X = \sigma_Y \Rightarrow \boxed{\rho = \frac{Cov(X,Y)}{Var(X)}} = \theta_1$ (SLOPE)

for optimal model that min MSE

How Well Does the Model Fit?

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

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

- Measures linear dependence between X and Y
- Positive coefficient implies positive correlation
  - The closer to 1 the coefficient is, the stronger the correlation
- Negative coefficient implies negative correlation
  - The closer to -1 the coefficient is, the stronger the correlation
- $\theta_1 = \frac{\sum \ (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\mathrm{Cov(X,Y)}}{Var[X]}$
- If σ_X = σ_Y, then $\theta_1 = \mathrm{Corr(X, Y)}$

How Well Does the Model Fit?

- Residual Sum of Squares
  - $RSS = \sum[R_i\ \ ]^2 = \sum[y_i - (\theta_0 + \theta_1 x_i)]^2$
- Total Sum of Squares
  - $TSS = \sum[y_i - \bar{y}]^2$ $= Var(Y)$
  - Total variance of the response
- Proportion of variability in Y that can be explained using X
  - $\boxed{R^2 = 1 - \frac{RSS}{TSS} \in [0,1]}$
- Correlation between feature and response

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

For simple regression R^2 is equal to ρ^2

Regression vs Correlation

- Correlation
  - Find a numerical value expressing the relationship between variables
  - Pearson correlation 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