THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

COMP4434 Big Data Analytics

Lecture 2 Basic Statistical Analysis

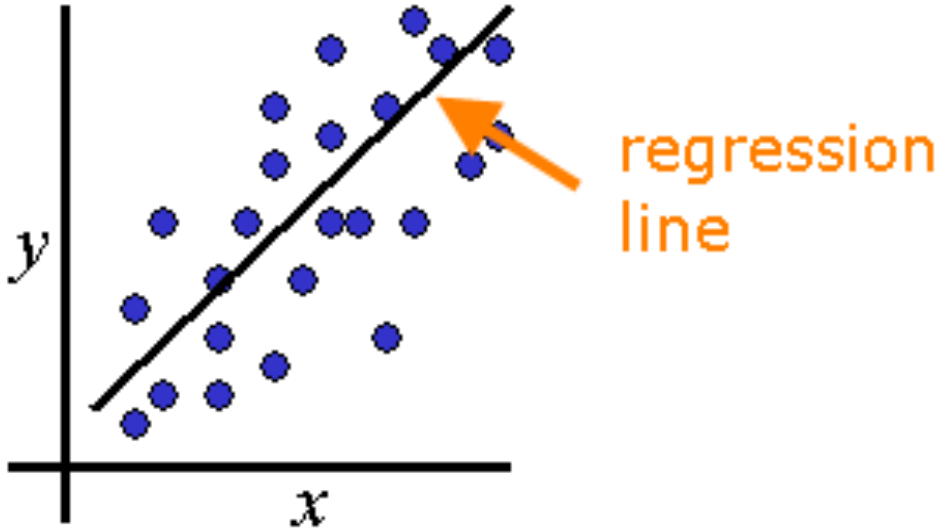
HUANG Xiao

xiaohuang@comp.polyu.edu.hk

Linear Regression

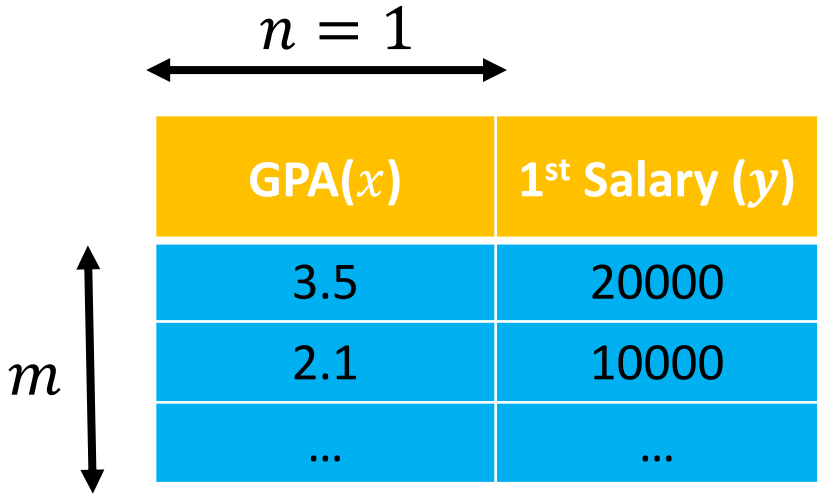
- A Regression Problem - Line Fitting
  - E.g., “My GPA is 2.9, what will be my salary?”

