

DS 4400

Machine Learning and Data Mining I Spring 2022

Alina Oprea

Associate Professor

Khoury College of Computer Science

Northeastern University

January 31 2022

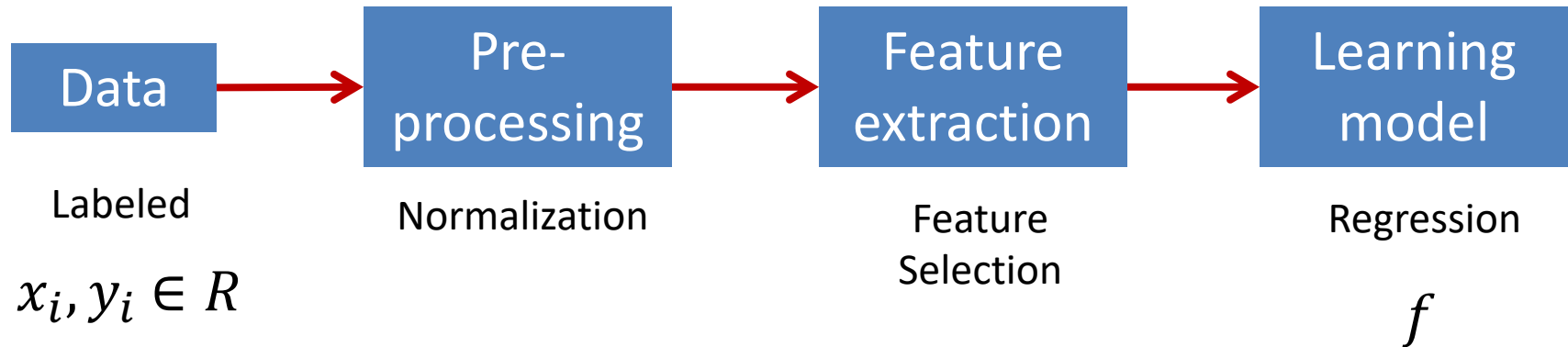
Today's Outline

- Announcements
 - HW 1 is due on Feb 7 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
 - Preferable 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

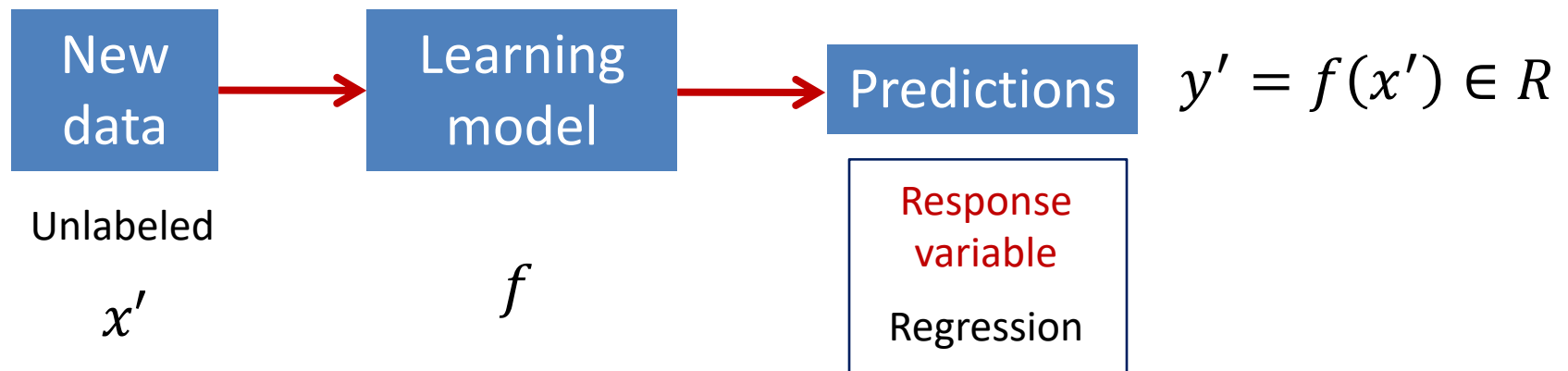
Linear regression

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)

