DS 4400

Machine Learning and Data Mining I
Spring 2024

David Liu

Khoury College of Computer Science

Northeastern University

Friday January 19, 2024

Today's Outline

- Announcements
  - HW 1 is due on Jan 30 at 11:59pm (Gradescope)
  - Pb 1,2,3, 4(a), and 4(b) are math
  - Pb 4(c) and 5 are coding
  - Submit PDF and include link to code
  - Preferrable to type solutions, but can also handwrite
- Linear regression
  - Simple and multiple linear regression
  - Derivation of optimal solution
  - Correlation coefficient, covariance, and connection to regression

OCTOBER 13, 2023 | 5 MIN READ

The AI Boom Could Use a Shocking Amount of Electricity

Powering artificial intelligence models takes a lot of energy. A new analysis demonstrates just how big the problem could become

BY LAUREN LEFFER

Credit: Erik Isakson/Getty Images

TECH · A.I.

A.I. tools fueled a 34% spike in Microsoft's water consumption, and one city with its data centers is concerned about the effect on residential supply

BY MATT O'BRIEN, HANNAH FINGERHUT AND THE ASSOCIATED PRESS
September 9, 2023 at 11:01 AM EDT

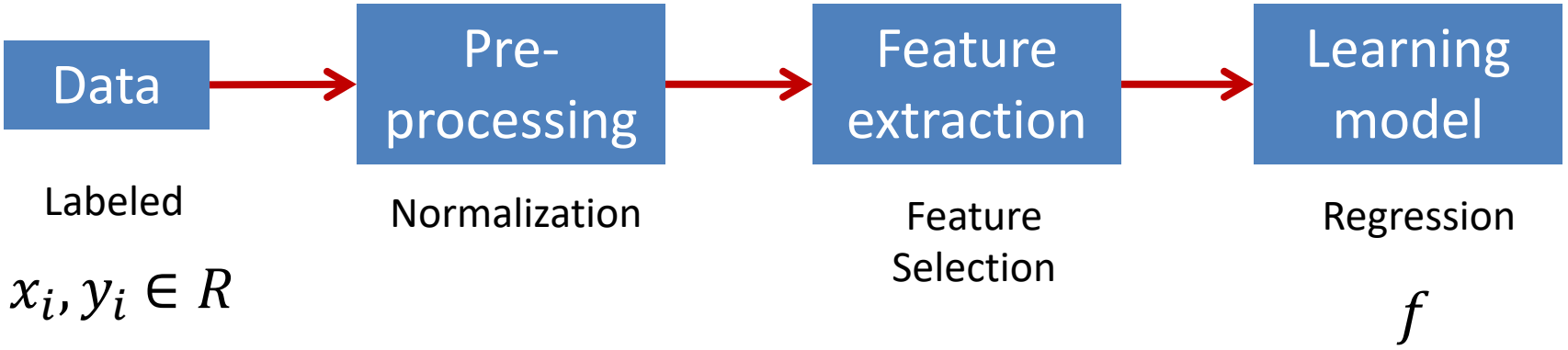
For much of the year, Iowa's weather is cool enough for Microsoft to use outside air to keep the supercomputer running properly and vent heat out of the building. Only when the temperature exceeds 29.3 degrees Celsius (about 85 degrees Fahrenheit) does it withdraw water, the company has said in a public disclosure.

That can still be a lot of water, especially in the summer. In July 2022, the month before OpenAI says it completed its training of GPT-4, Microsoft pumped in about 11.5 million gallons of water to its cluster of Iowa data centers, according to the West Des Moines Water Works. That amounted to about 6% of all the water used in the district, which also supplies drinking water to the city's residents.

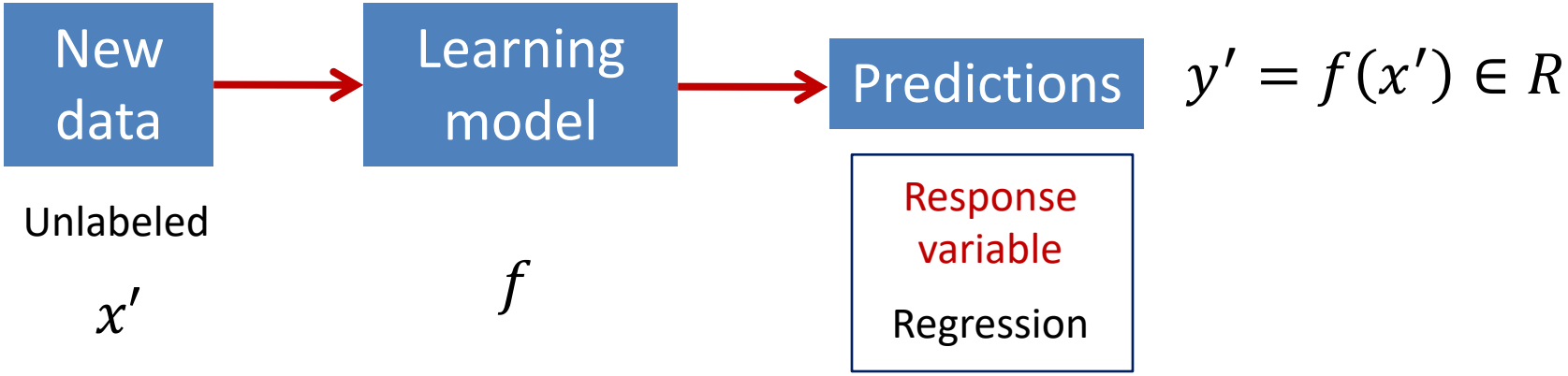
Linear regression

Supervised Learning: Regression

Training



Testing



Steps to Learning Process

- Define problem space
- Collect data
- Extract features
- Pick hypothesis space
- Develop a learning algorithm
  - Train and learn model parameters
- Make predictions on new data
  - Testing phase
- In practice, usually re-train when new data is available and use feedback from deployment