GPA (x)	1st Salary (y)
3.5	20000
2.1	10000
…	…

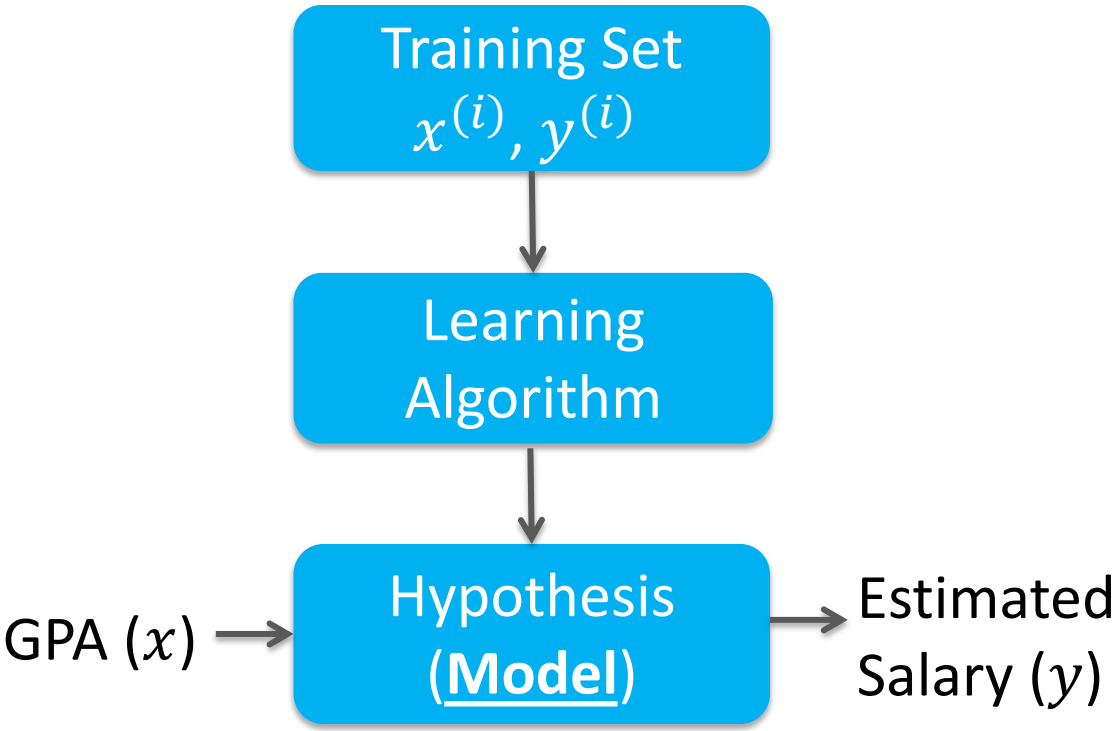


Notations

- x : input variables/attributes/**features**
- y : output variable/attribute/target variable
- m : number of training examples
- n : number of input variables
  - Univariate: $n = 1$
  - Multivariate: $n > 1$

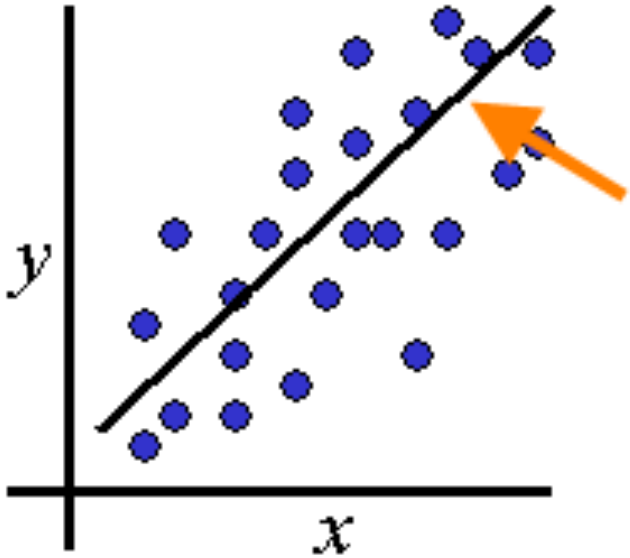


Model $h_\theta(x) = \theta_0 + \theta_1 x$

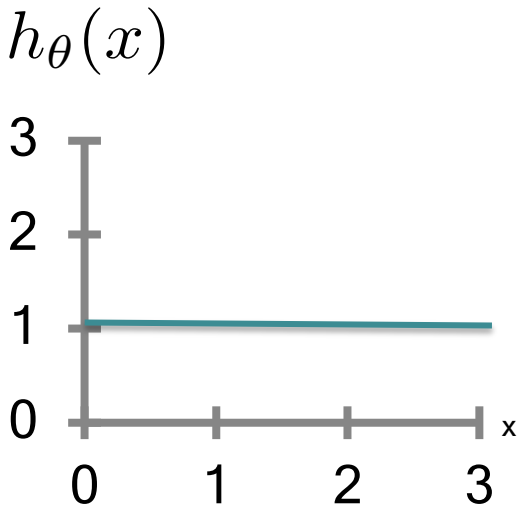


How do we represent h?