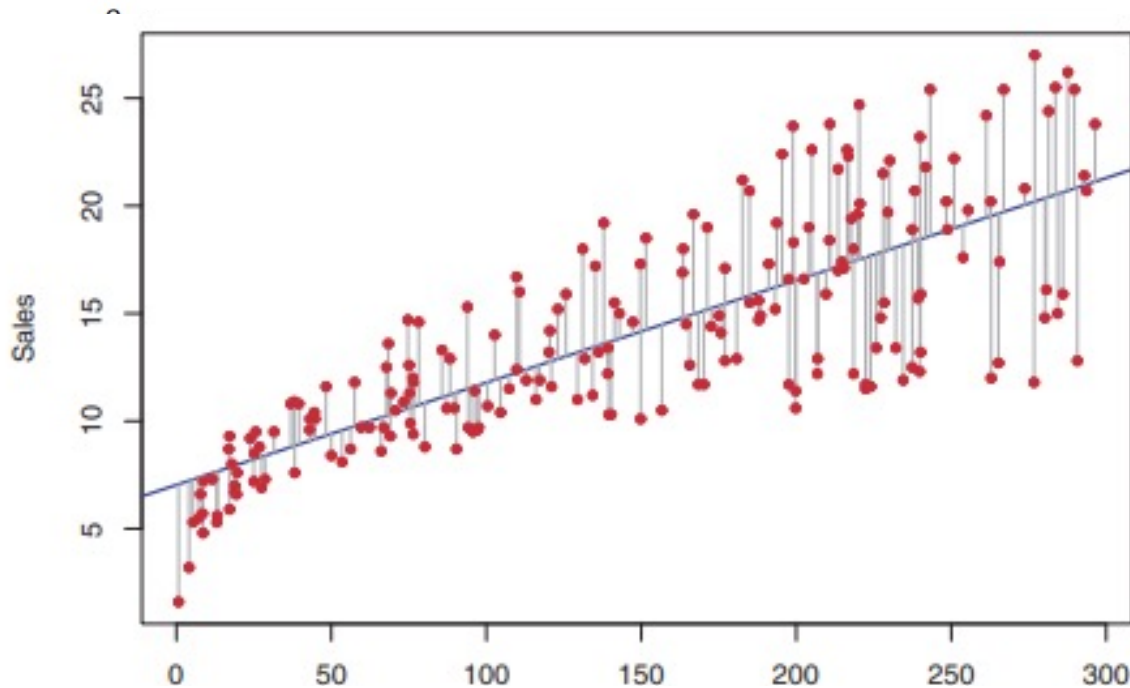
Linear regression

Given:

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

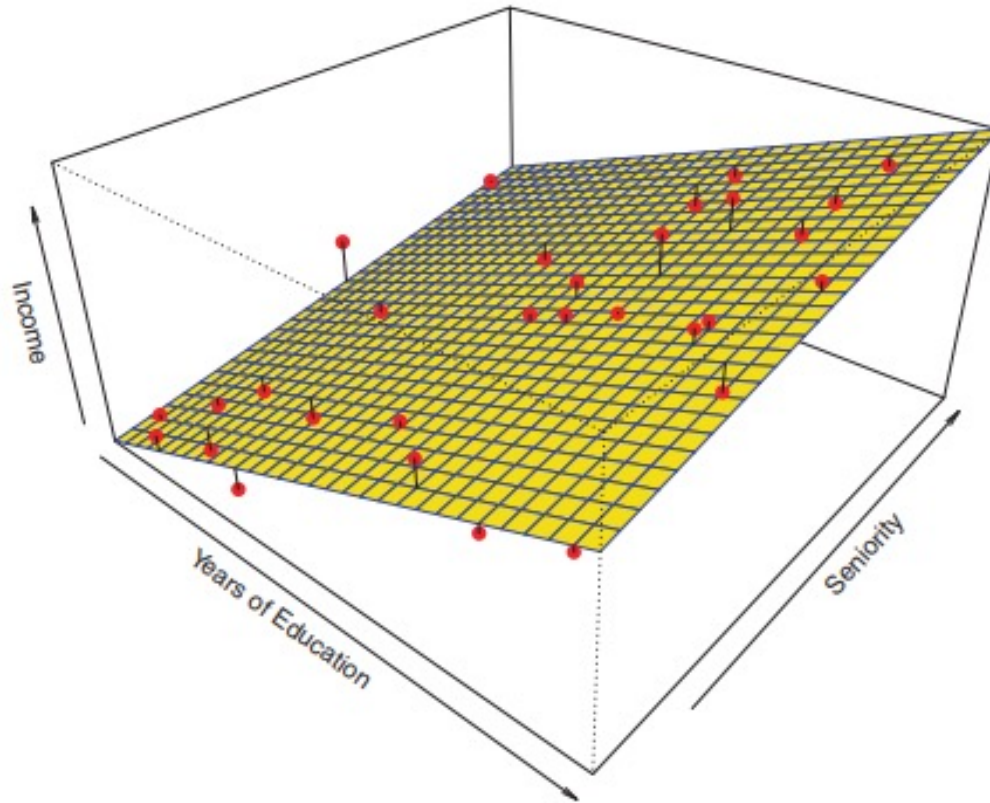
Features

Response
variables



Simple Linear Regression: 1 predictor

Income Prediction



Linear Regression with 2 predictors
Multiple Linear Regression

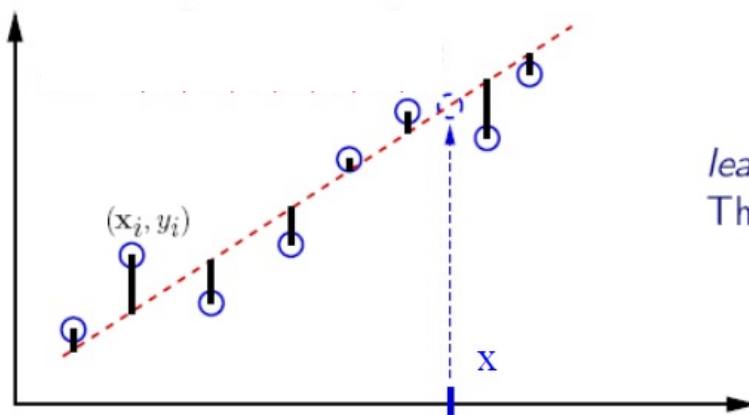
Hypothesis: linear model

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

Simple linear regression

Regression model is a line with 2 parameters: θ_0, θ_1

- Fit model by minimizing sum of squared errors



least squares (LSQ)
The fitted line is used as a predictor

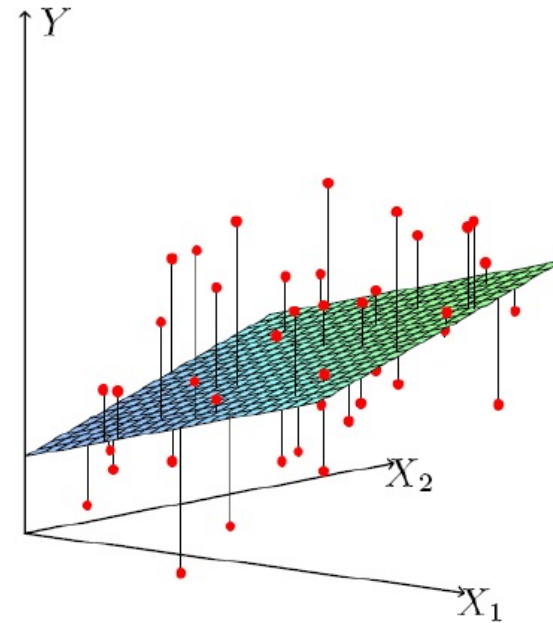
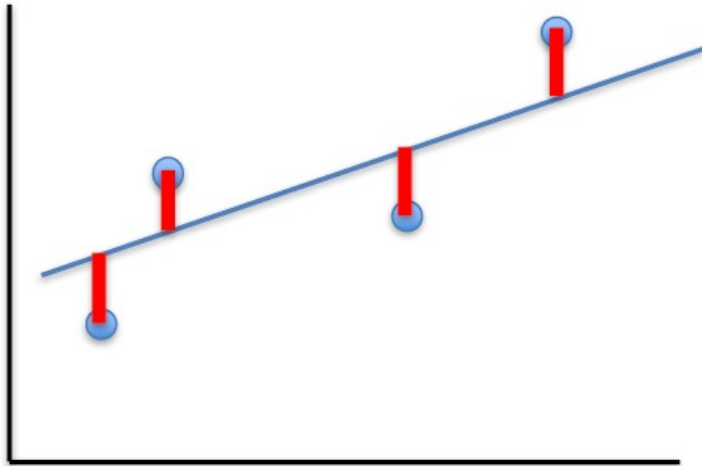
Least-Squares Linear Regression

- 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 $\min_{\theta} J(\theta)$