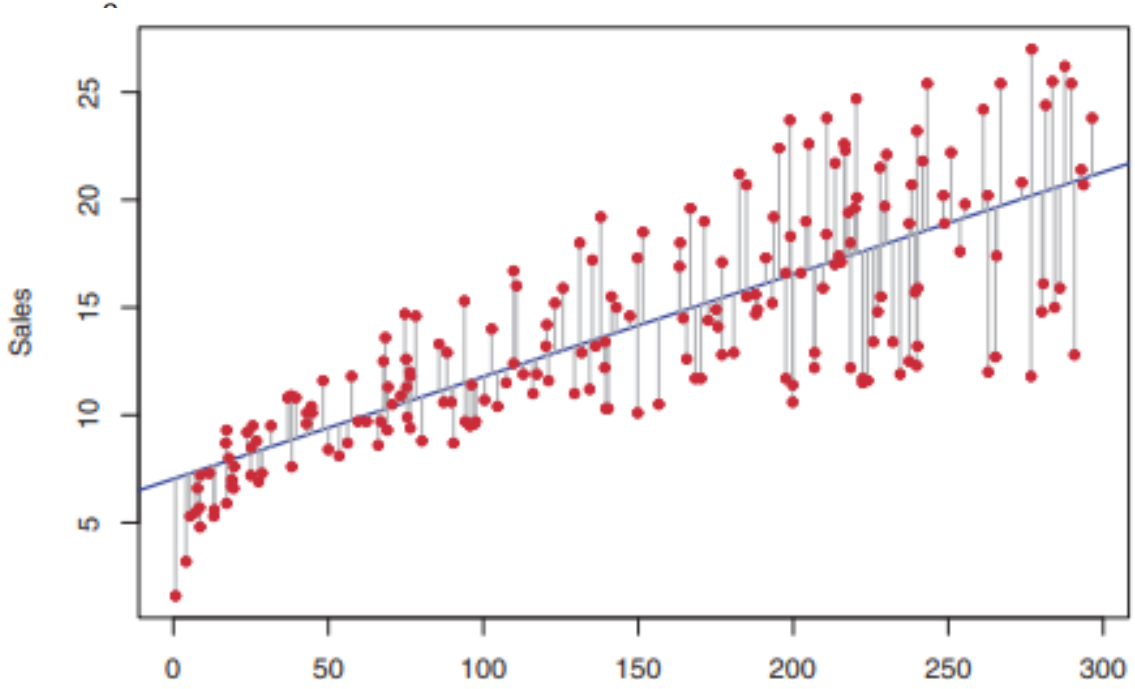
Linear regression

- The oldest statistical learning method
  - Legendre and Gauss 1805
- 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 find optimal solution in closed form
- Efficient practical algorithm (gradient descent)

Linear regression

Given:

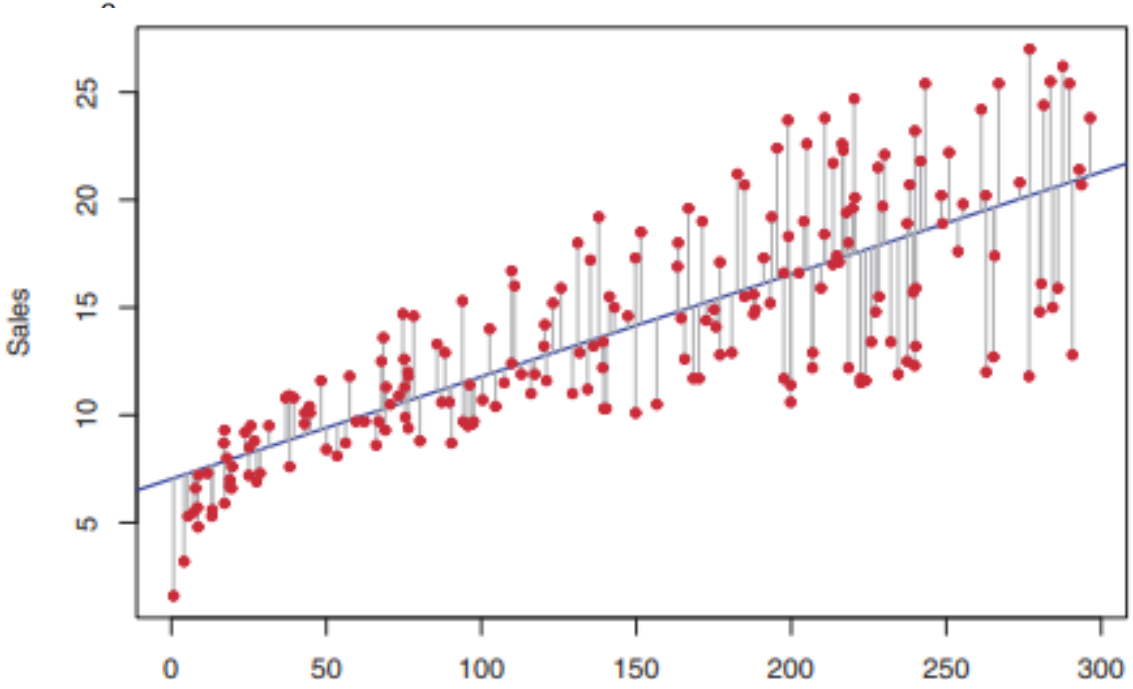
- Data $X = \{x_1, \ldots x_N\}$, where $x_i \in R^d$
- Corresponding labels $Y = \{y_1, \ldots y_N\}$, where $y_i \in R$



