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 $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

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

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 $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linear independence

- A set of vectors is linearly independent if none of them can be written as a linear combination of the others.
- Vectors $v_1,\ldots,v_k$ are linearly independent if $c_1v_1+\ldots+c_kv_k = 0$ implies $c_1=\ldots=c_k=0$
- Otherwise they are linearly dependent

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

(c_1, c_2)=(0,0), i.e. the columns are linearly independent.

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

Linearly dependent

$$x_3 = -2x_1 + x_2$$

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)

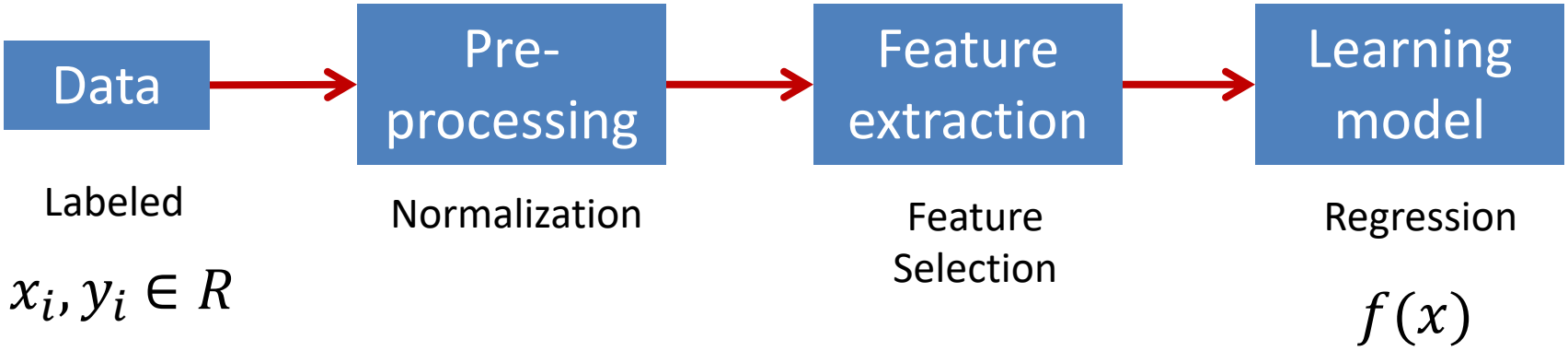
- Examples

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 \\ 4 & 2 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 3 \\ 0 & 5 & 2 \end{pmatrix}$$

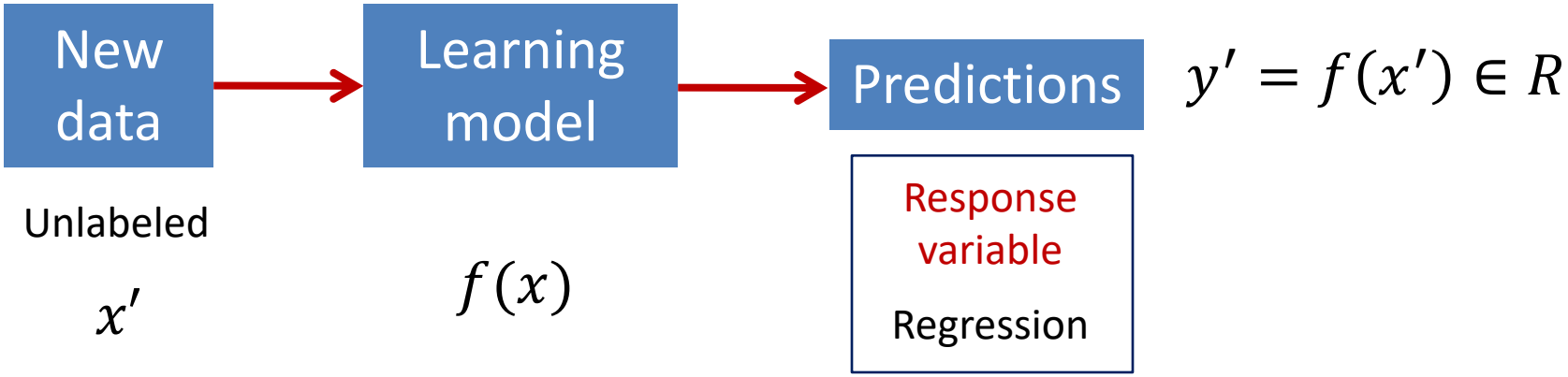
Linear regression

Supervised Learning: Regression

Training



Testing



Steps to Learning Process

- Define problem space
- Collect data
- Extract feature
- Pick a model (hypothesis)
- 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