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

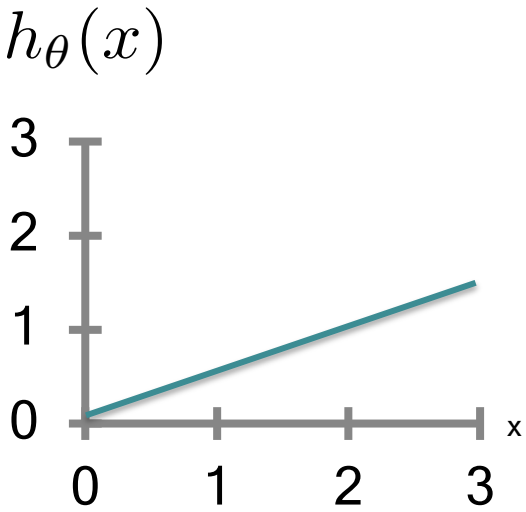


How to Set θ



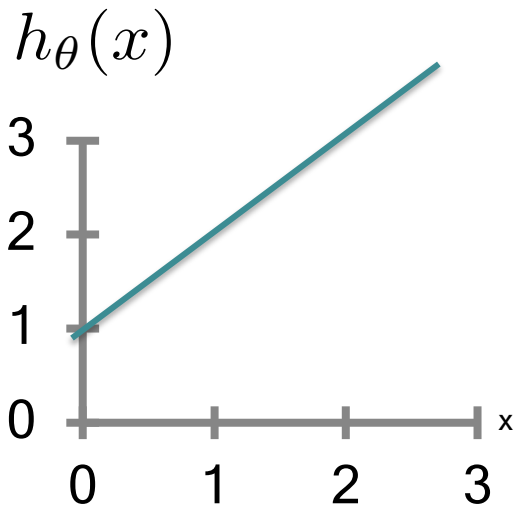
$\theta_0 = 1$

$\theta_1 = 0$



$\theta_0 = 0$

$\theta_1 = 0.5$



$\theta_0 = 1$

$\theta_1 = 1$

What is the Best Fitting Line?

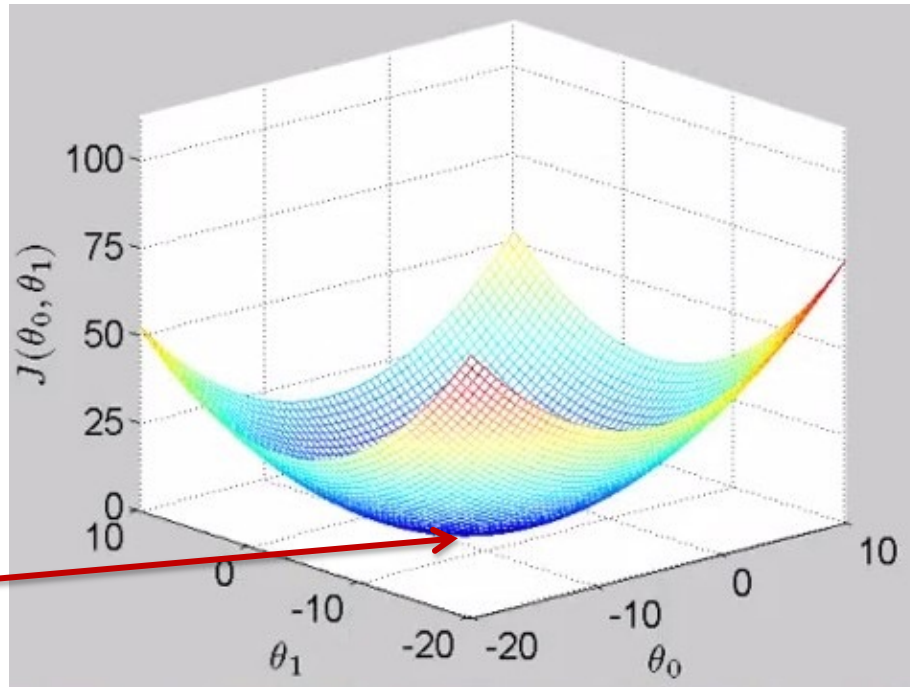
- Finding θ, which makes $h_\theta(x)$ closest to y for all training data $\left(x^{(i)}, y^{(i)}\right)$
- Mathematical definition: **Cost Function $J(\boldsymbol{\theta_0}, \boldsymbol{\theta_1})$**