Linear regression

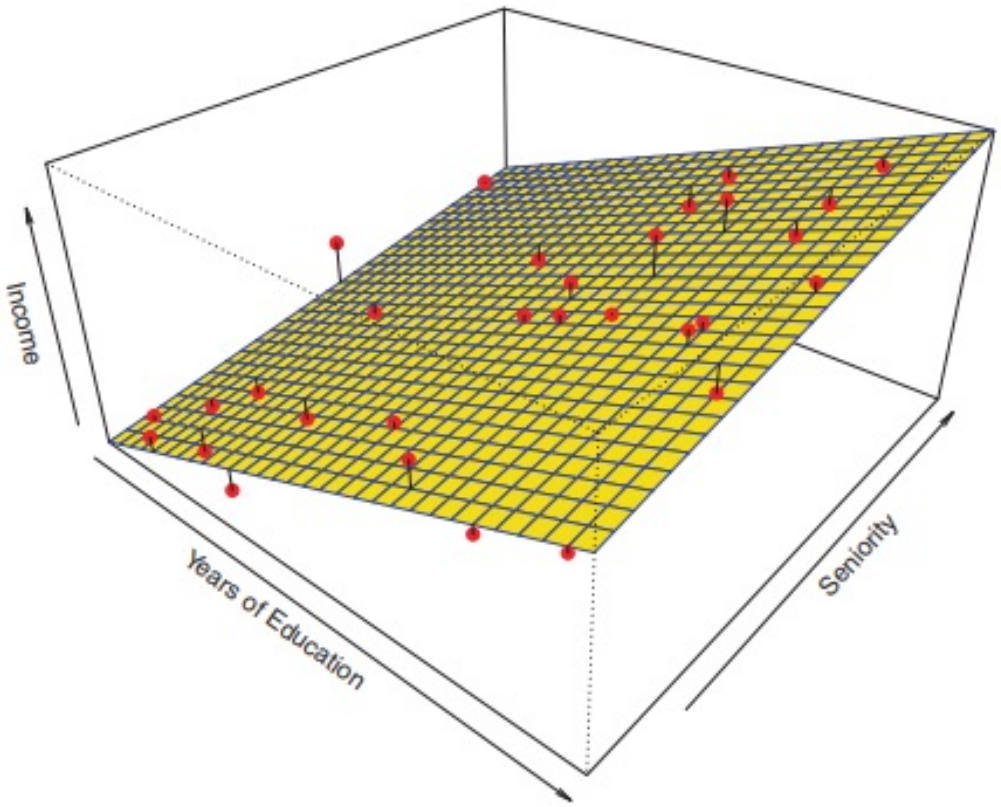
Given:

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

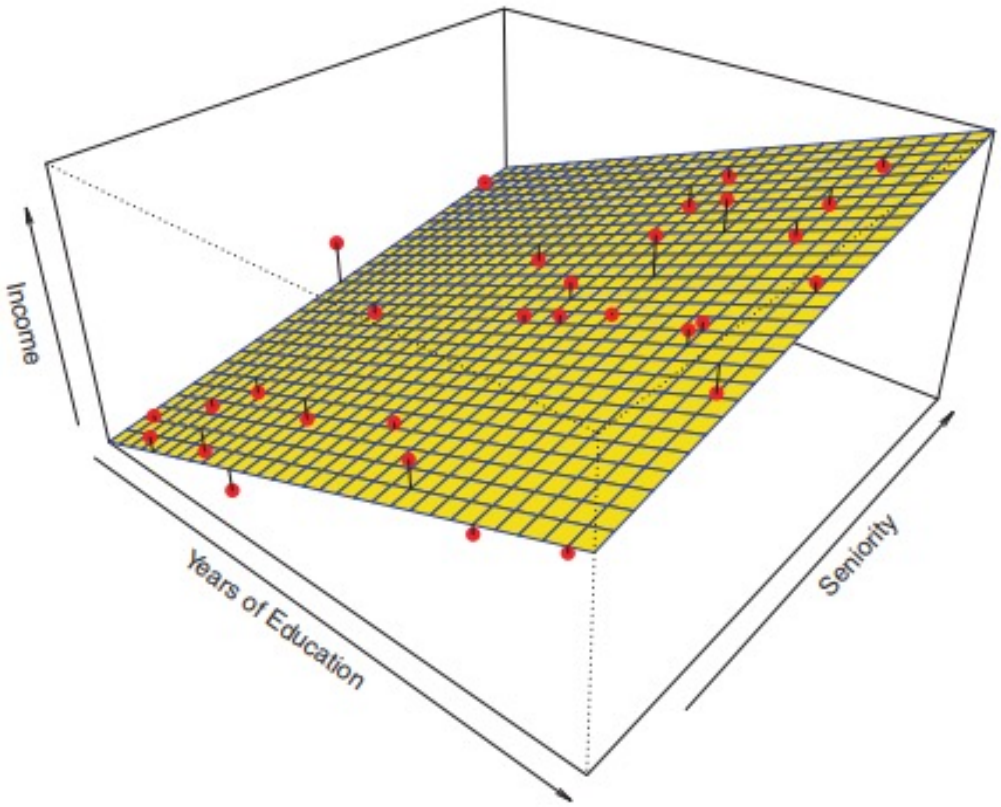


Simple Linear Regression: 1 predictor

Income Prediction



Income Prediction



Linear Regression with 2 predictors

Multiple Linear Regression

Hypothesis: Linear Model

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

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

Least-Squares Linear Regression

- Cost Function

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

Mean Square Error (MSE)