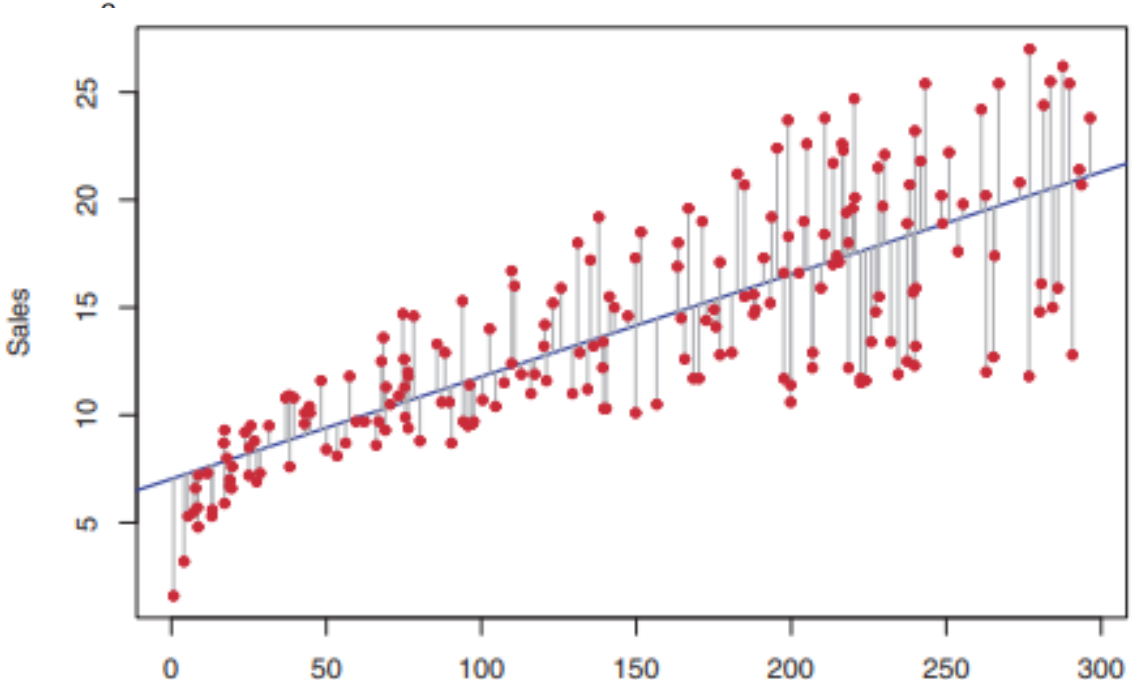
Linear regression

- 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

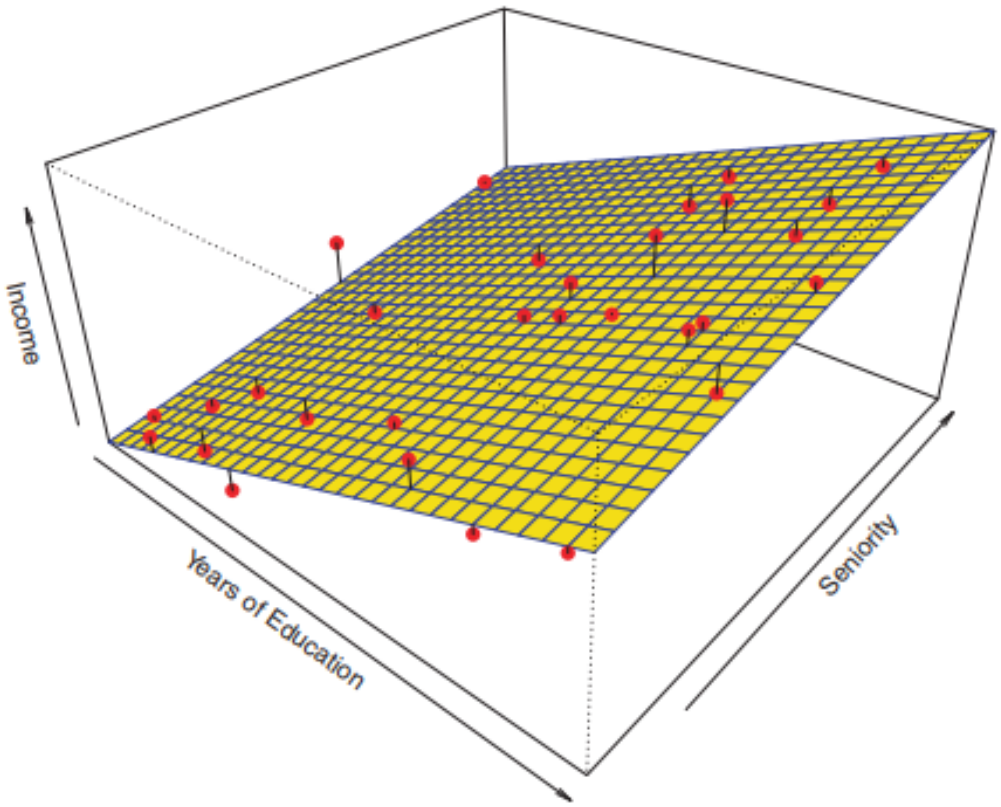
Given:

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



Simple Linear Regression: 1 predictor

Income Prediction



Linear Regression with 2 predictors

Multiple Linear Regression

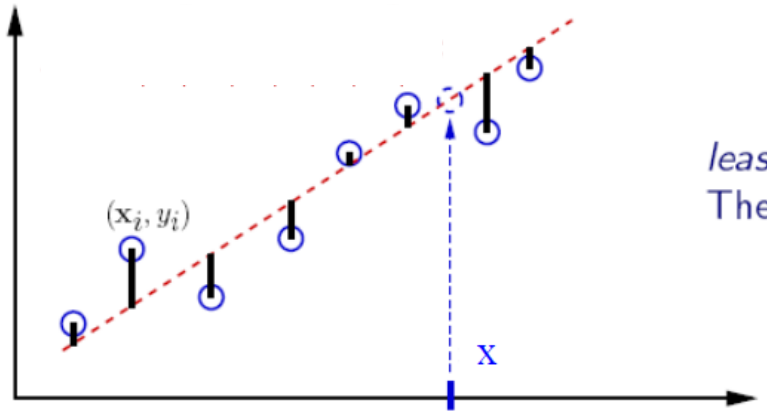
Hypothesis: linear model

- 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

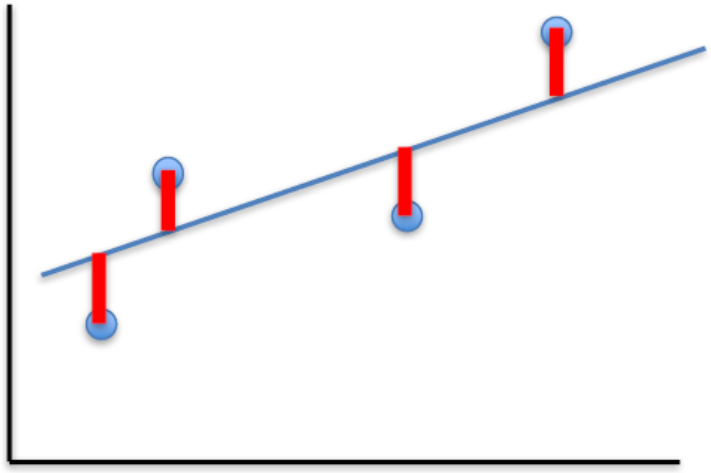
Least-Squares Linear Regression

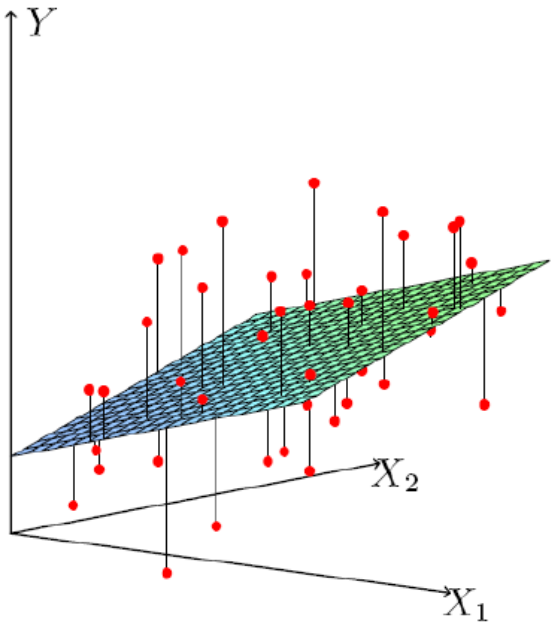
- Cost Function

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

Mean Square Error (MSE)

