DS 4400

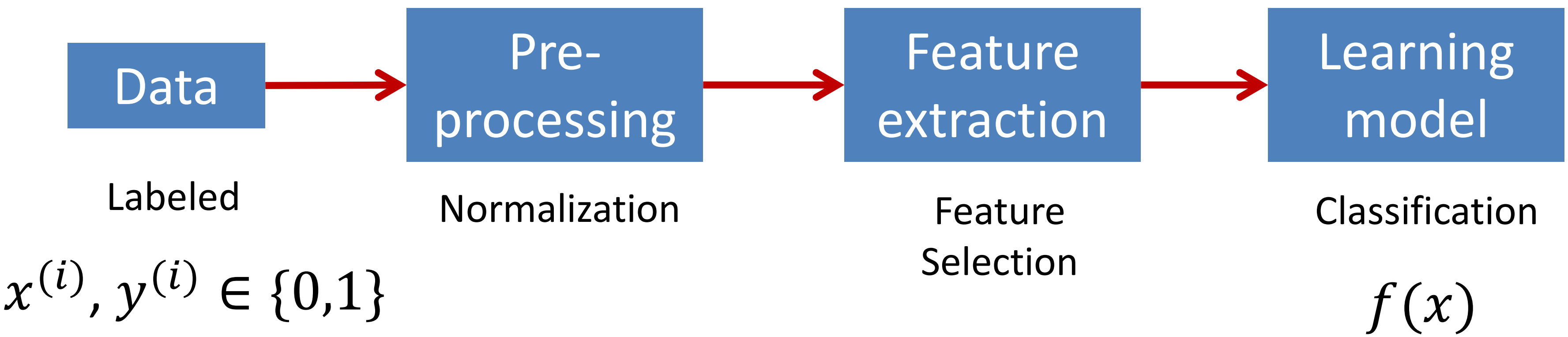
Machine Learning and Data Mining I

Alina Oprea
Associate Professor, CCIS
Northeastern University

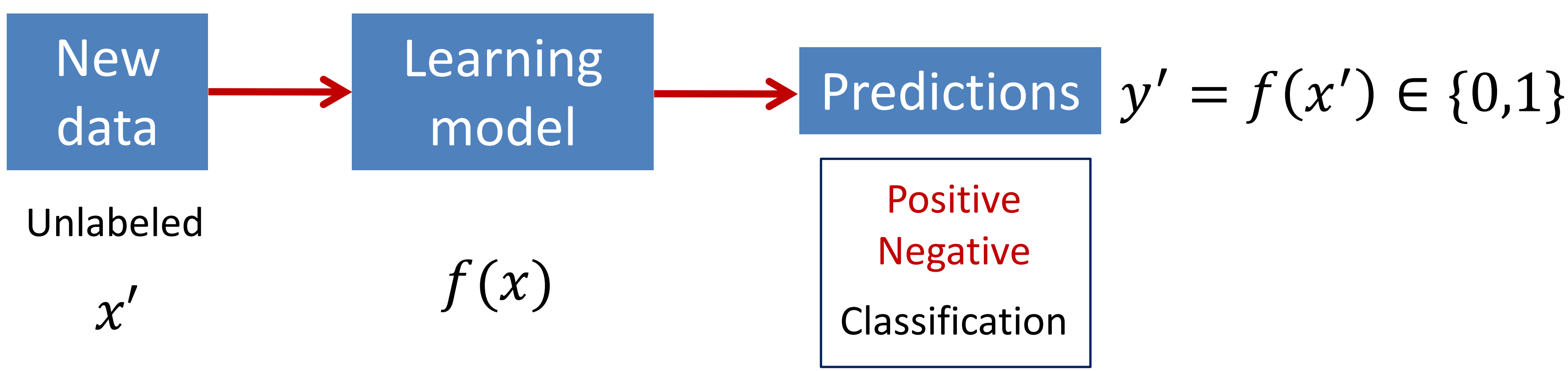
January 15 2019

Supervised Learning: Classification

Training

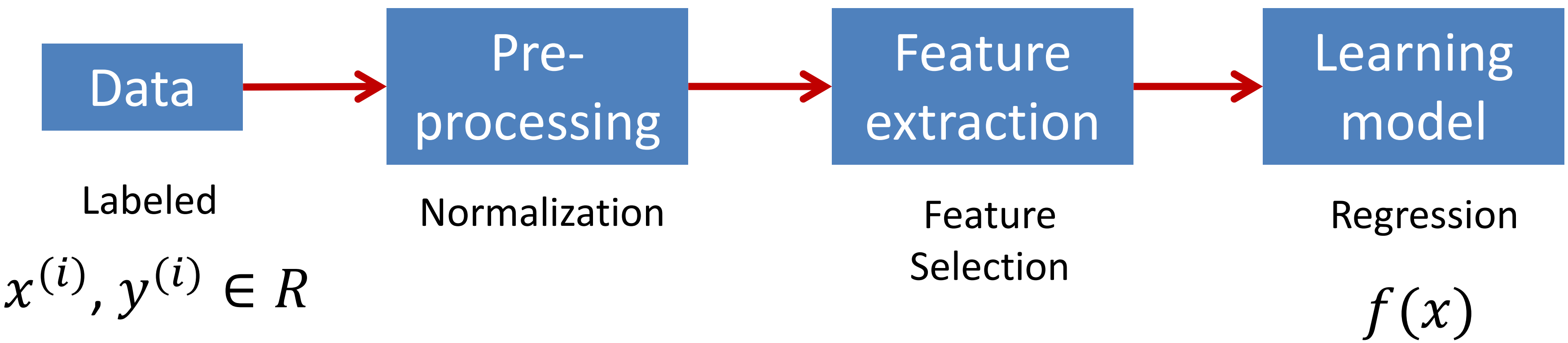


Testing

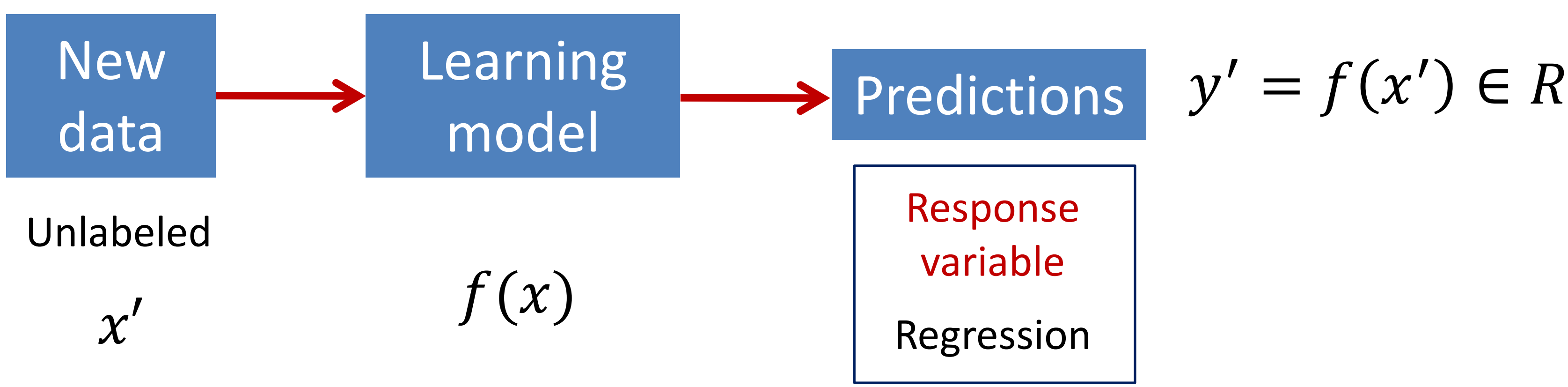


Supervised Learning: Regression

Training



Testing

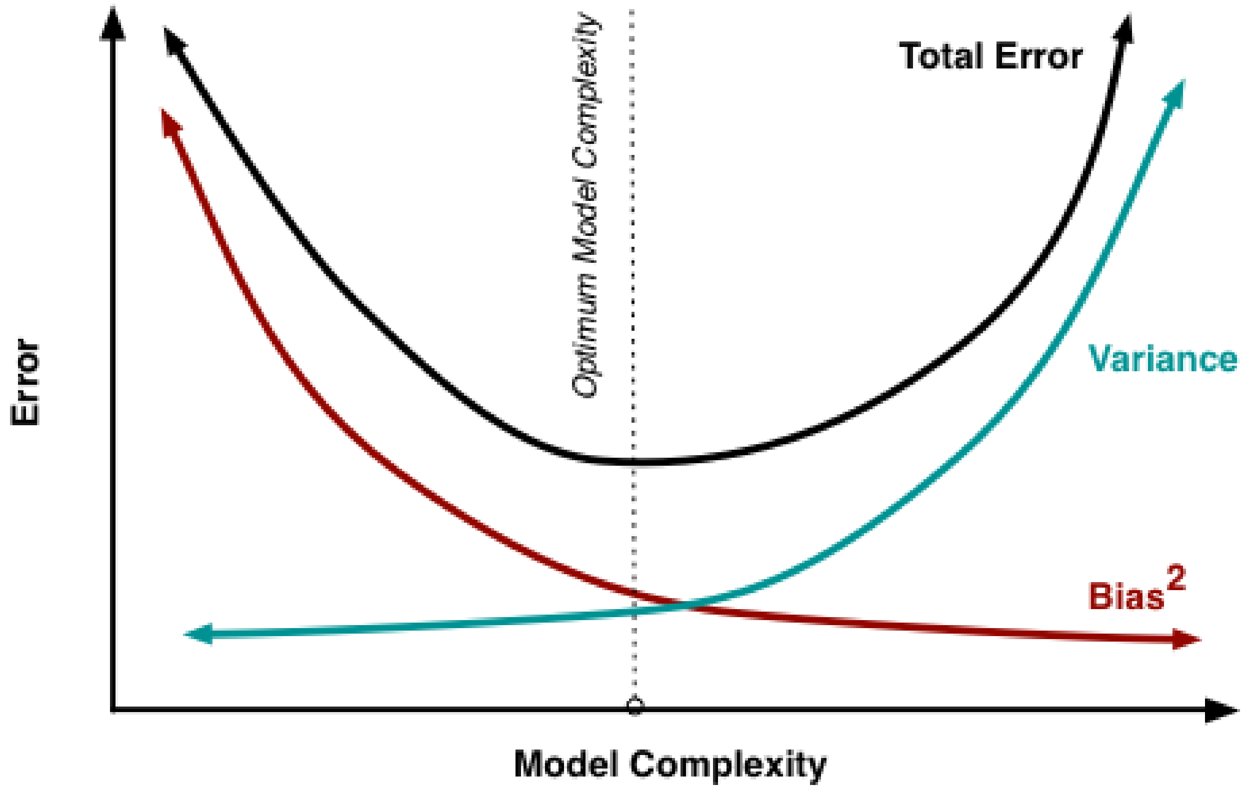


Learning Challenges

- Goal
  - Classify well new testing data
  - Model generalizes well to new testing data
- Variance
  - Amount by which model would change if we estimated it using a different training data set
  - More complex models result in higher variance
- Bias
  - Error introduced by approximating a real-life problem by a much simpler model
  - E.g., assume linear model (linear regression), then error is high
  - More complex models result in lower bias

Bias-Variance tradeoff

Bias-Variance Tradeoff



Model underfits the data

Model overfits the data

Outline

- Probability review
  - Conditional probabilities, Bayes Theorem
- Linear algebra review
  - Matrix and vector operations
- Linear regression
  - Simple linear regression
  - Optimal simple linear regression model
  - Correlation coefficient
  - Lab

Resources

Probability

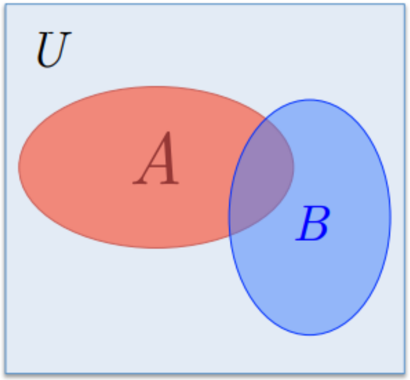
- Review notes from Stanford's machine learning class
- Sam Roweis's probability review

Linear algebra

- Review notes from Stanford's machine learning class
- Sam Roweis's linear algebra review

Conditional Probability

- $\mathrm{P}(A \mid B)$ = Fraction of worlds in which B is true that also have A true



What if we already know that B is true?

That knowledge changes the probability of A

- Because we know we're in a world where B is true

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

Def: Events A and B are **independent** if and only if

$$\Pr[\, A \cap B \,] = \Pr[A] \cdot \Pr[B]$$

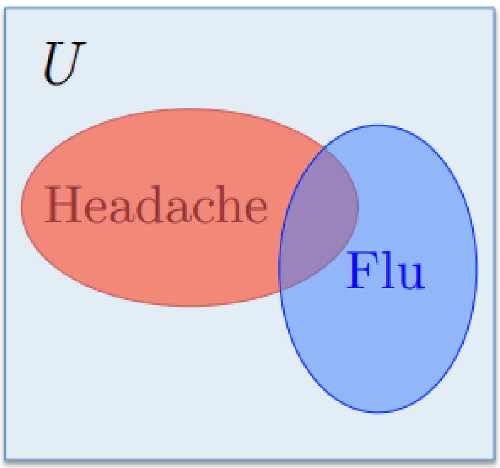
If A and B are independent

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]} = \frac{\Pr[A]\Pr[B]}{\Pr[B]} = \Pr[\mathrm{A}]$$