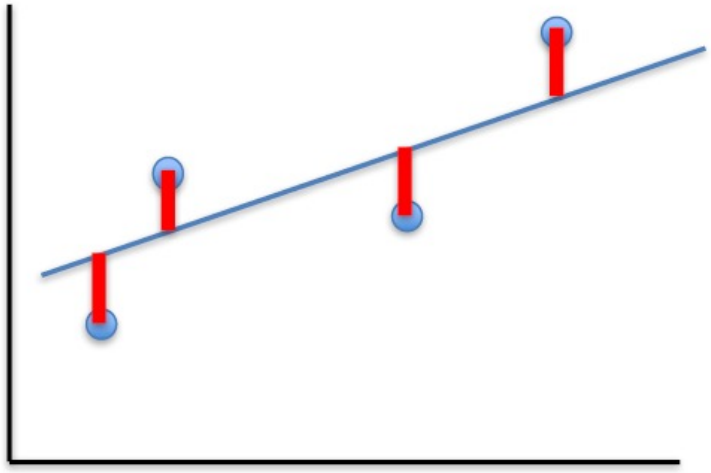
Least-Squares Linear Regression

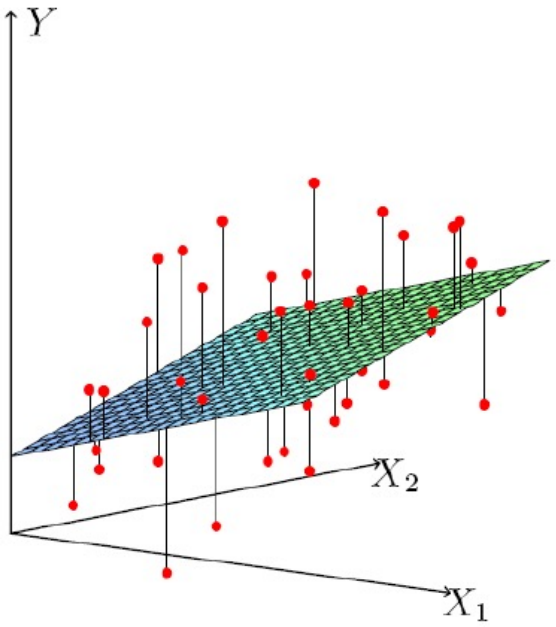
- Cost Function

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

Mean Square Error (MSE)