Terminology and Metrics

- **Residuals**

- Difference between predicted values and actual values

- Predicted value for example i is: $\hat{y}_i = h_{\theta}(x_i)$

- $R_i = |y_i - \hat{y}_i| = |y_i - (\theta_0 + \theta_1 x_i)|$

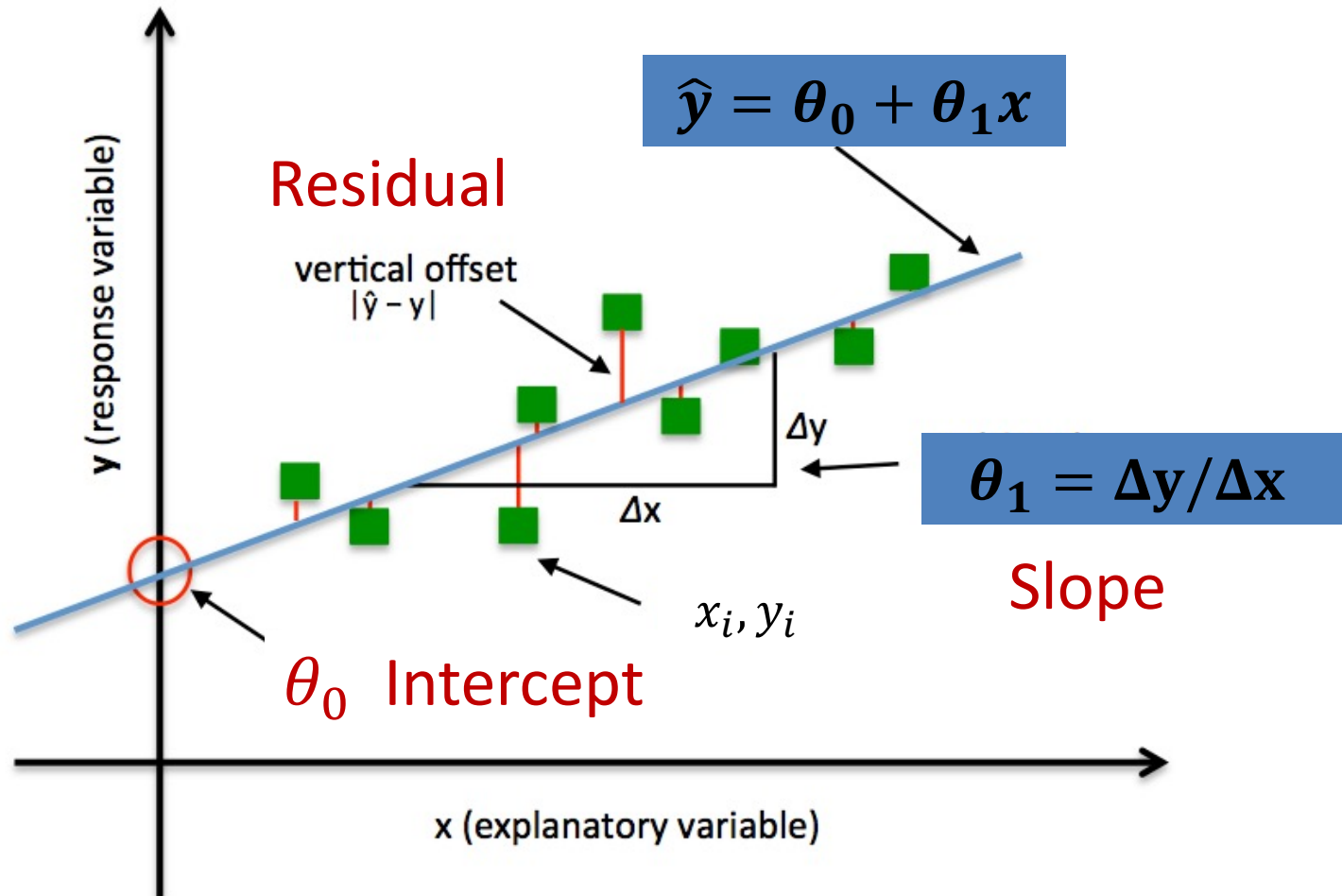
- **Residual Sum of Squares (RSS)**

- $RSS = \sum R_i^2 = \sum [y_i - (\theta_0 + \theta_1 x_i)]^2$

- **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



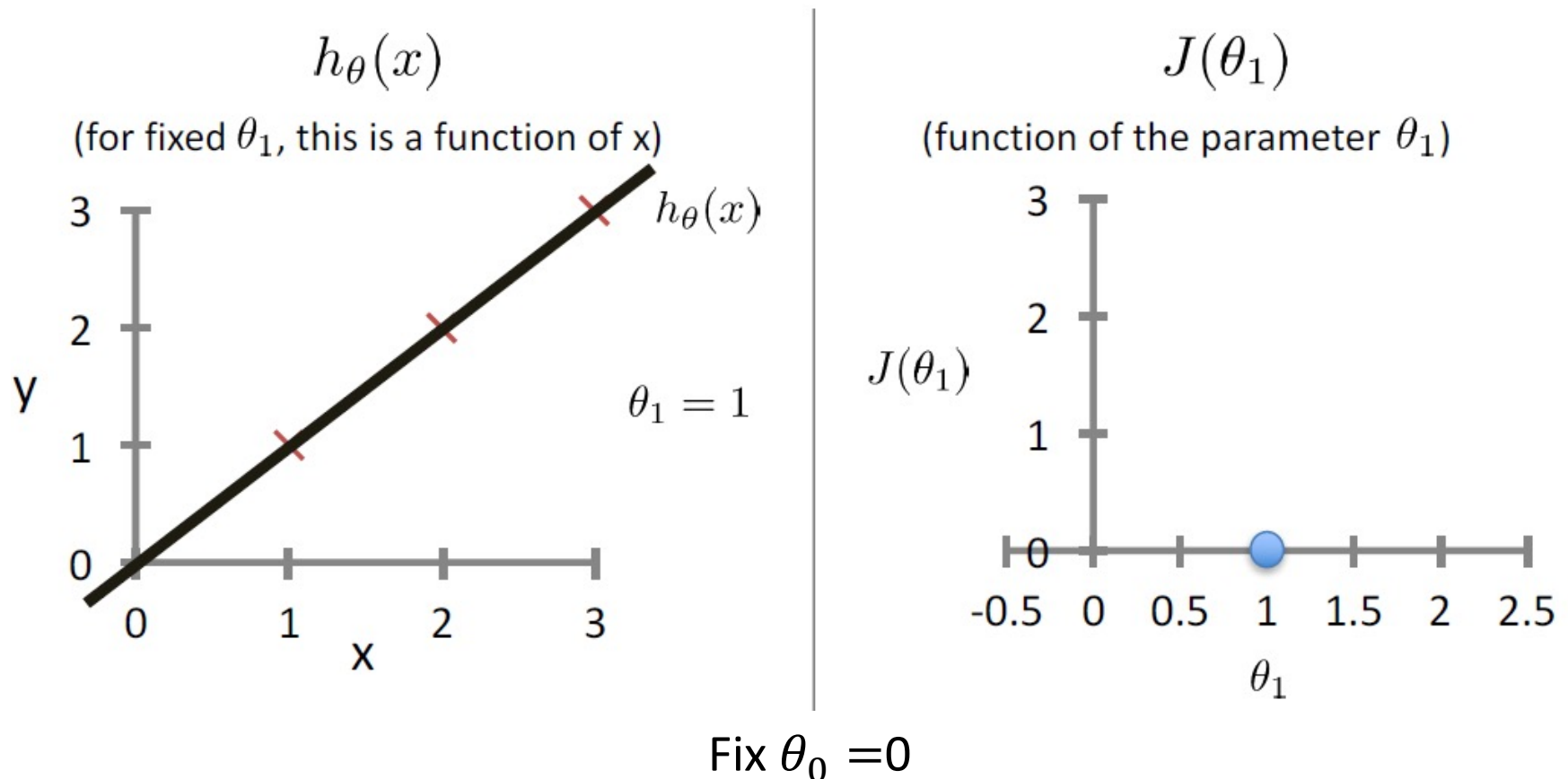
$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