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

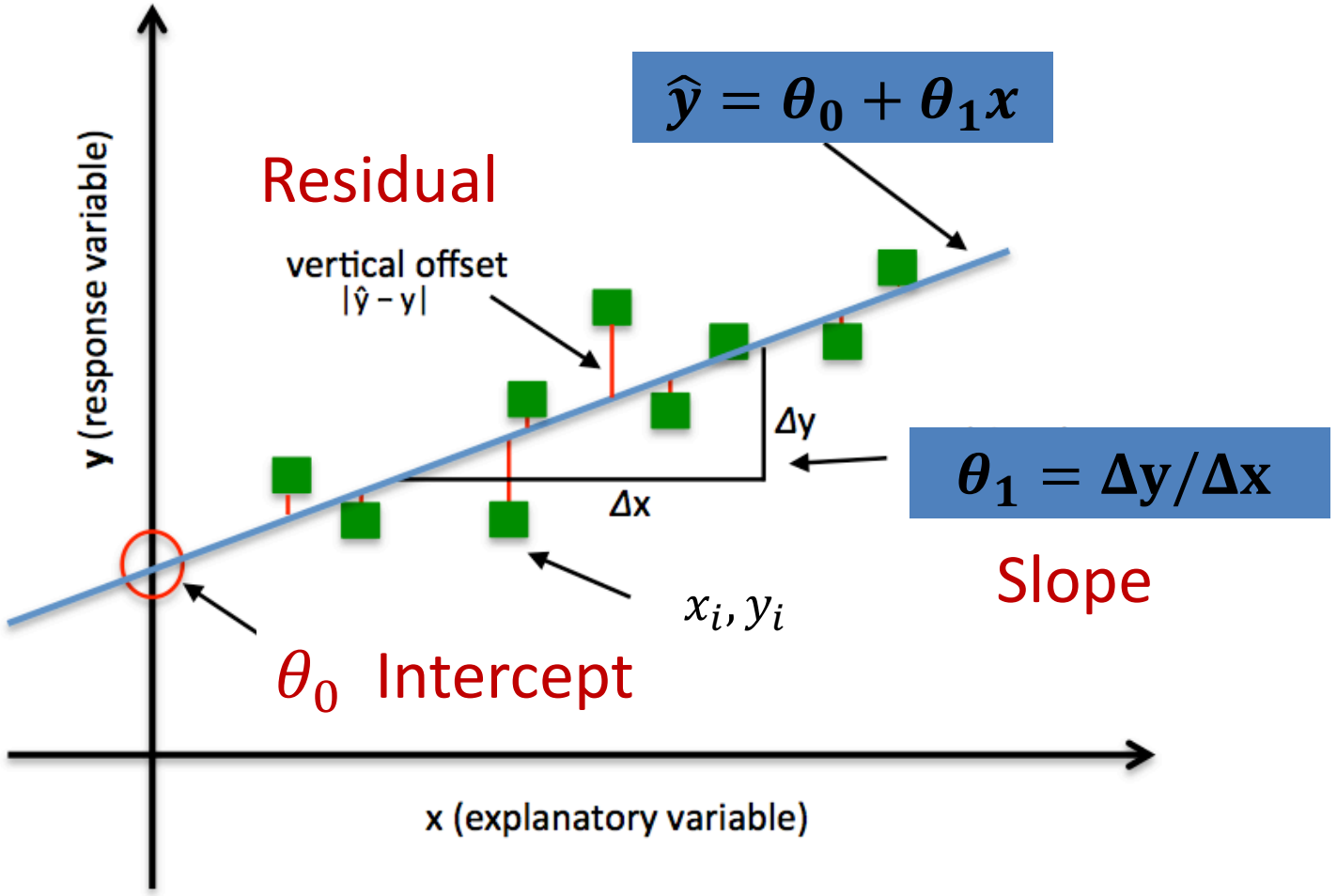




Terminology and Metrics

- 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



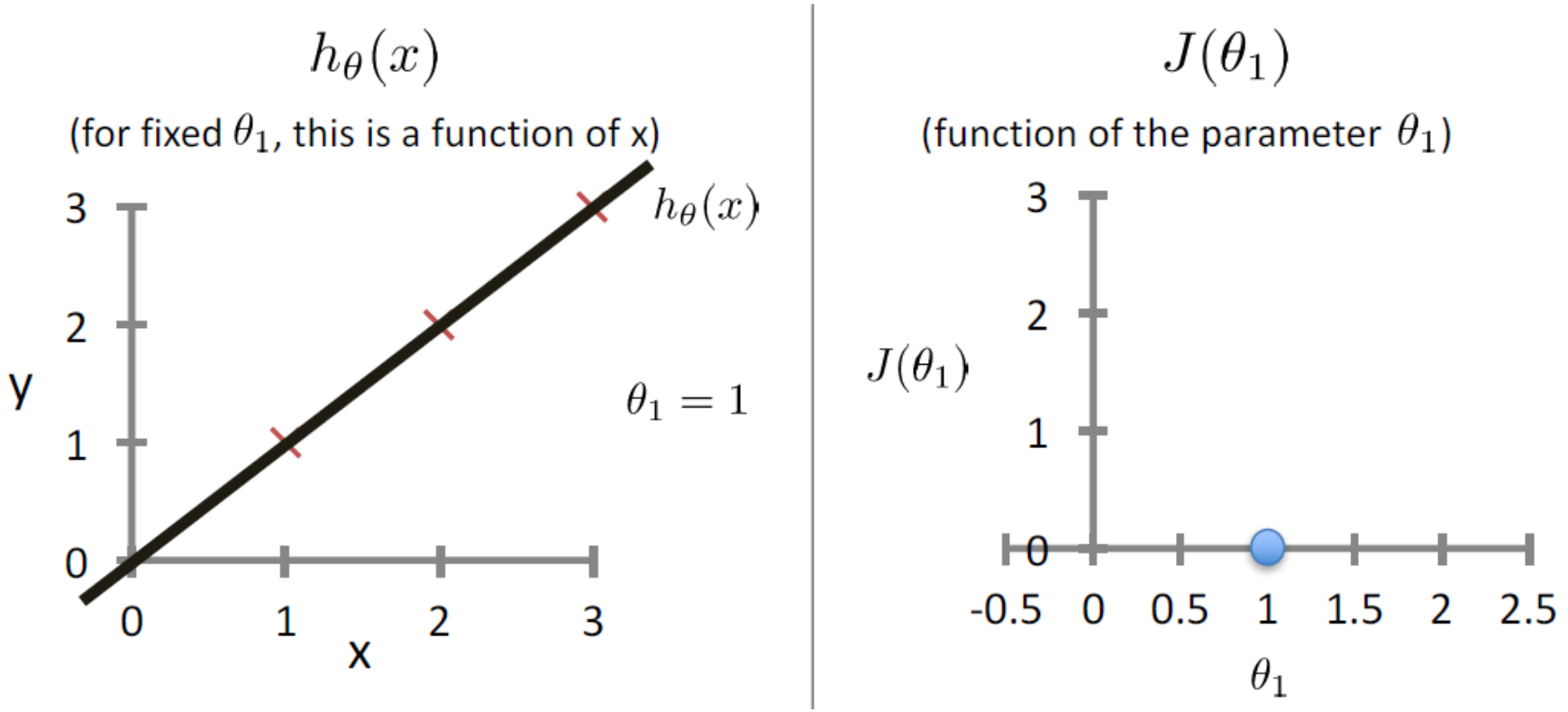
$$h_\theta(x) = \theta_0 + \theta_1 x$$

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

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



Fix θ_0 =0

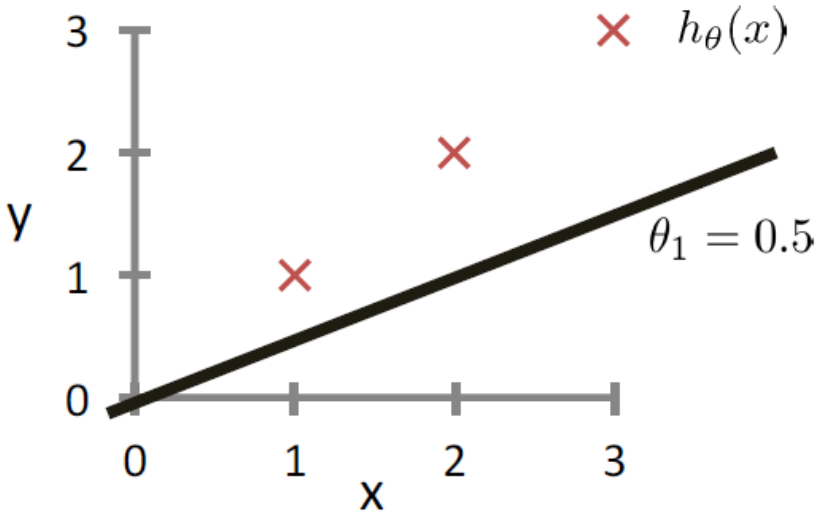
Intuition on MSE