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



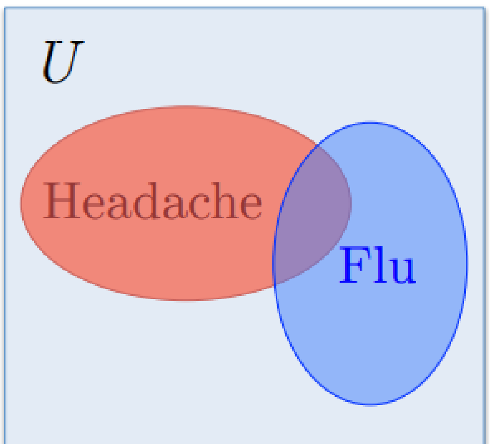
P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with the flu there's a 50-50 chance you'll have a headache."

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$



P(headache) = 1/10
P(flu) = 1/40
P(headache | flu) = 1/2

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu."

Is this reasoning good?

Inference from Conditional Probability

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$P(A \wedge B) = P(A \mid B) \times P(B)$$

$P(\text{headache}) = 1/10$
$P(\text{flu}) = 1/40$
$P(\text{headache} \mid \text{flu}) = 1/2$

Want to solve for:
$P(\text{headache} \wedge \text{flu}) = ?$
$P(\text{flu} \mid \text{headache}) = ?$

$$P(\text{headache} \wedge \text{flu}) = P(\text{headache} \mid \text{flu}) \times P(\text{flu}) = 1/2 \times 1/40 = 0.0125$$

$$P(\text{flu} \mid \text{headache}) = P(\text{headache} \wedge \text{flu}) / P(\text{headache}) = 0.0125 / 0.1 = 0.125$$

Bayes Theorem

Bayes' Rule

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)}$$

- Exactly the process we just used
- The most important formula in probabilistic machine learning

(Super Easy) Derivation:

$$\boxed{\begin{aligned} P(A \wedge B) \\ P(B \wedge A) \end{aligned}} \begin{aligned} &= P(A \mid B) \times P(B) \\ &= P(B \mid A) \times P(A) \end{aligned}$$

these are the same

Just set equal...

$$P(A \mid B) \times P(B) = P(B \mid A) \times P(A)$$

and solve...

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

Vectors and matrices

- Vector in R^n is an ordered set of n real numbers.
  - e.g. v = (1,6,3,4) is in R^4
  - A column vector:
  - A row vector:

$$\begin{pmatrix} 1 \\ 6 \\ 3 \\ 4 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 6 & 3 & 4 \end{pmatrix}$$

- m-by-n matrix is an object in R^{mxn} with m rows and n columns, each entry filled with a (typically) real number:

$$\begin{pmatrix} 1 & 2 & 8 \\ 4 & 78 & 6 \\ 9 & 3 & 2 \end{pmatrix}$$

Matrix multiplication

We will use upper case letters for matrices. The elements are referred by $A_{i,j}$.