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

Intuition on MSE

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

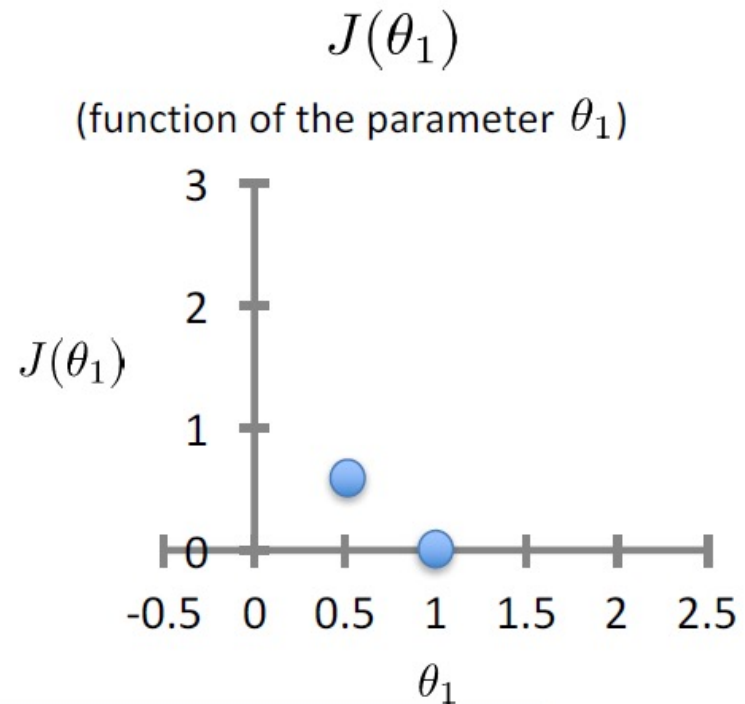
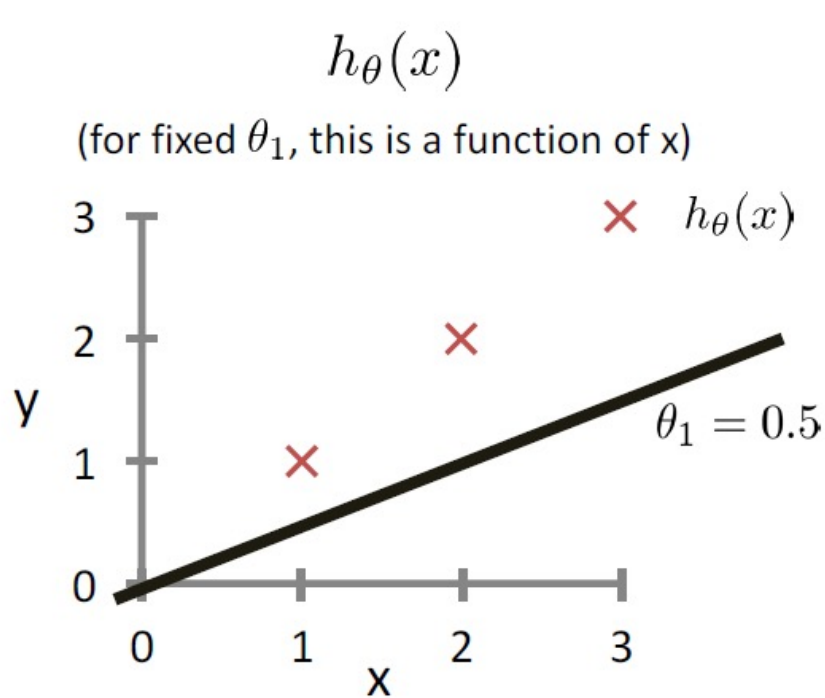
For insight on $J()$, let's assume $x \in \mathbb{R}$ so $\theta = [\theta_0, \theta_1]$



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 $\theta = [\theta_0, \theta_1]$