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

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

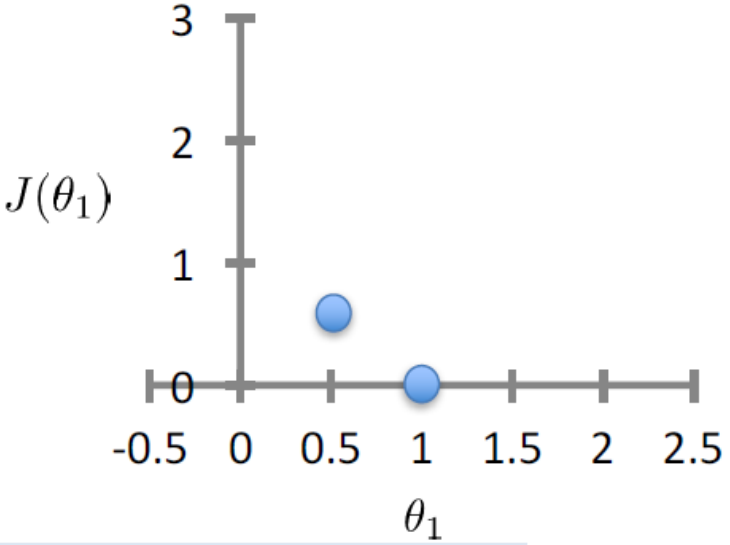
$h_\theta(x)$

(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

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

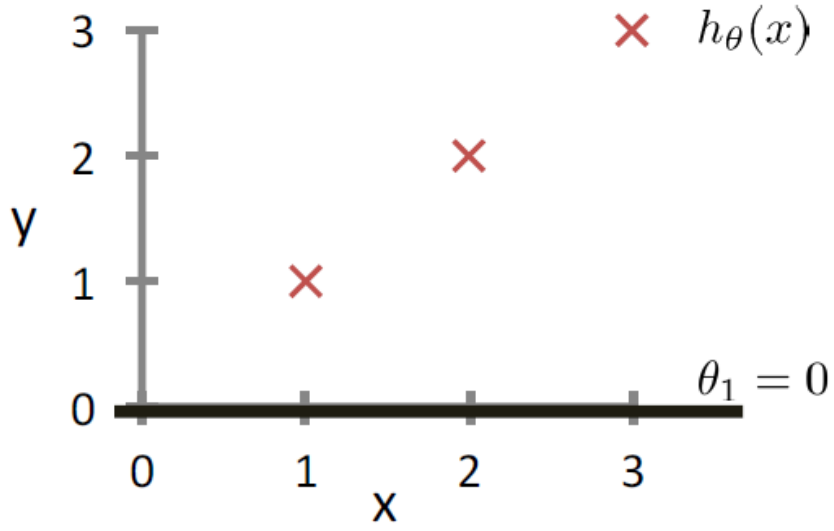
Intuition on MSE

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

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