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

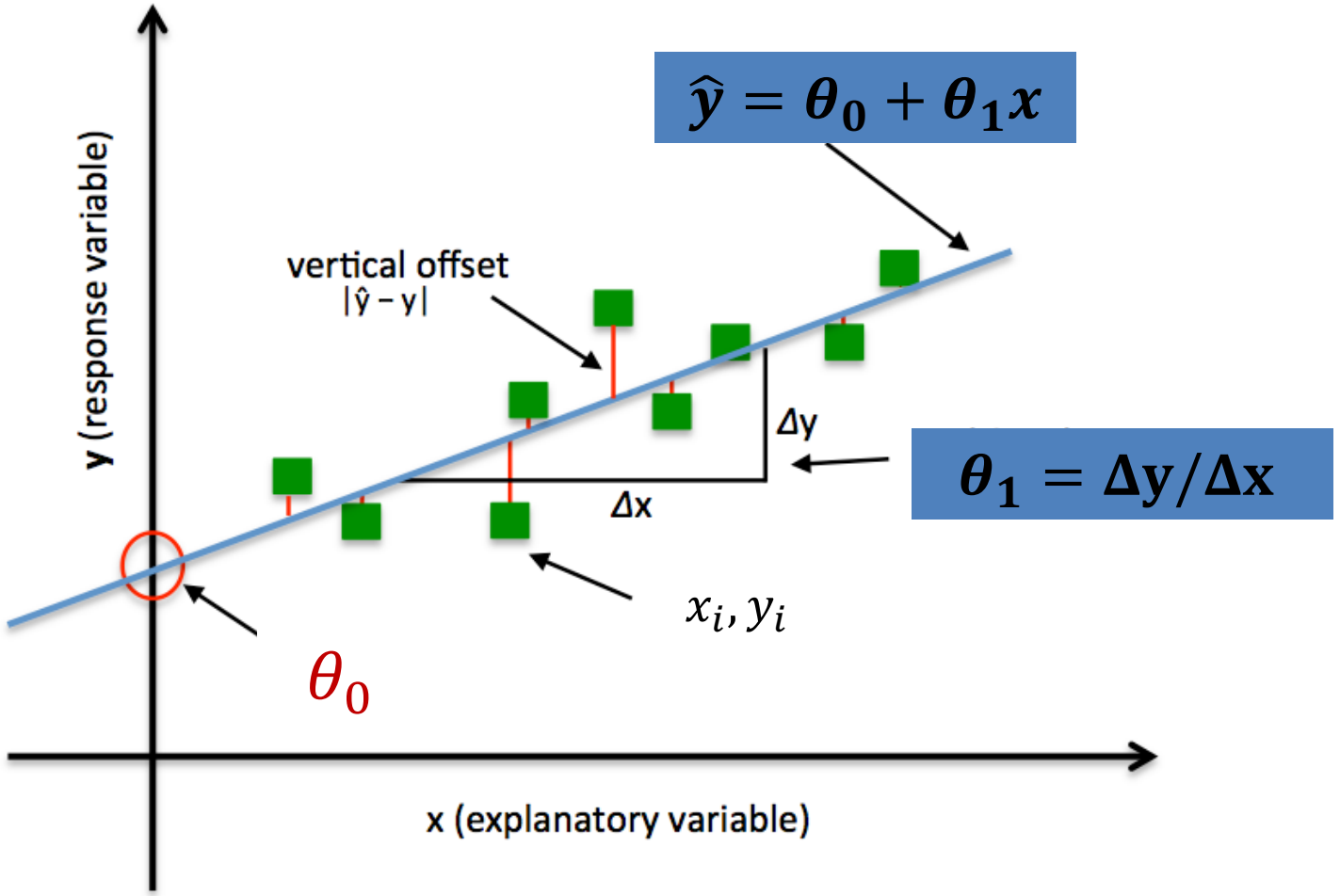




Terminology and Metrics

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

Interpretation



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

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

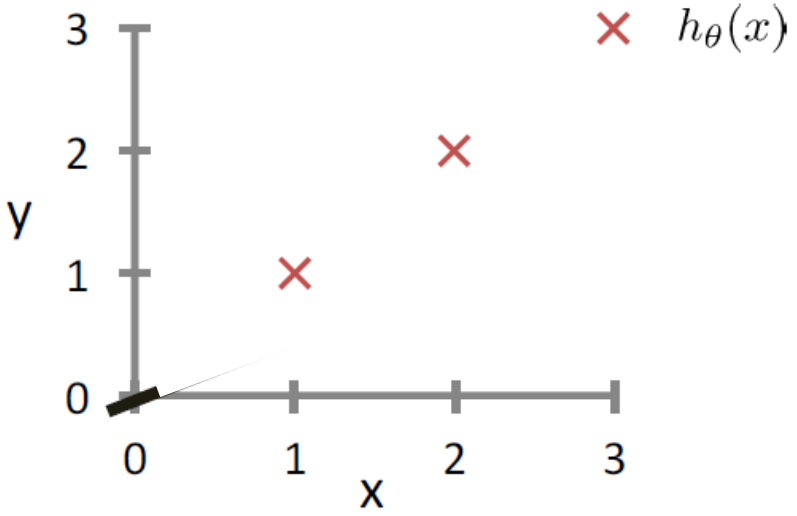
Intuition on MSE

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

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