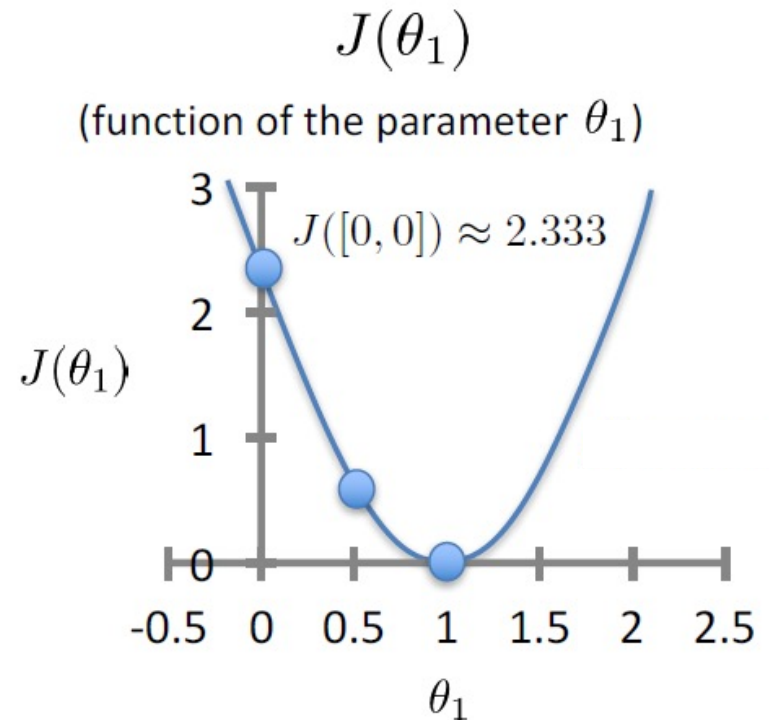
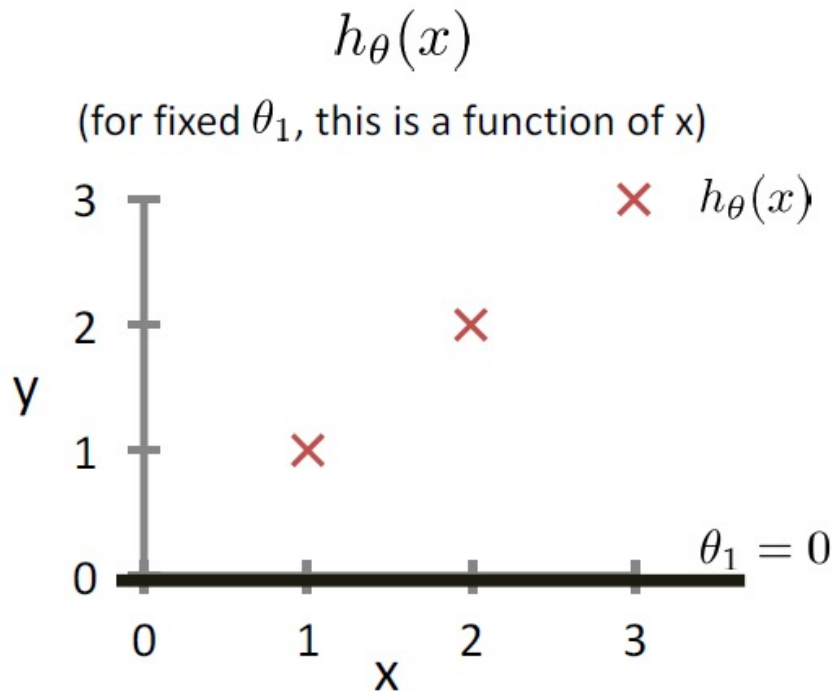
Based on example
by Andrew Ng