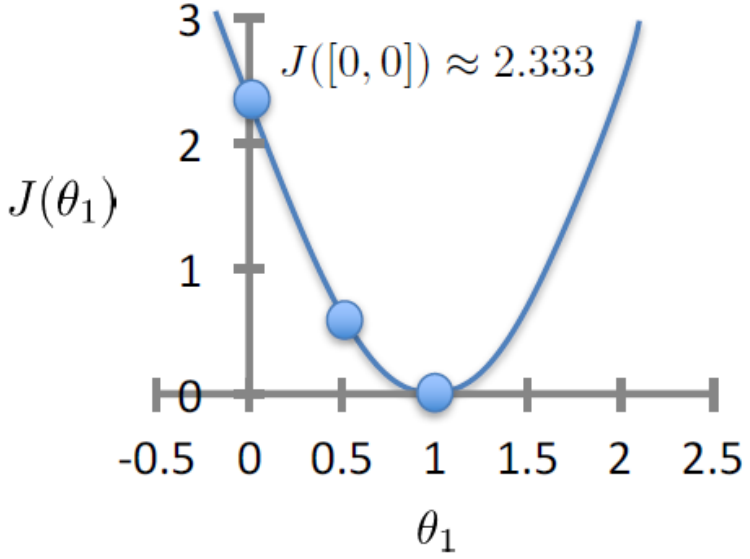
$h_\theta(x)$

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



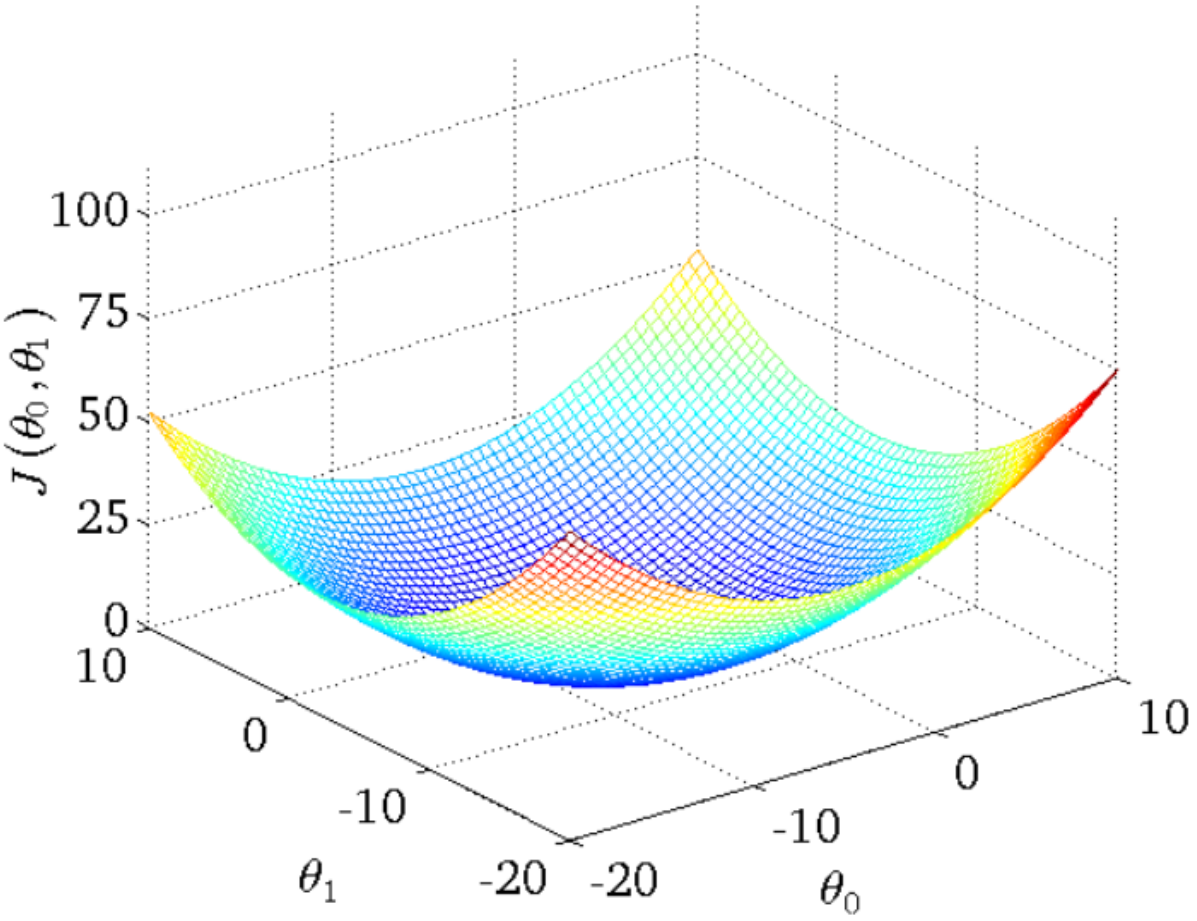
$J(\theta_1)$

(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

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

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

- 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

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