$$\min_{\theta_0,\theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Error: (Estimated – Actual)
Squared Error: to be positive

Cost Function $J(\theta_0, \theta_1)$

$$J(\theta_0, \theta_1) = \min_{\theta_0, \theta_1} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$



Our Target

- Input: $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$
- Start with some θ_0, θ_1 (e.g., $\theta_0 = 0, \theta_1 = 0$)
- Keep refining θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we hopefully end up at a minimum

Gradient Descent Algorithm

Repeat until convergence {

$$\theta_0 = \theta_0 - \alpha \cdot \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0}$$

$$\theta_1 = \theta_1 - \alpha \cdot \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1}$$

}

Derivative
Slope
Gradient

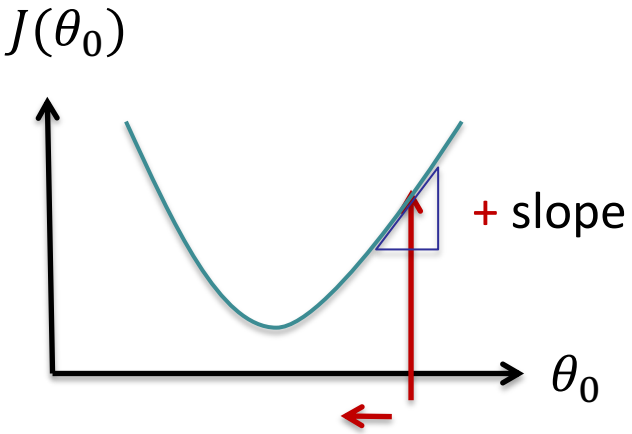
Learning Rate
(or Step Size)

Three Problems:

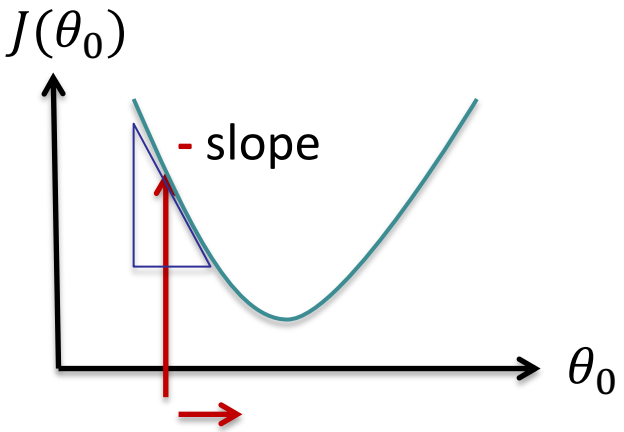
1. How to compute the derivative?
2. How to set the learning rate?
3. What is the convergence criteria?