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

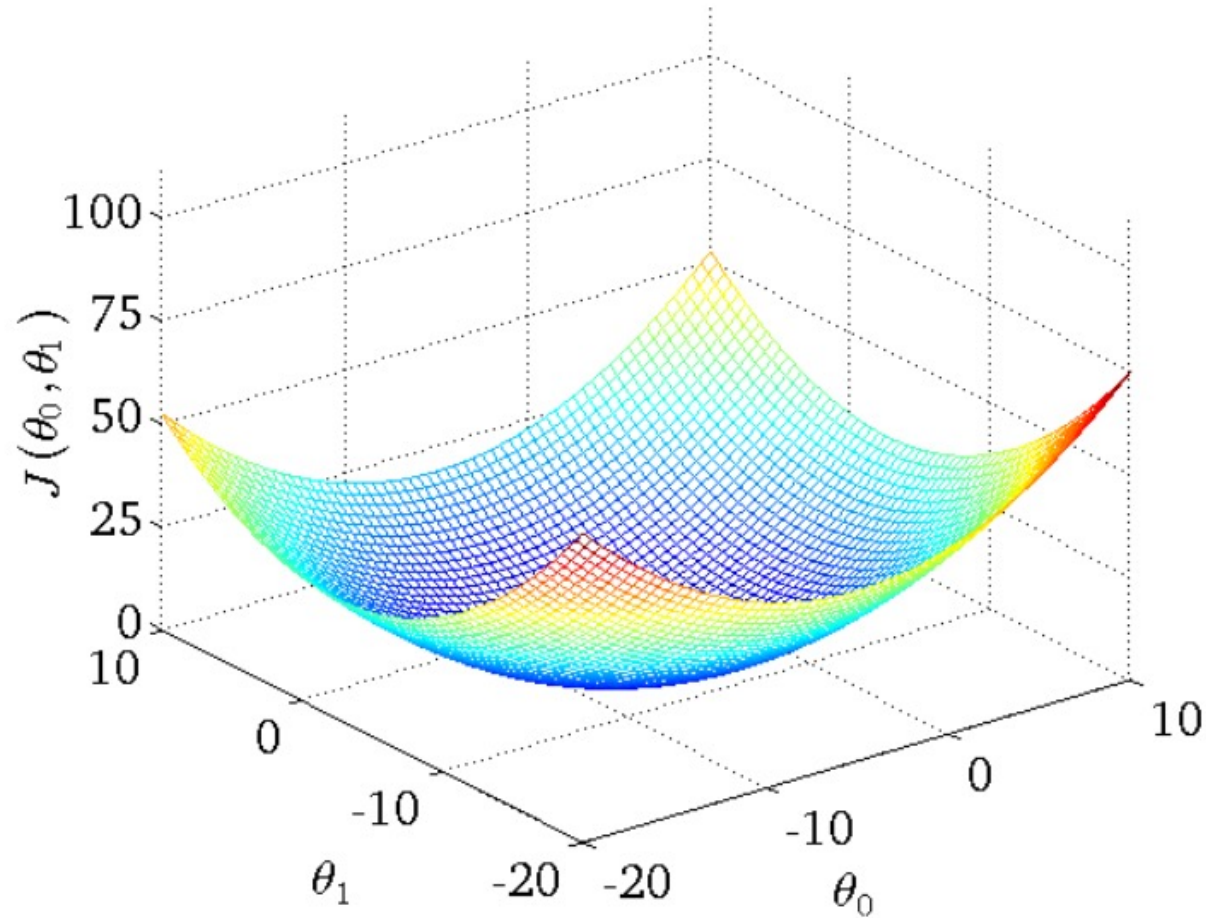
Intuition on MSE

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

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



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**

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

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

- Solution of min loss

$$\begin{aligned} -\theta_0 &= \bar{y} - \theta_1 \bar{x} \\ -\theta_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{aligned}$$

$$\begin{aligned} \bar{x} &= \frac{\sum_{i=1}^N x_i}{N} \\ \bar{y} &= \frac{\sum_{i=1}^N y_i}{N} \end{aligned}$$

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

$$(i) \quad Cov(X, Y) = Cov(Y, X)$$

$$(ii) \quad Cov(X, X) = Var(X)$$

$$(iii) \quad Cov(aX, Y) = a Cov(X, Y)$$

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

Covariance

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

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

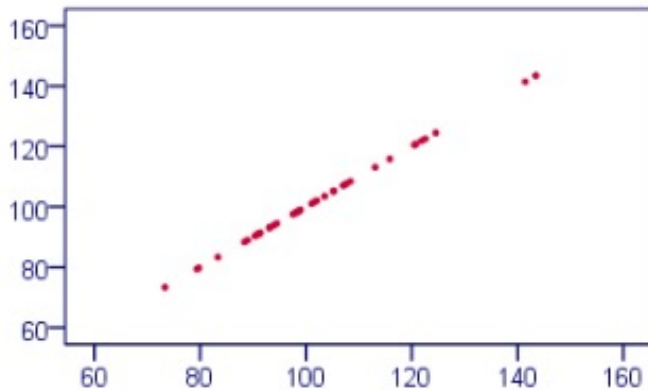
Standard deviation
 $\sigma_X = \sqrt{\text{Var}(X)}$

Pearson Correlation

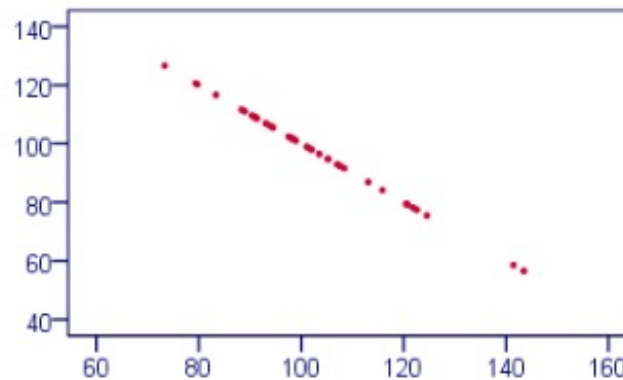
$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \in [-1, 1]$$