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

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

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

Covariance

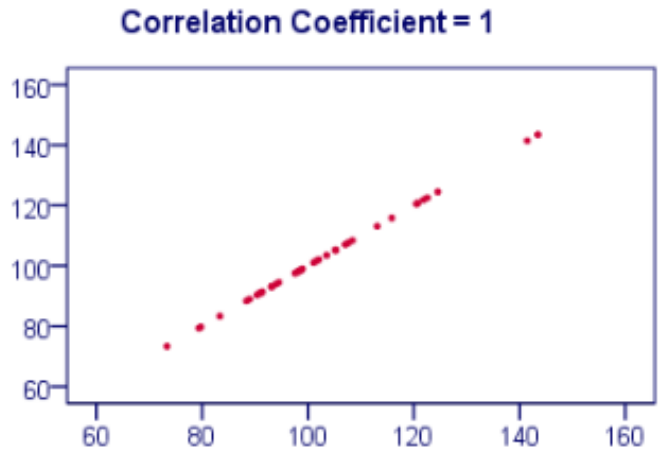
- X and Y are random variables
- Definition:
  - $Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$
- Can derive that:
  - $Cov(X,Y) = E[XY] - E[X]E[Y]$
- If X and Y are independent then:
  - $E[XY] = E[X]E[Y]$
  - $Cov(X,Y) = 0$