- Matrix product:

$$A \in \mathbb{R}^{m \times n} \qquad B \in \mathbb{R}^{n \times p}$$

$$C = AB \in \mathbb{R}^{m \times p}$$

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

e.g.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{pmatrix}$$

Matrix transpose

Transpose: You can think of it as

- "flipping" the rows and columns

OR

- "reflecting" vector/matrix on line

e.g. $\begin{pmatrix} a \\ b \end{pmatrix}^T = \begin{pmatrix} a & b \end{pmatrix}$

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^T = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

A is a symmetric matrix if $A = A^T$

Inverse of a matrix

- Inverse of a square matrix A, denoted by A^{-1} is the *unique* matrix s.t.
  - $AA^{-1} = A^{-1}A = I$ (identity matrix)

- If A^{-1} and B^{-1} exist, then
  - $(AB)^{-1} = B^{-1}A^{-1}$,
  - $(A^T)^{-1} = (A^{-1})^T$

- For diagonal matrices $\mathbf{D}^{-1} = \mathrm{diag}\{d_1^{-1}, \ldots, d_n^{-1}\}$

Linear independence

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

$$\begin{pmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{pmatrix}\begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

e.g. $\begin{pmatrix} 1 & 0 \\ 2 & 3 \\ 1 & 3 \end{pmatrix}$

(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

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

System of linear equations

$$\begin{array}{rcrcl} 4x_1 & - & 5x_2 & = & -13 \\ -2x_1 & + & 3x_2 & = & 9. \end{array}$$

Matrix formulation

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

If A has an inverse, solution is $x = A^{-1}b$

Linear regression

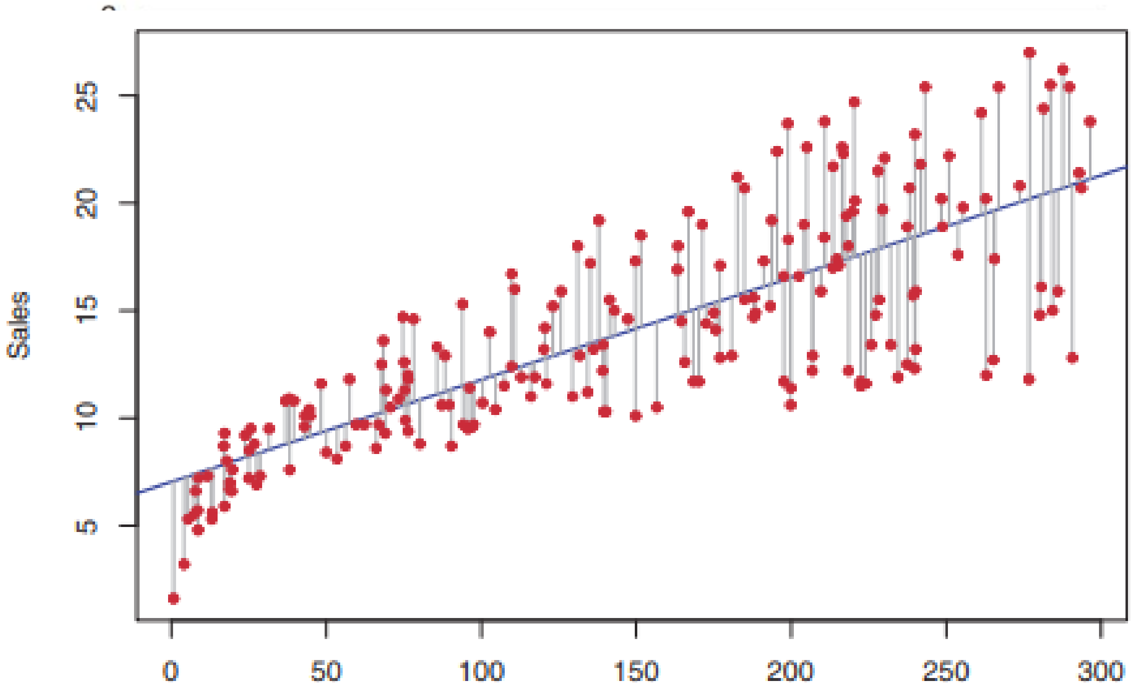
- One of the most widely used techniques
- Fundamental to many complex models
  - Generalized Linear Models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to solve in closed form
- Efficient practical algorithm (gradient descent)

Linear regression

Given:

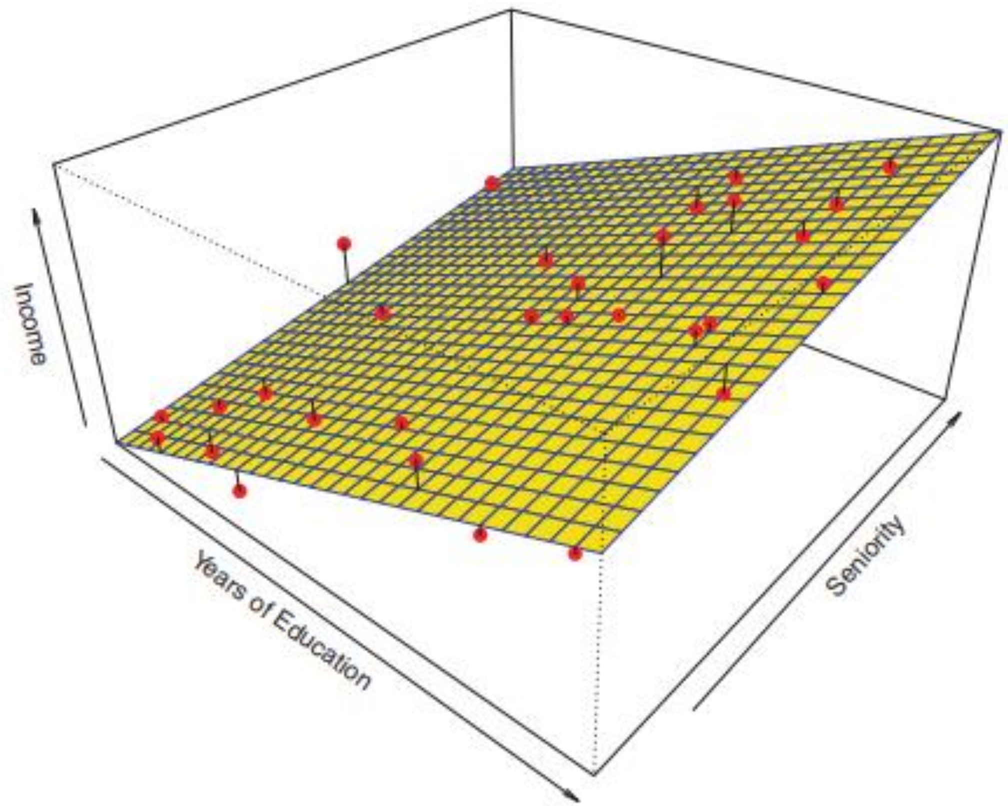
- Data $\boldsymbol{X} = \left\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right\}$ where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ Features
- Corresponding labels $\boldsymbol{y} = \left\{y^{(1)}, \ldots, y^{(n)}\right\}$ where $y^{(i)} \in \mathbb{R}$

Response variables



Simple Linear Regression: 1 predictor

Income Prediction



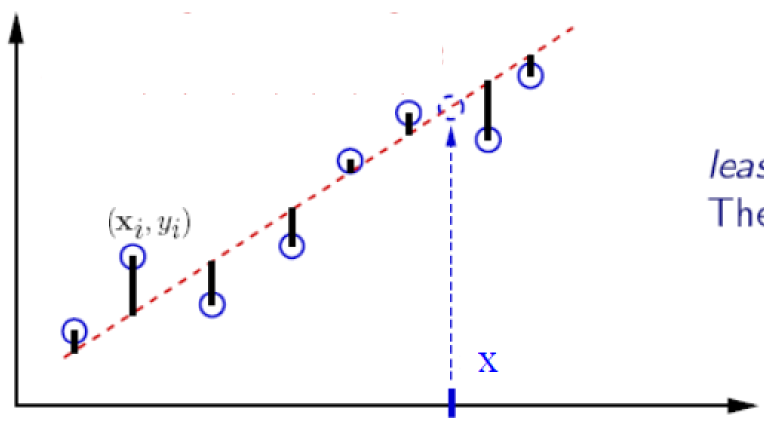
Linear Regression with 2 predictors
Multiple Linear Regression