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

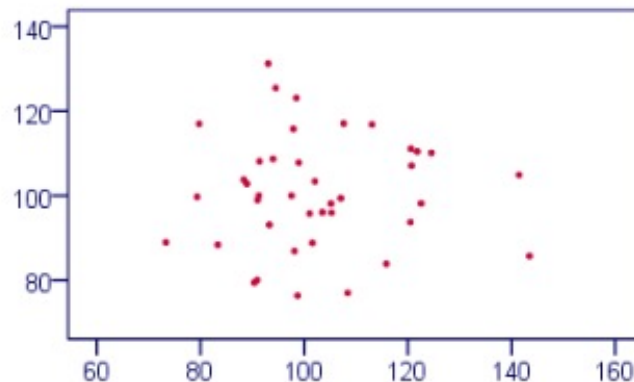
Correlation Coefficient = 1



Correlation Coefficient = -1

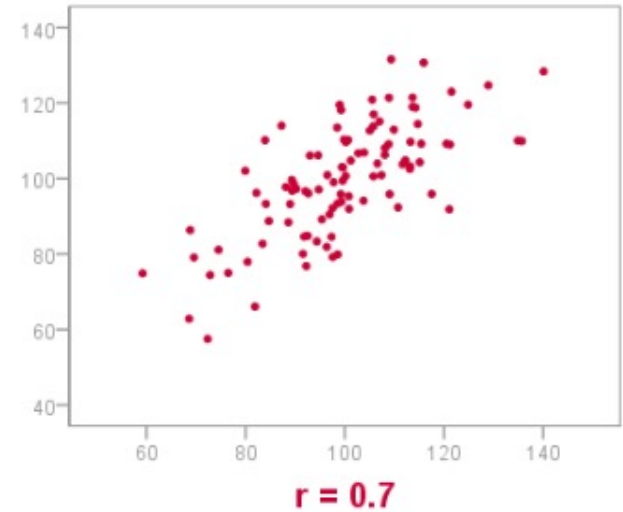
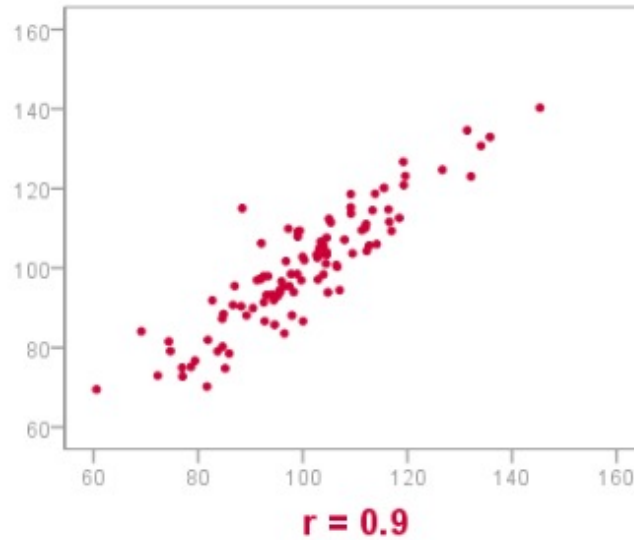


Correlation Coefficient = 0

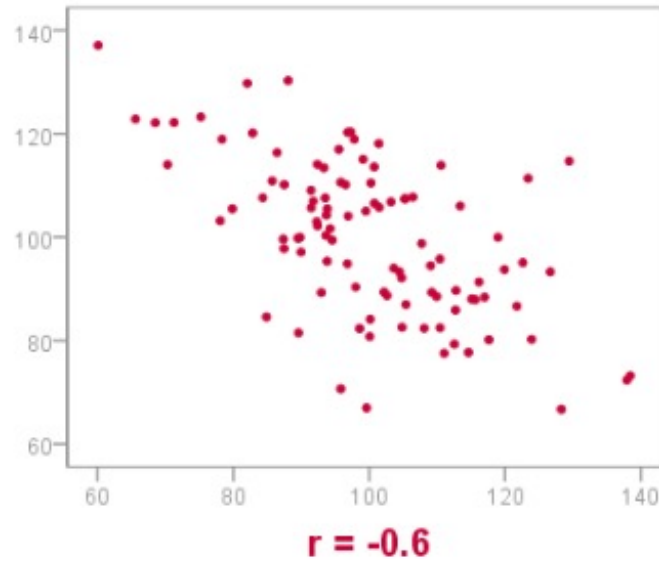


Positive/Negative Correlation

Positive
Correlation



Negative
Correlation



How Well Does the Model Fit?

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

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

- 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
- $\theta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}[X]}$
- If $\sigma_X = \sigma_Y$, then $\theta_1 = \text{Corr}(X, Y)$

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

- $TSS = \sum [y_i - \bar{y}]^2$

- Total variance of the response

- Proportion of variability in Y that can be explained using X

- $R^2 = 1 - \frac{RSS}{TSS} \in [0,1]$

- Correlation between feature and response

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

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

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

$$\begin{aligned} -\theta_0 &= \bar{y} - \theta_1 \bar{x} \\ -\theta_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \end{aligned}$$

Acknowledgements

- Slides made using resources from:
 - Andrew Ng
 - Eric Eaton
 - David Sontag
- Thanks!