$h_\theta(x)$

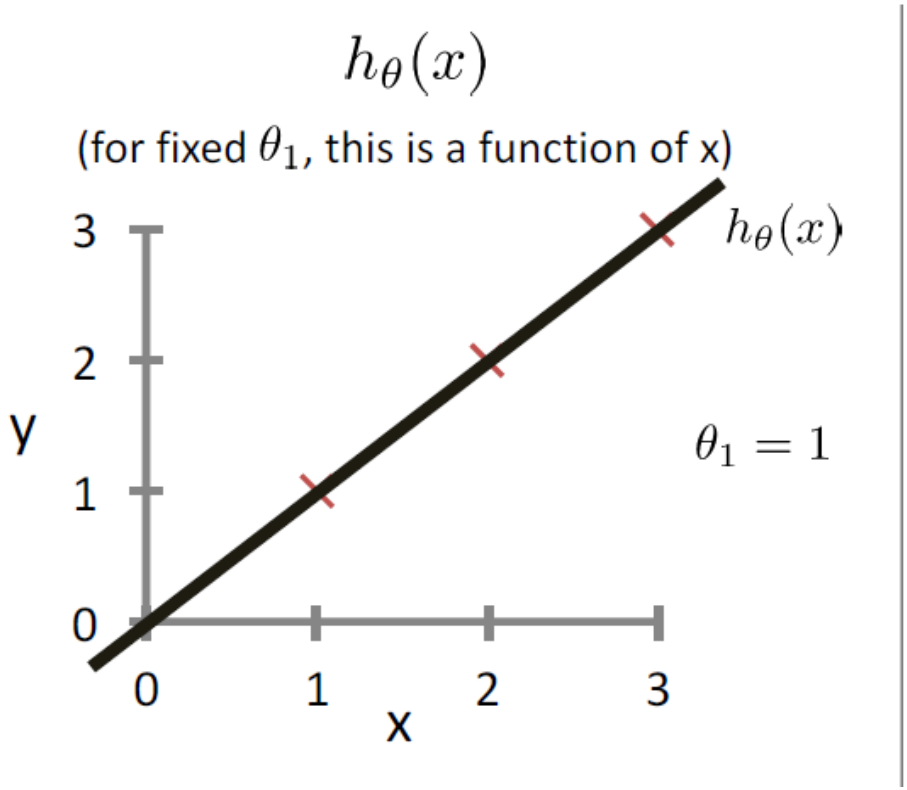
(for fixed θ_1, this is a function of x)



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 $\boldsymbol{\theta} = [\theta_0, \theta_1]$



Fix θ_0 =0

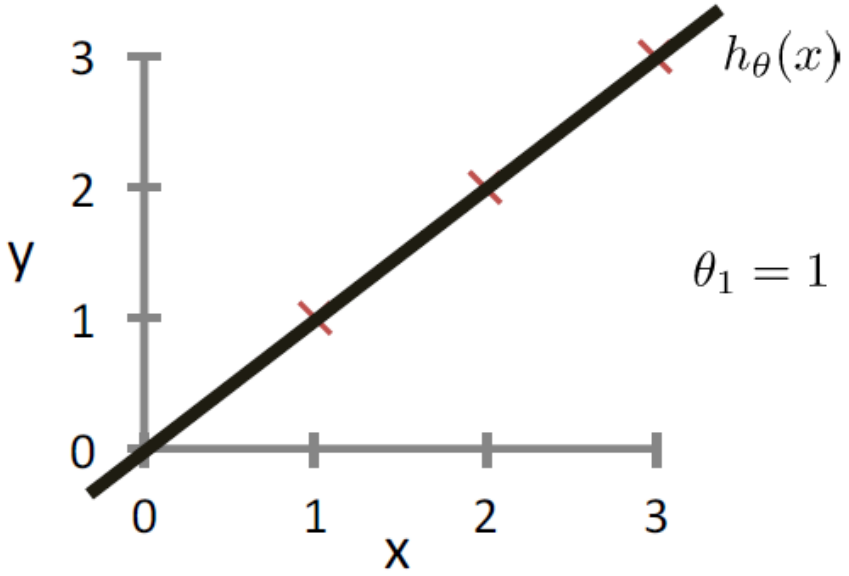
Intuition on MSE

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

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

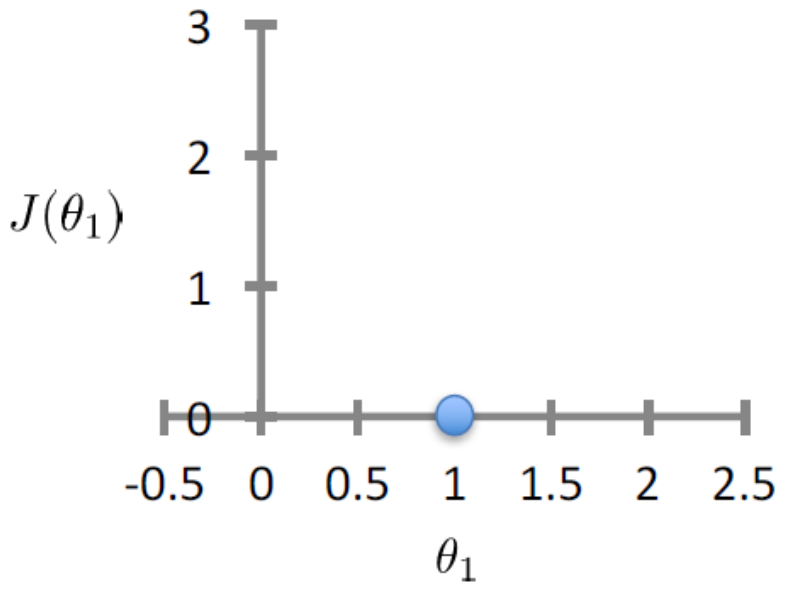
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Fix θ_0 =0