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

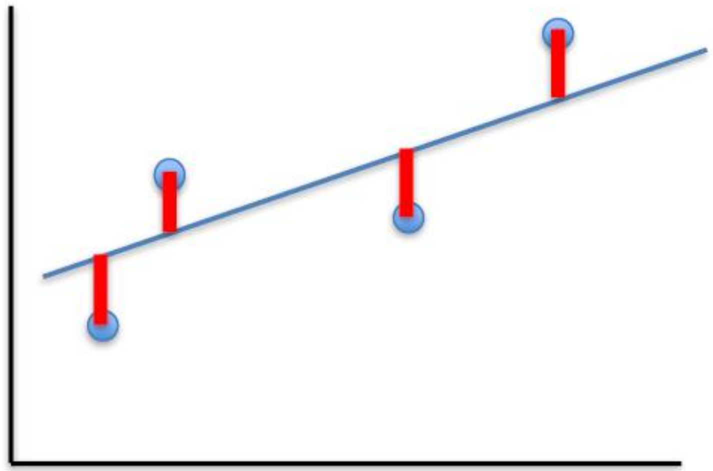
Least squares Linear Regression

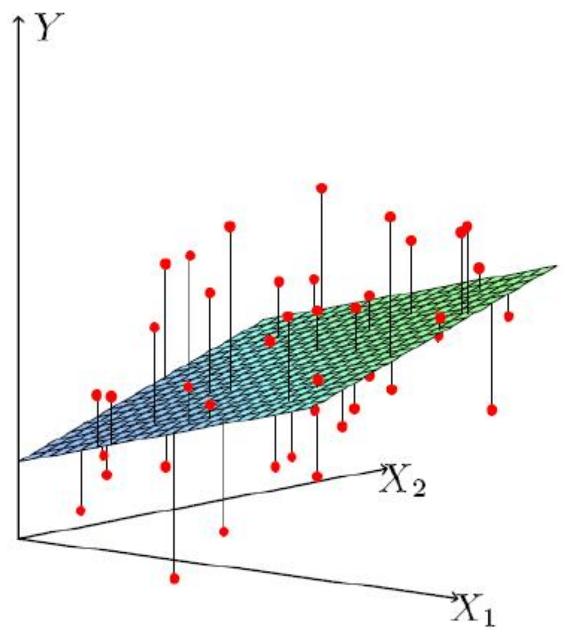
- Cost Function

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) - y^{(i)} \right)^2$$

Mean Square Error (MSE)

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

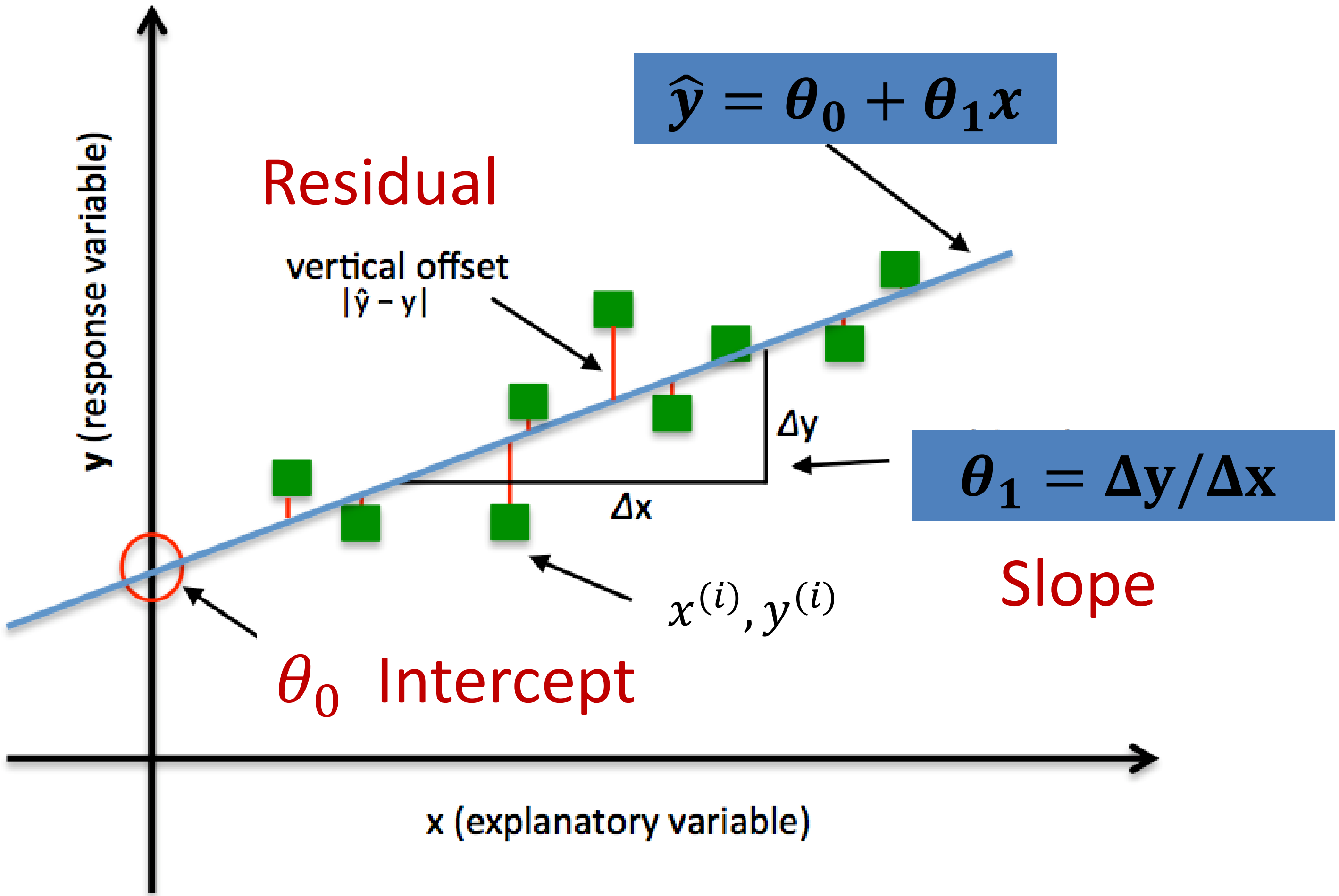




Terminology and Metrics

- Residuals
  - Difference between predicted values and actual values
  - Predicted value for example i is: $\hat{y}^{(i)} = h_\theta\left(x^{(i)}\right)$
  - $R^{(i)} = \left|y^{(i)} - \hat{y}^{(i)}\right| = |y^{(i)} - (\theta_0 + \theta_1 x^{(i)})|$
- 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



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

MSE$= \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$

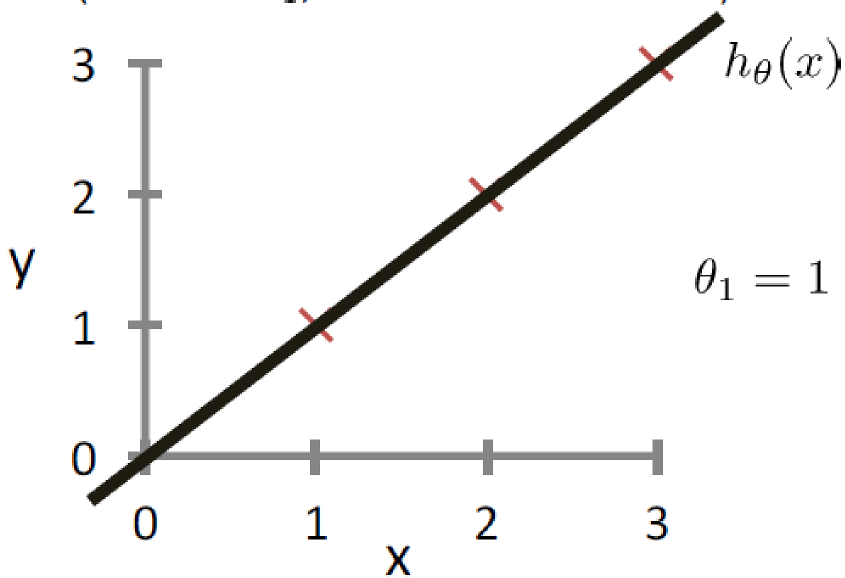
Intuition on MSE

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right)^2$$

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

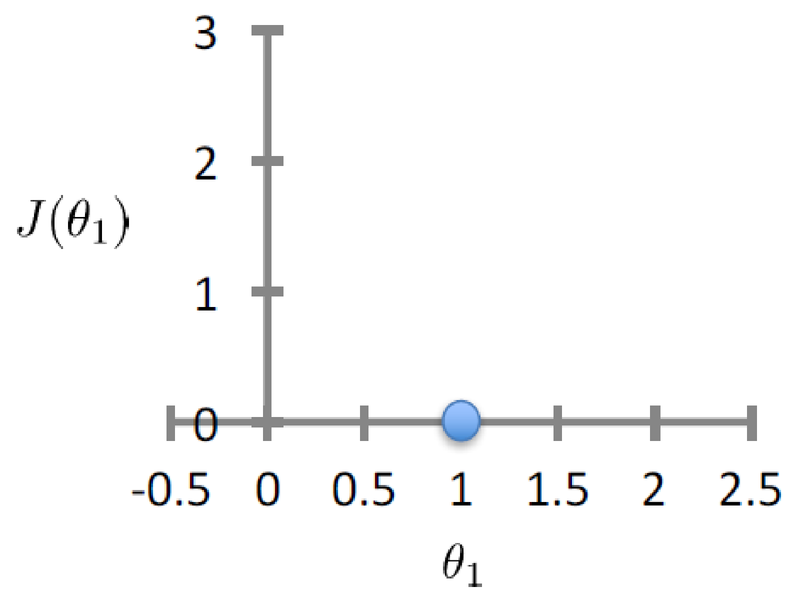
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Fix $\theta_0 = 0$