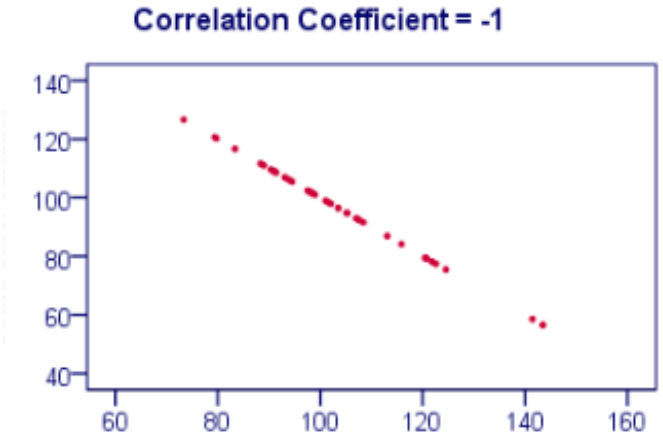
Pearson Correlation

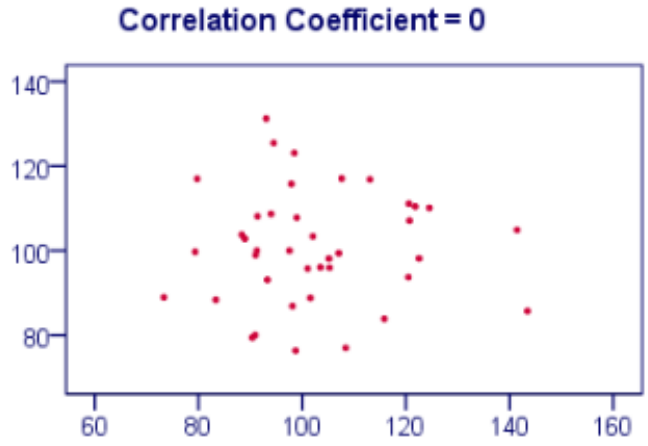
$$\rho = \text{Corr}(X, Y) = \frac{\text{Cov(X, Y)}}{\sigma_X \sigma_Y} \in [-1,1]$$

Standard deviation

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

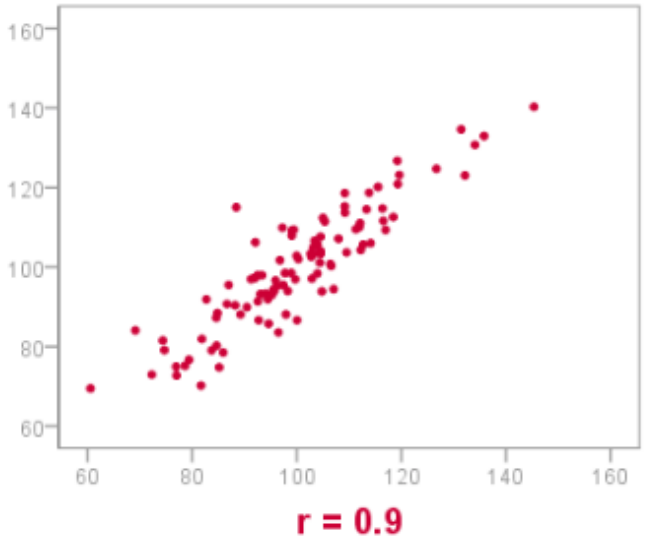


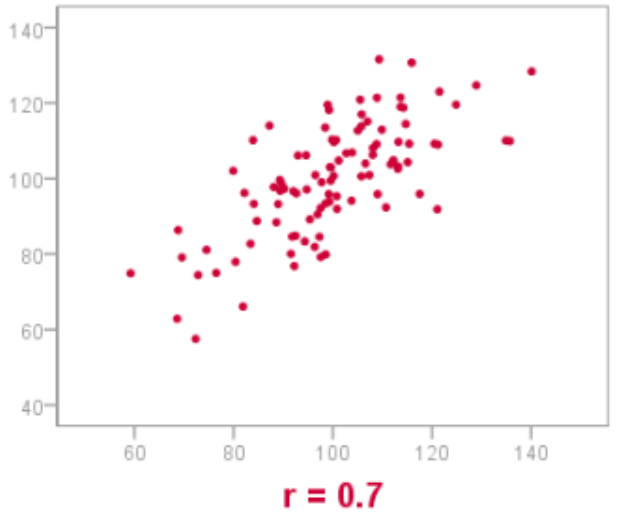




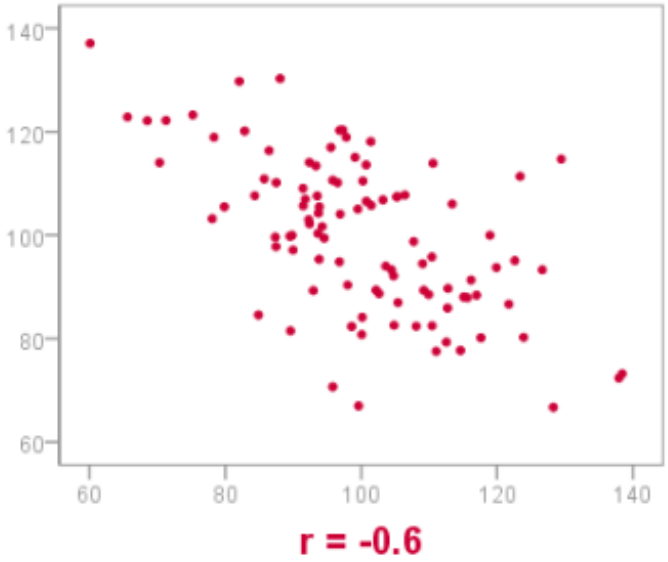
Positive/Negative Correlation

Positive Correlation





Negative Correlation



26

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

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

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