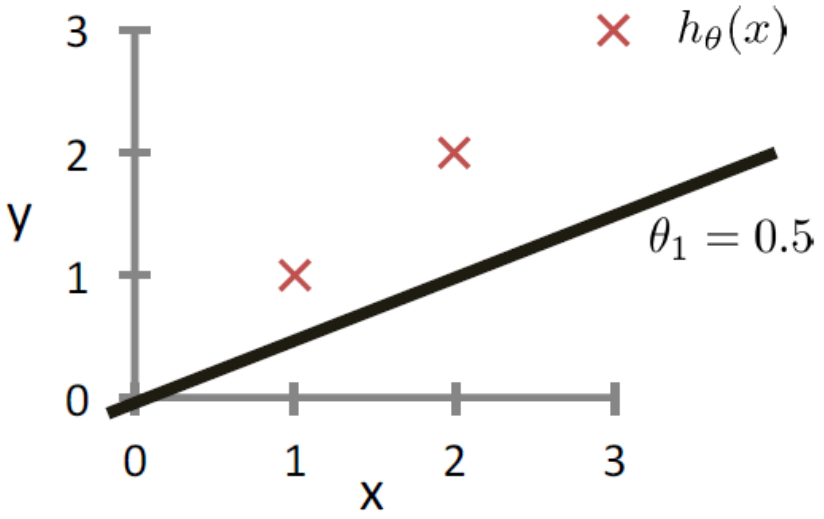
Intuition on MSE

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

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

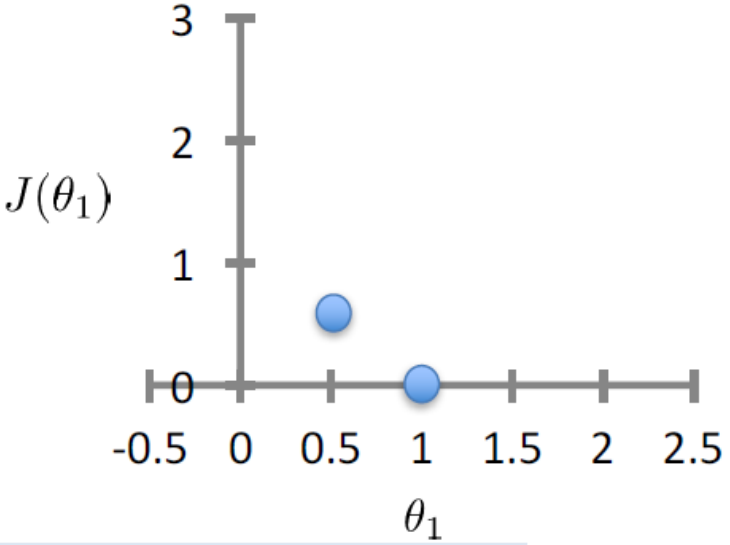
$h_\theta(x)$

(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

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

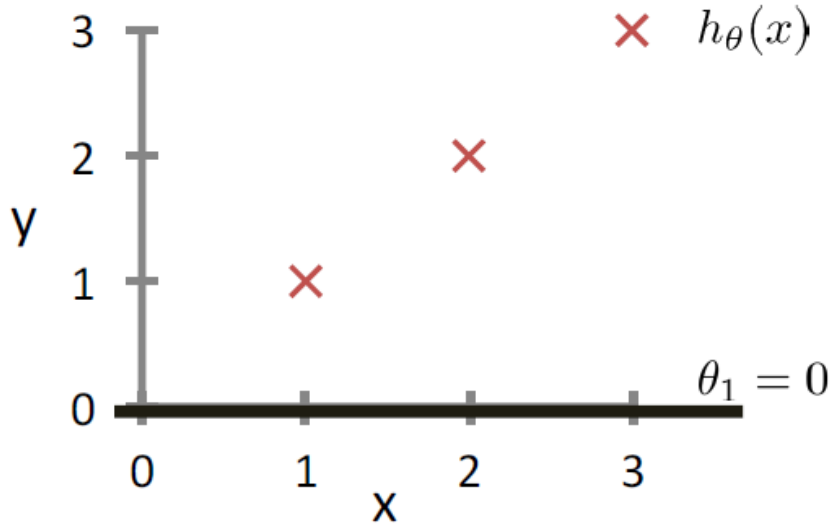
Intuition on MSE

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

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

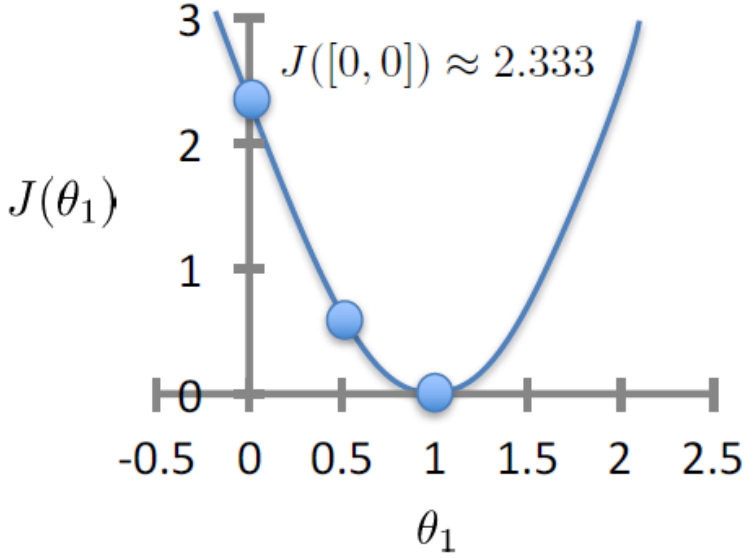
$h_\theta(x)$

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



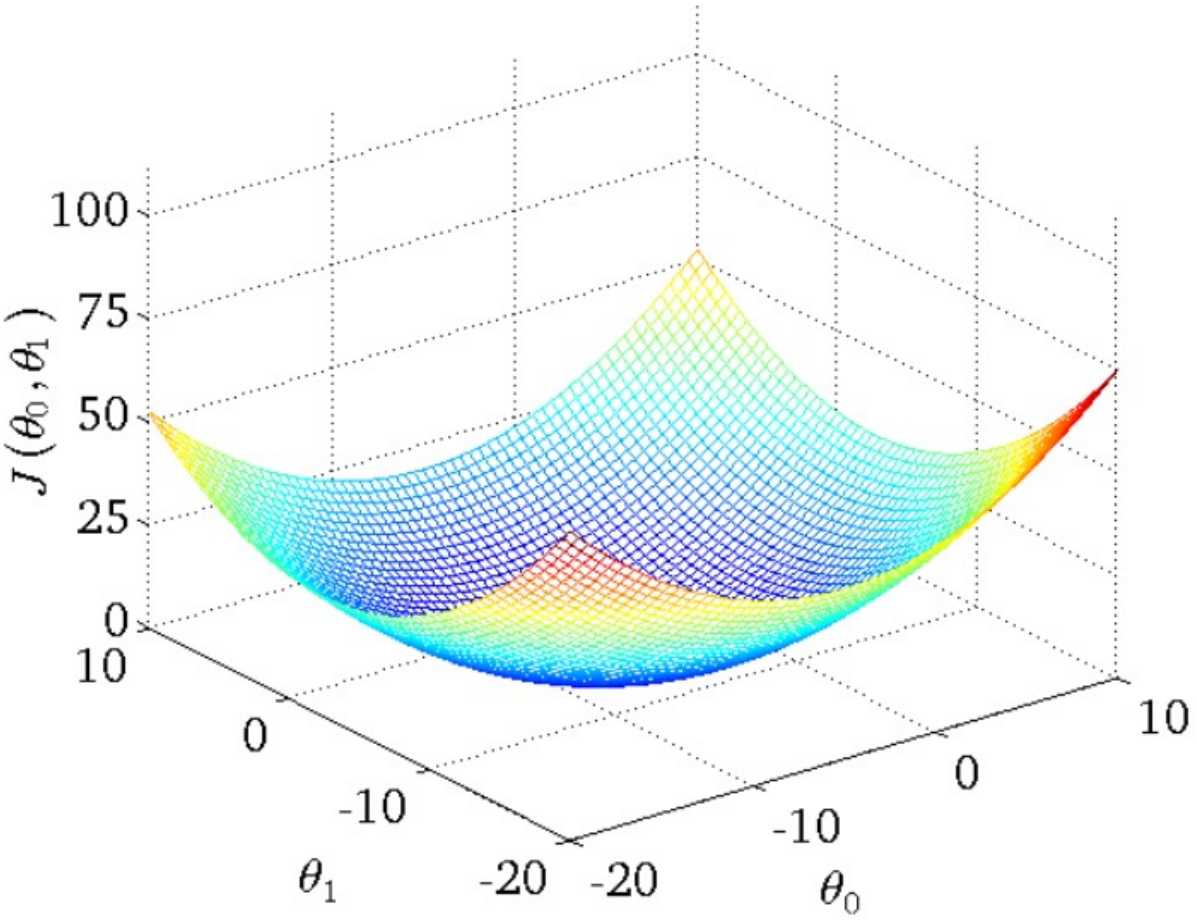
$J(\theta_1)$

(function of the parameter θ_1)



Based on example by Andrew Ng

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

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

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

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

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

Covariance

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

$$\text{(i)}\quad \mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$$

$$\text{(ii)}\quad \mathrm{Cov}(X,X) = \mathrm{Var}(\mathrm{X})$$

$$\text{(iii)}\quad \mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y)$$

$$\text{(iv)}\quad \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)$$

Covariance

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

Covariance

- X and Y are random variables
- Definition:
  - $Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$
- Can derive that:
  - $Cov(X,Y) = E[XY] - E[X]E[Y]$
- If X and Y are independent then:
  - $E[XY] = E[X]E[Y]$
  - $Cov(X,Y) = 0$