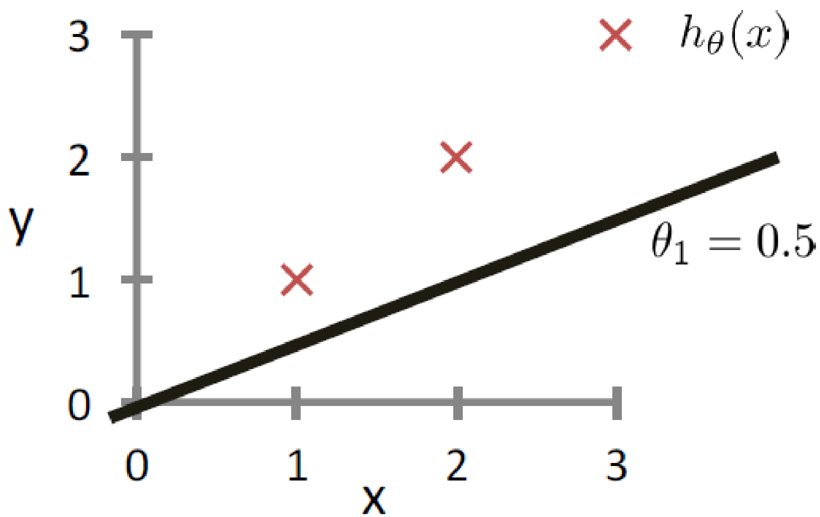
Intuition on MSE

$$J(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2$$

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

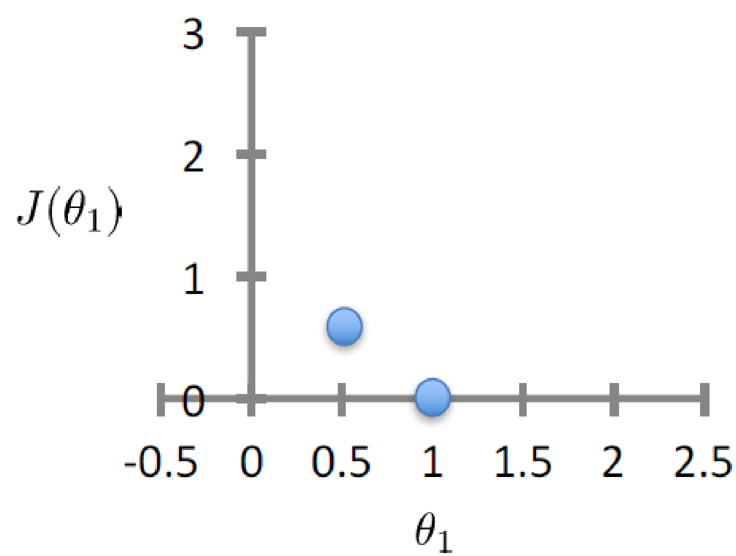
$h_\theta(x)$

(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

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

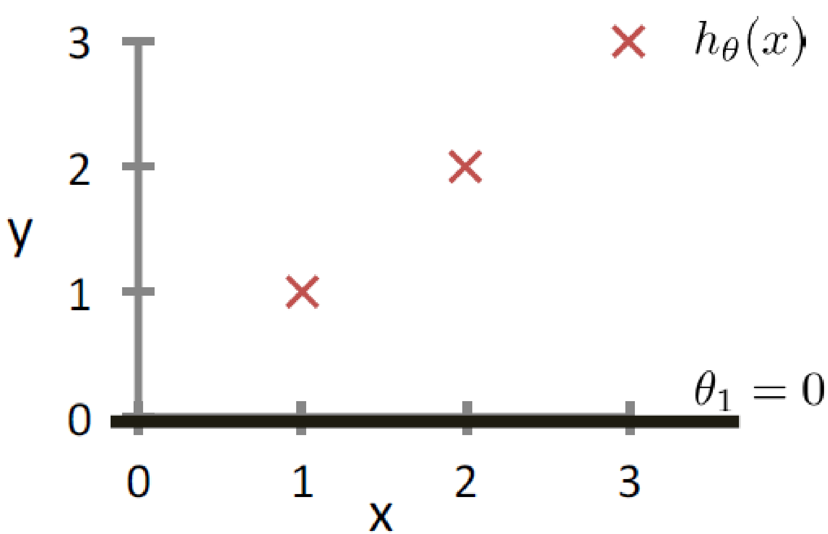
Intuition on MSE

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)} \right)^2$$

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

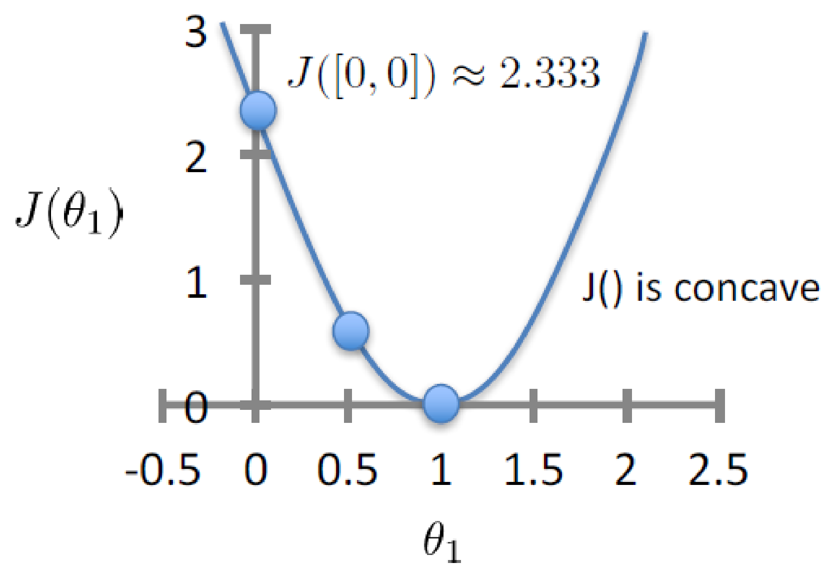