Derivative

$$\theta_j = \theta_j - \alpha \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_j}$$



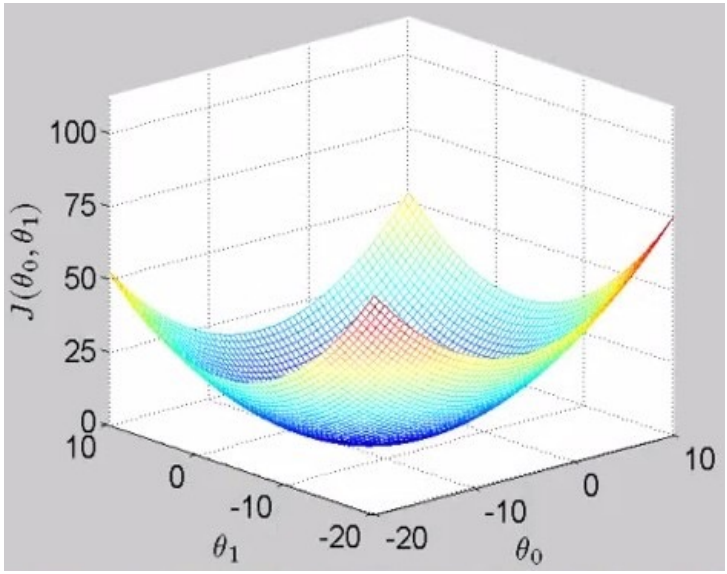
$\theta_0 := \theta_0 - \alpha$ (+ number)

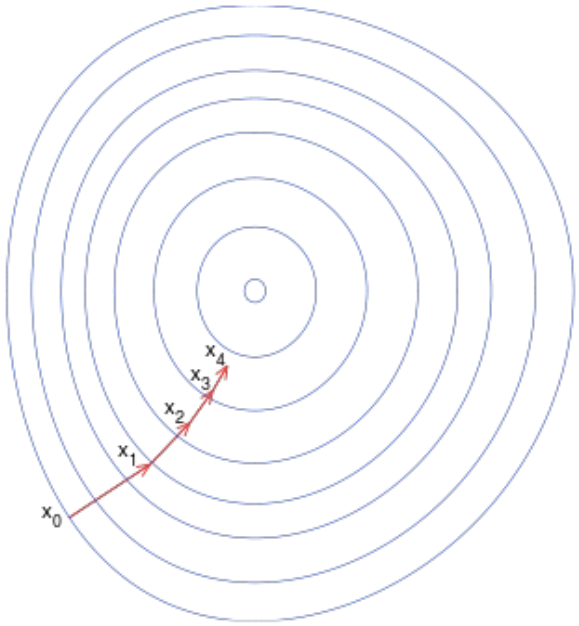


$\theta_0 := \theta_0 - \alpha$ (- number)

Geometric Interpretation: Gradient Decent

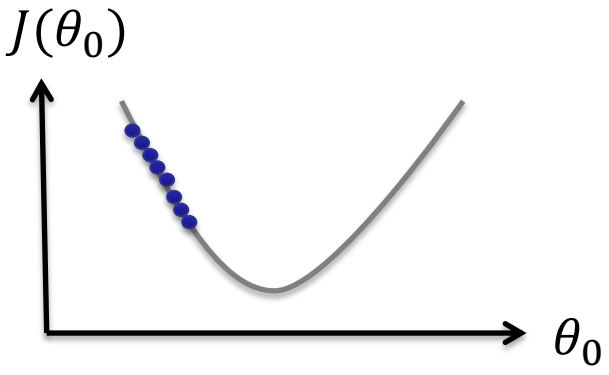
- The gradient can be interpreted as the direction and rate of fastest increase
- Parameters update in reverse direction of gradient