Pearson Correlation

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

Standard deviation

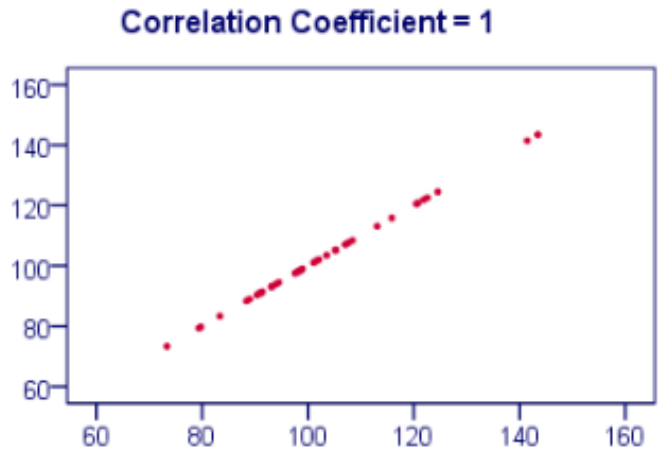
$$\sigma_X = \sqrt{Var(X)}$$

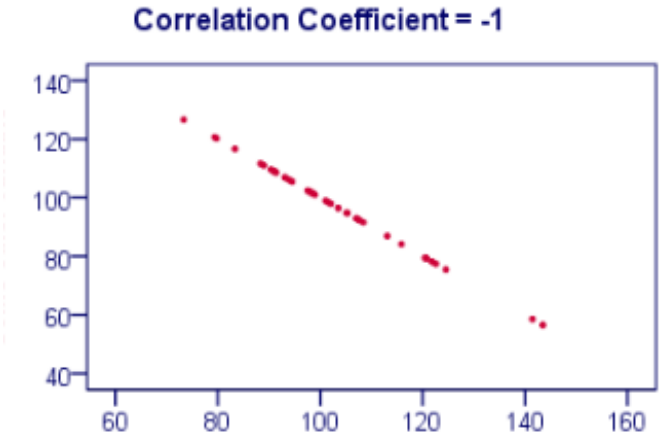
Pearson Correlation

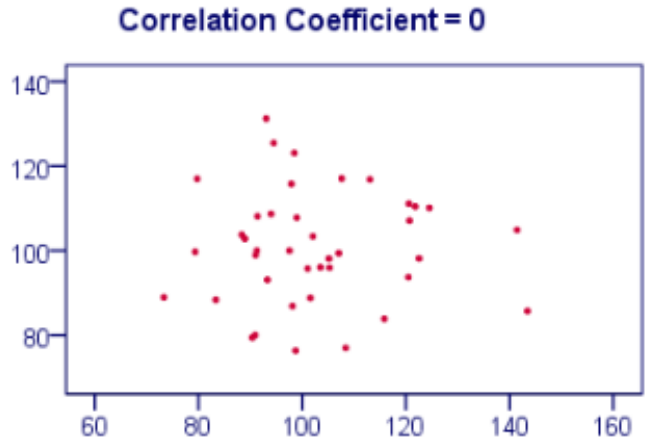
$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov(X, Y)}}{\sigma_X \sigma_Y} \in [-1,1]$$

Standard deviation

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

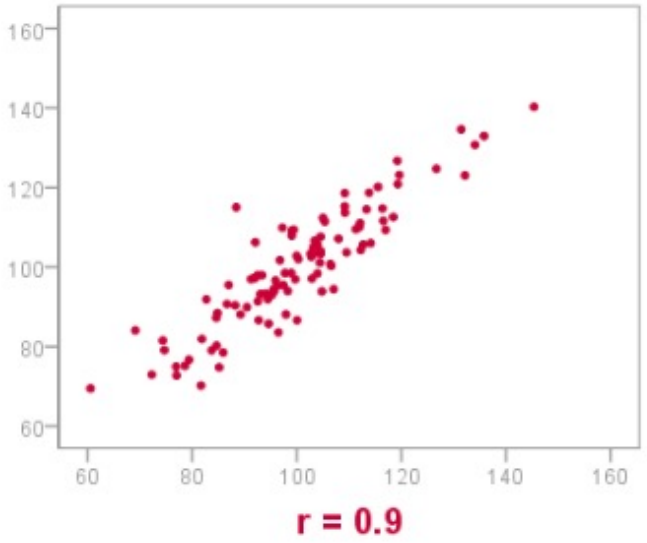


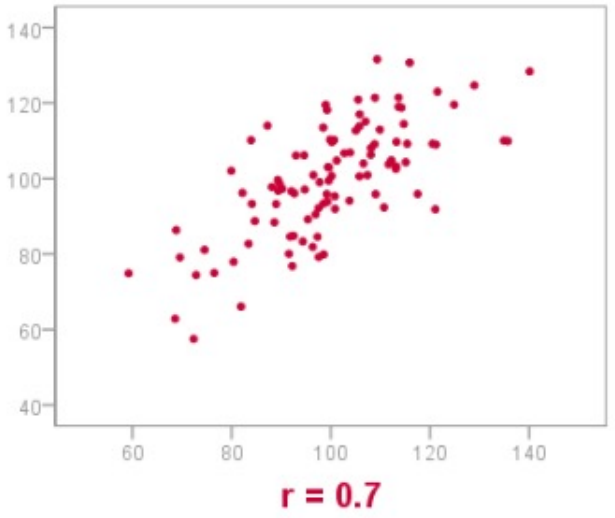




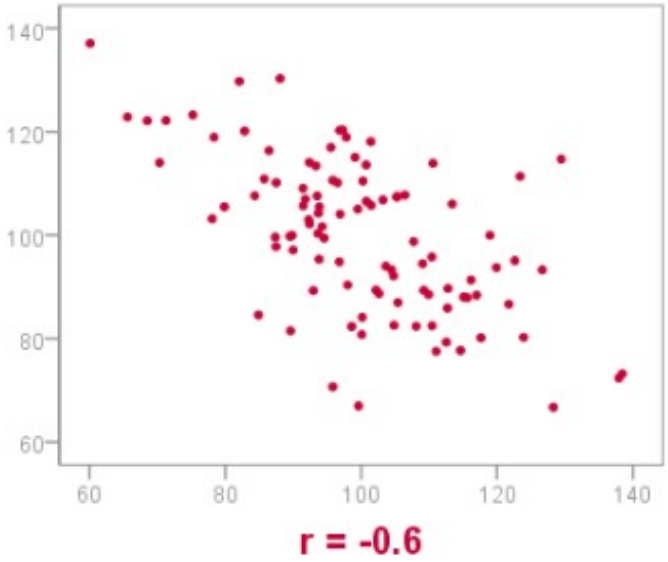
Positive/Negative Correlation

Positive Correlation





Negative Correlation



34

How Well Does the Model Fit?

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

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

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
  - $TSS = \sum[y_i - \bar{y}]^2$
  - Total variance of the response
- Proportion of variability in Y that can be explained using X
  - $R^2 = 1 - \frac{RSS}{TSS} \in [0,1]$

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

Recap Linear Regression

- Least Square Regression, OLS
- Simple linear regression: one dimension
- Multiple linear regression: multiple dimensions
- Minimize cost (loss) function
  - MSE: average of squared residuals
- Can derive closed-form solution for simple LR

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

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