$h_\theta(x)$

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



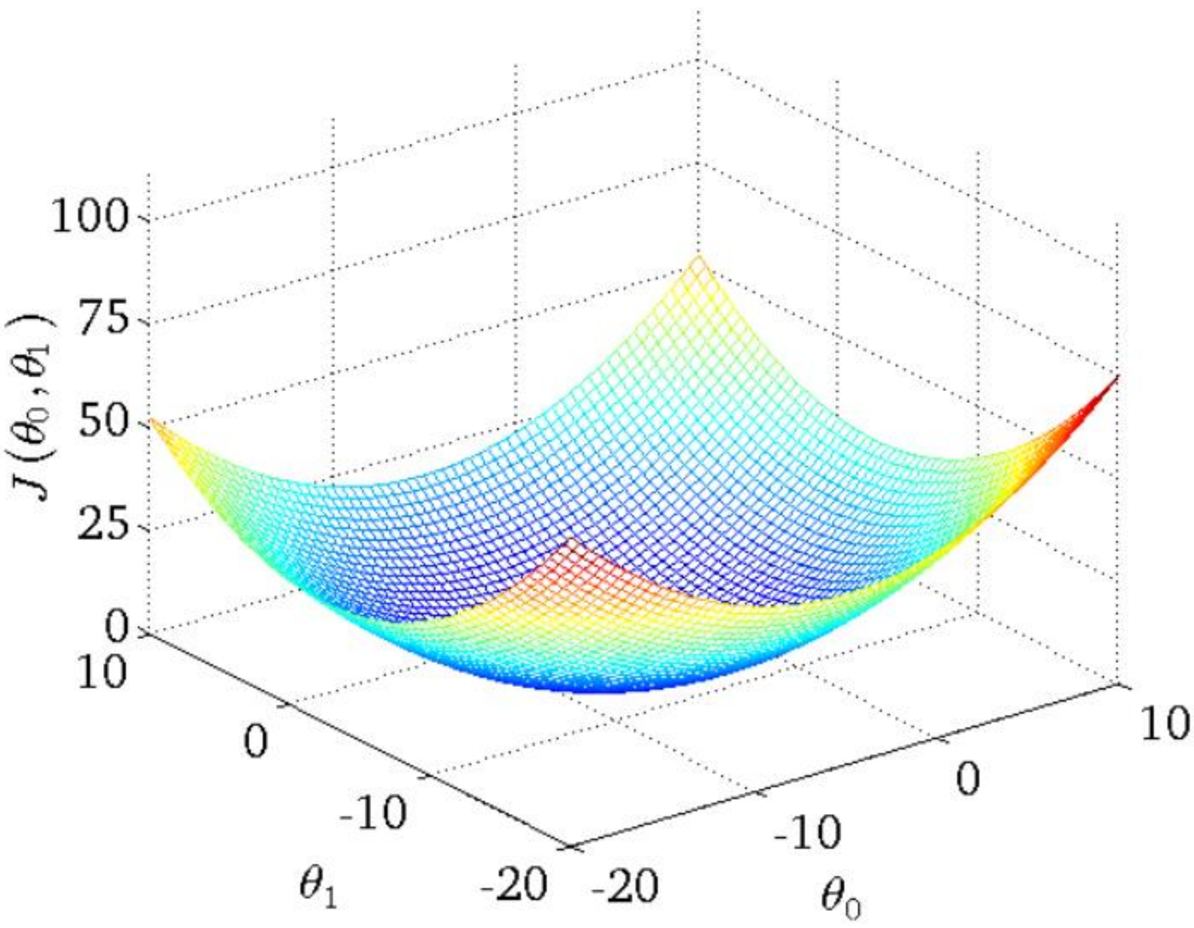
$J(\theta_1)$

(function of the parameter θ_1)



Based on example
by Andrew Ng

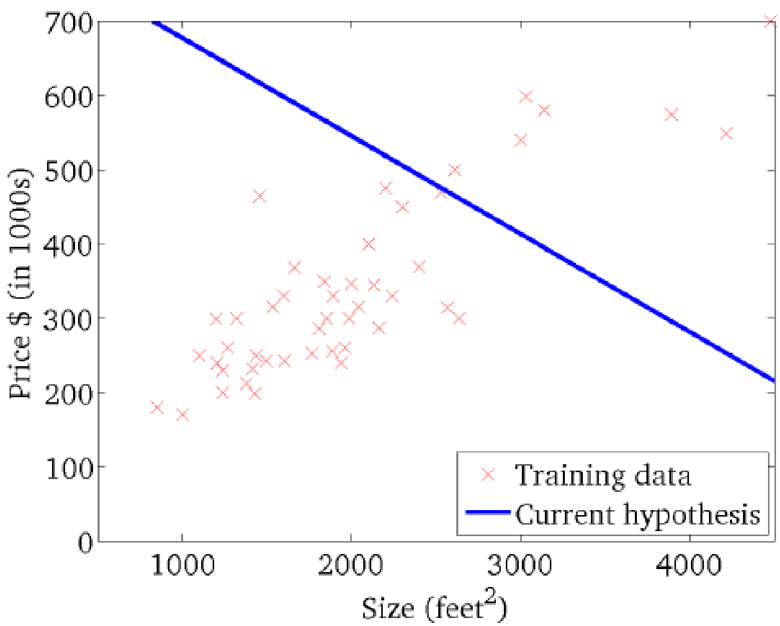
MSE function



Relation between h and J

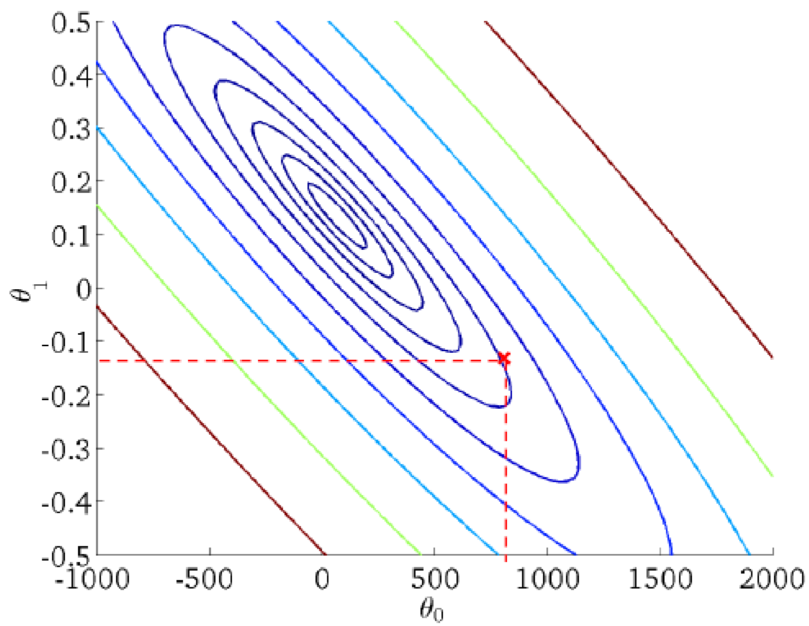
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

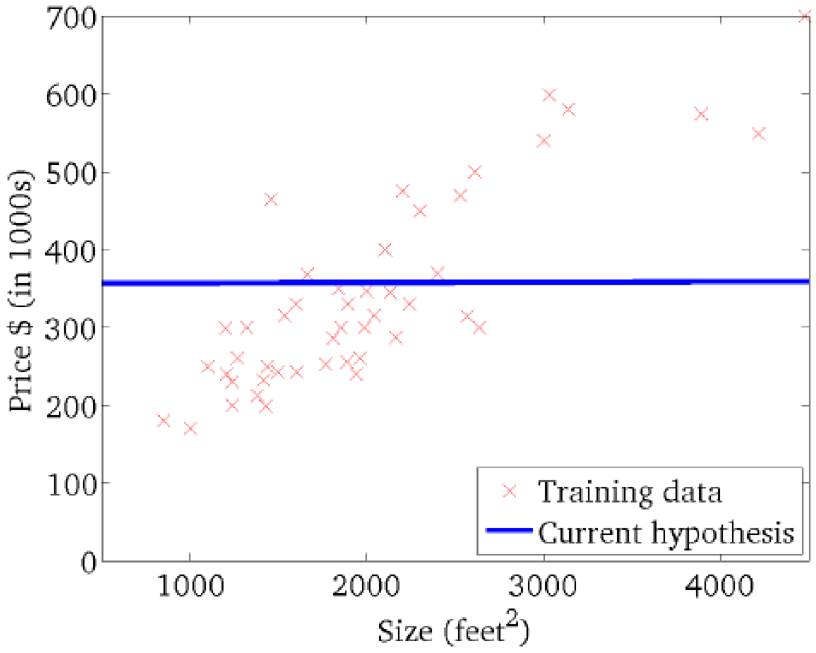
(function of the parameters θ_0, θ_1)



Relation between h and J

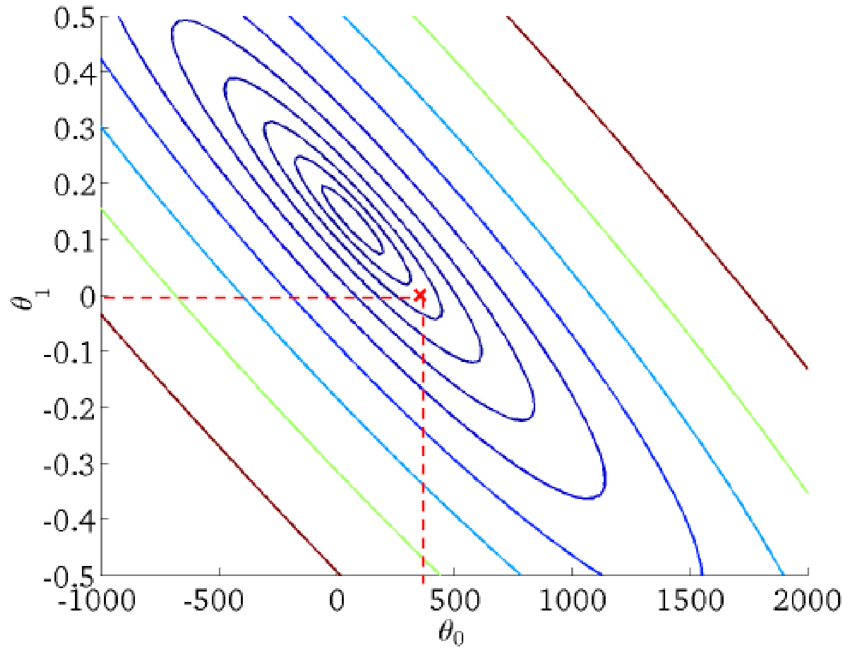
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

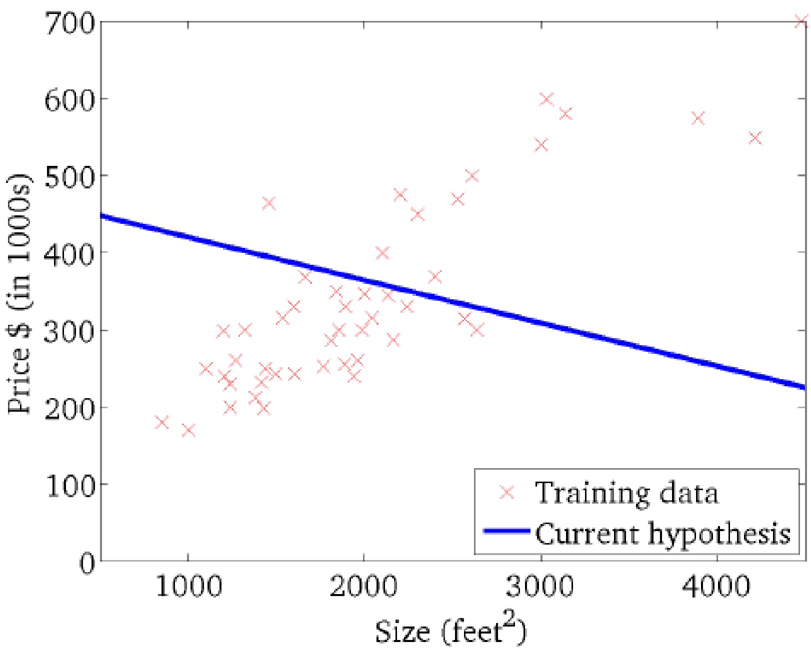
(function of the parameters θ_0, θ_1)



Relation between h and J

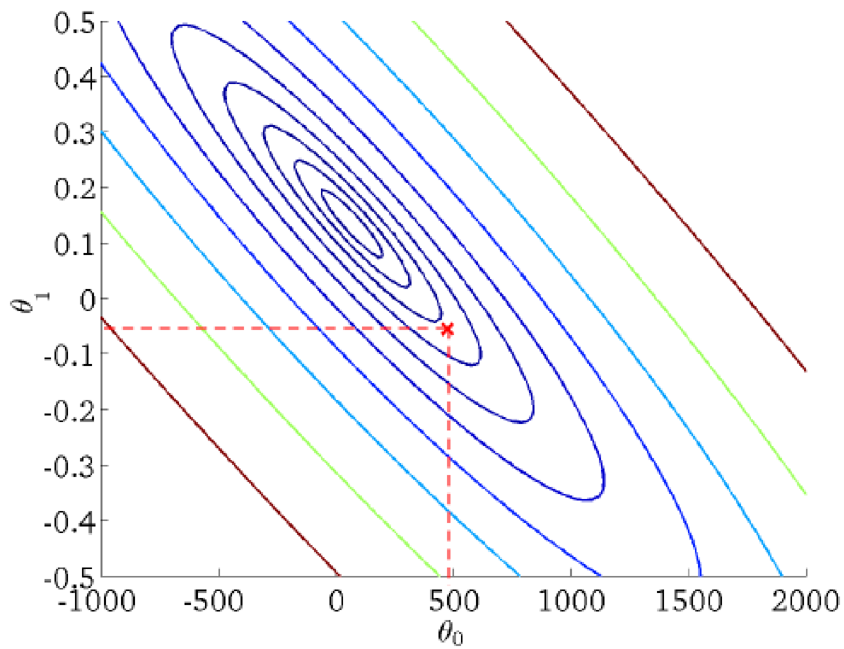
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

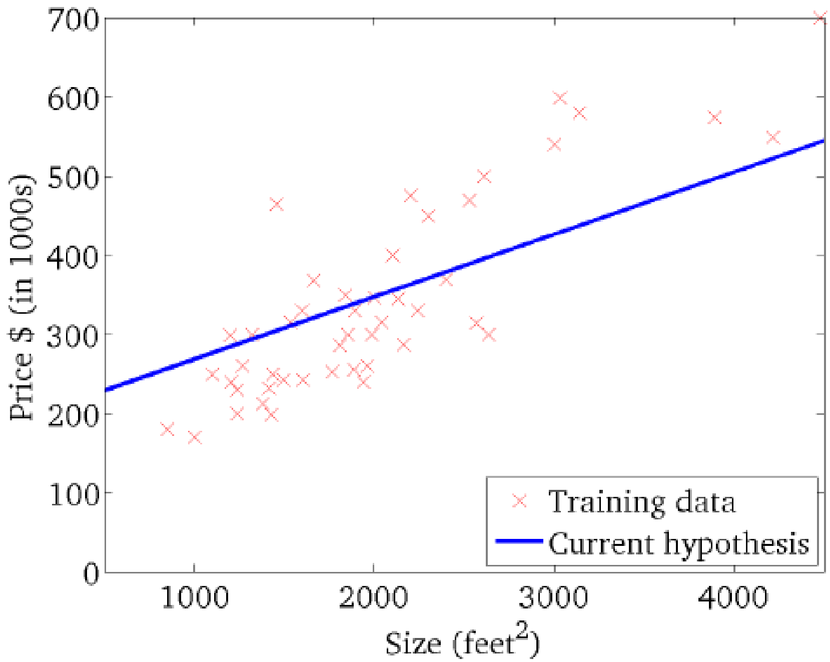
(function of the parameters θ_0, θ_1)



Relation between h and J

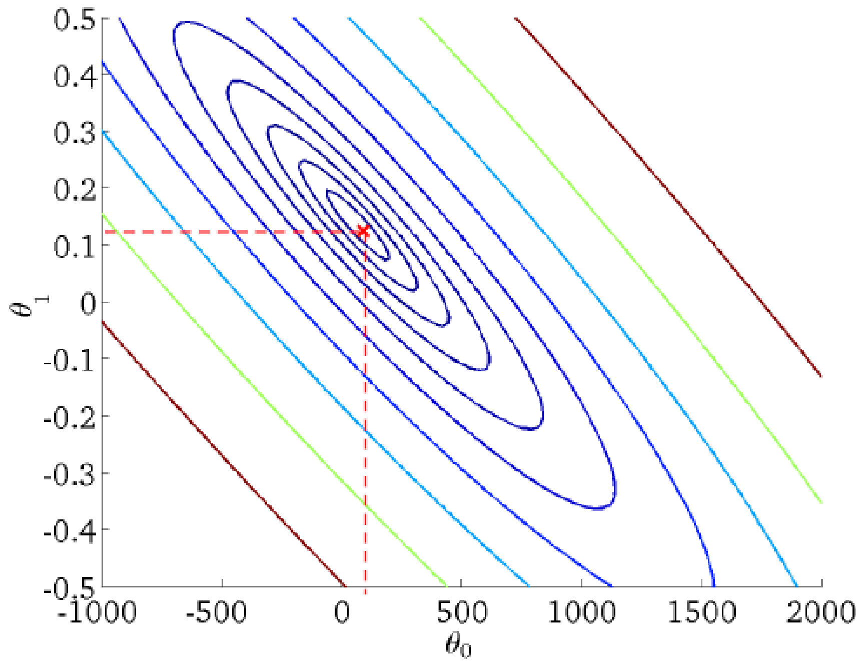
$h_\theta(x)$

(for fixed θ_0, θ_1, this is a function of x)



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



How to find optimal model parameters θ to minimize MSE J?

Simple linear regression

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

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{2}{n}\sum_{i=1}^{n}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{2}{n}\sum_{i=1}^{n} x^{(i)}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = 0$$

- Solution of min loss

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

$$-\ \theta_1 = \frac{\sum (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum \left(x^{(i)} - \bar{x}\right)^2}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x^{(i)}}{n}$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y^{(i)}}{n}$$

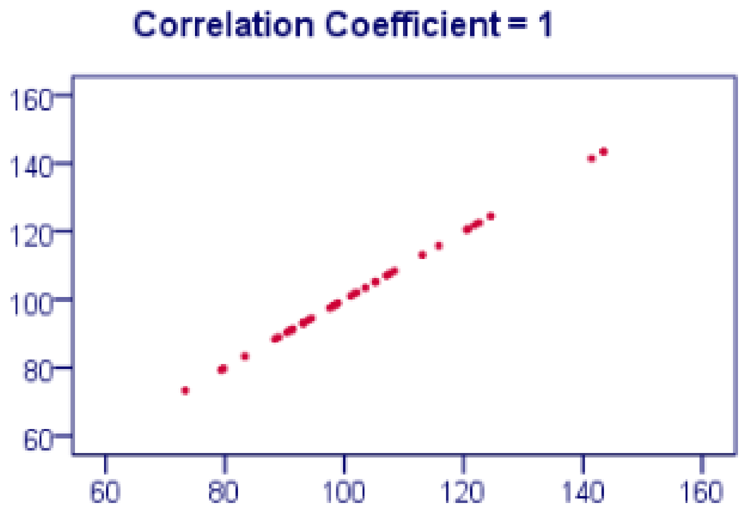
How Well Does the Model Fit?

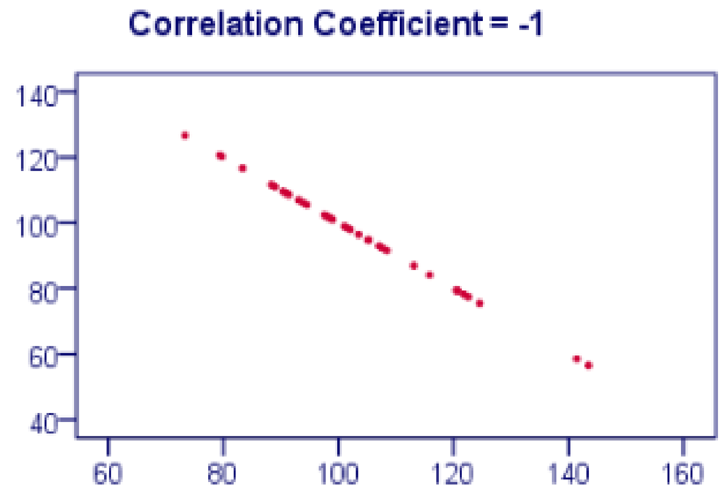
- Correlation between feature and response
  - Pearson's correlation coefficient

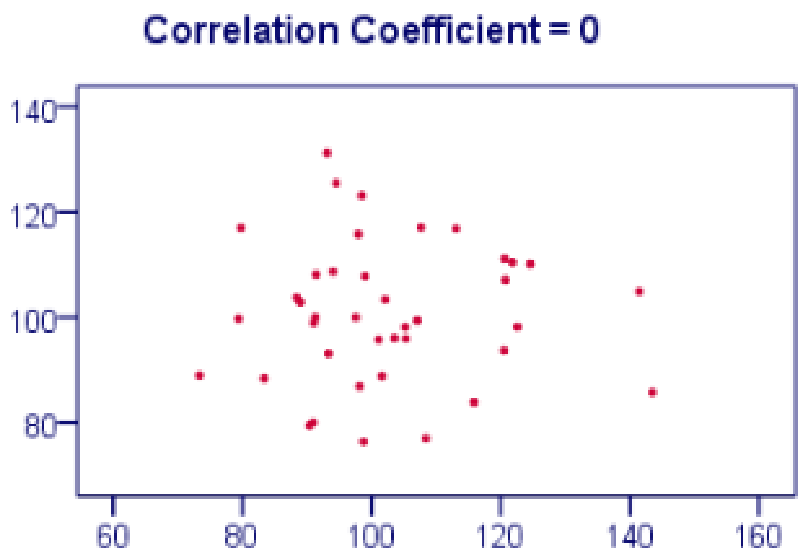
$$\mathrm{Cor}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}},$$

- 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 the coefficient is, the stronger the correlation

Correlation Coefficient

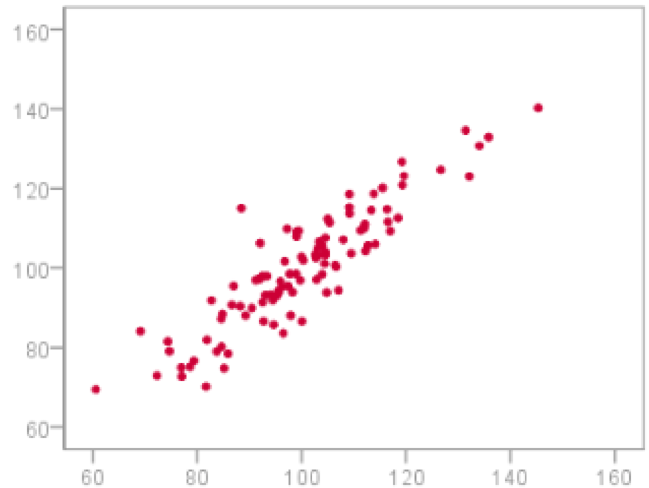




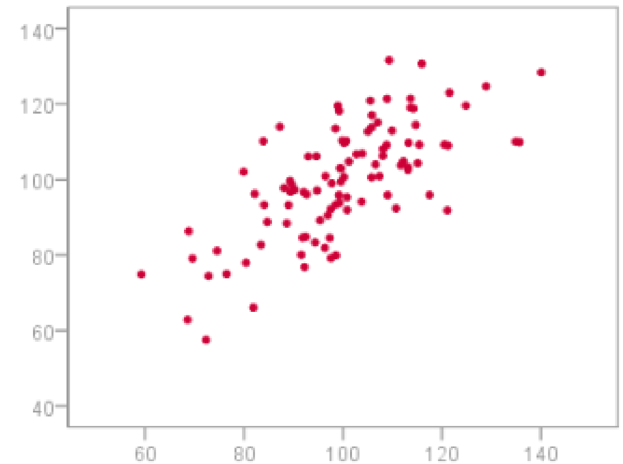


Positive/Negative Correlation

Positive Correlation

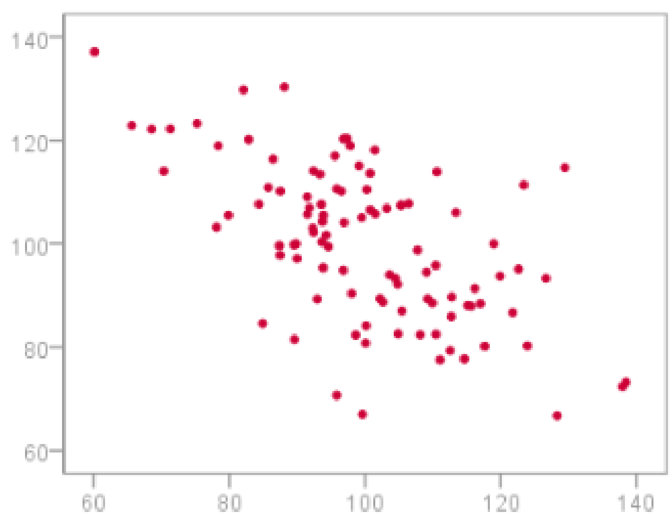


r = 0.9

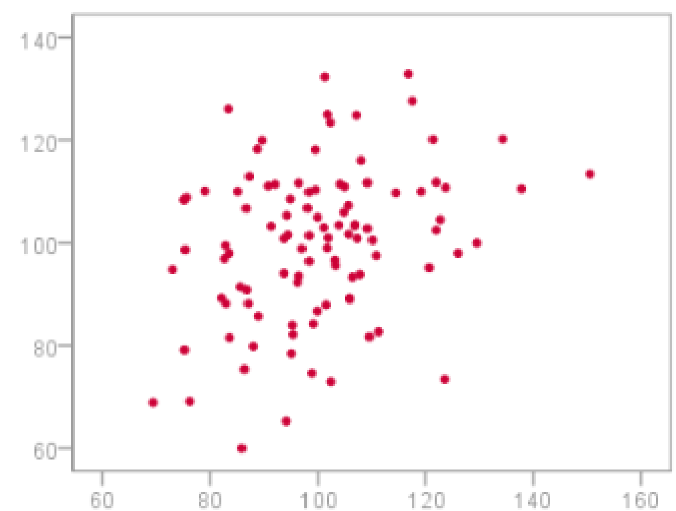


r = 0.7

Negative Correlation



r = -0.6



r = 0.3

Review linear regression

- 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
  - $\theta_0 = \bar{y} - \theta_1 \bar{x}$
  - $\theta_1 = \frac{\sum (x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\sum (x^{(i)} - \bar{x})^2}$

Acknowledgements

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