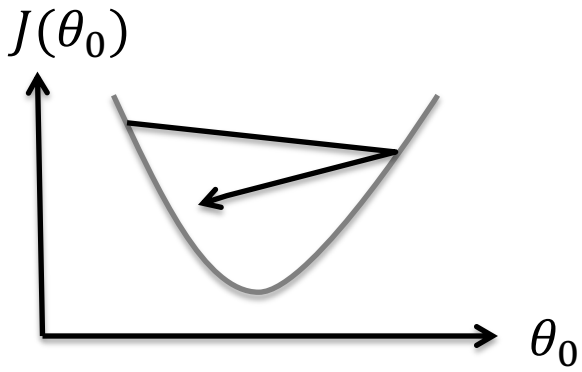




Step Size

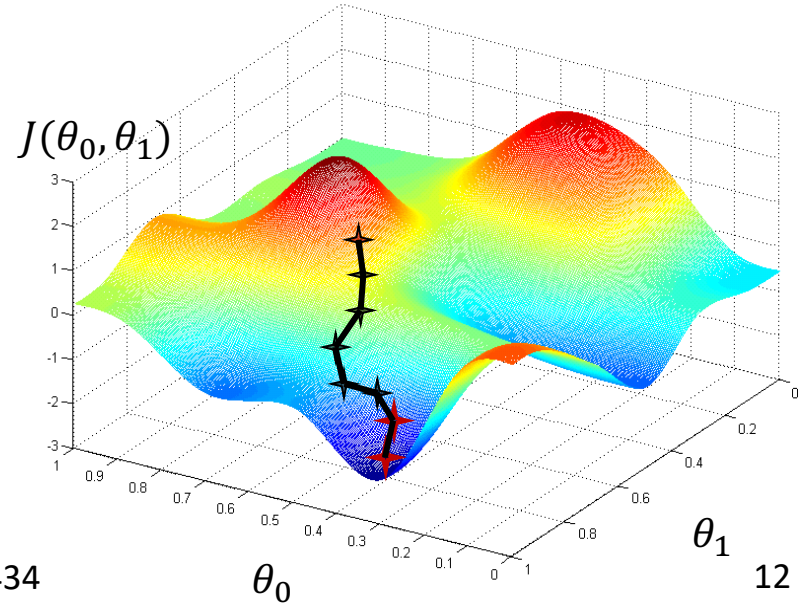
- Learning rate / Step size α is a user parameter.





Small α => slow convergence

Large α => fail to converge



Gradient

- Gradient of J is the vector

$$\nabla J(\theta_0, \theta_1) = \begin{bmatrix} \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0} \\ \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} \end{bmatrix}$$

- Partial derivative ∂ of a function $J(\theta_0, \theta_1)$ of several variables θ_0, θ_1 is its derivative with respect to one of those variables, with the others held constant

Gradient $\partial J(\theta_0, \theta_1)/\partial \theta_0$

$$\frac{\partial J(\theta_0,\theta_1)}{\partial \theta_0} = \frac{\partial}{\partial \theta_0}\left(\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2\right)$$

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

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

$$= \frac{1}{2m}\sum_{i=1}^{m} 2\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$

$$y = f(u) = f(g(x))$$
$$\frac{dy}{dx} = \frac{dy}{du}\frac{du}{dx} = f'(u)g'(x)$$

Chain Rule

Gradient $\partial J(\theta_0, \theta_1)/\partial \theta_1$

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{\partial}{\partial \theta_1}\left(\frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2\right)$$

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

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

$$= \frac{1}{2m}\sum_{i=1}^{m} 2\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \cdot x^{(i)}$$

$$= \frac{1}{m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)x^{(i)}$$

Gradient Descent Algorithm

$n = 0, \theta_0[n] = \theta_1[n] = 0$

REPEAT {

$$\theta_0[n+1] = \theta_0[n] - \frac{\alpha}{m}\sum_{i=1}^{m}\left(\theta_0[n] + \theta_1[n]x^{(i)} - y^{(i)}\right)$$

$$\theta_1[n+1] = \theta_1[n] - \frac{\alpha}{m}\sum_{i=1}^{m}\left(\theta_0[n] + \theta_1[n]x^{(i)} - y^{(i)}\right)x^{(i)}$$

$n = n + 1$

} UNTIL $J(\theta_0[n-1], \theta_1[n-1]) - J(\theta_0[n], \theta_1[n]) < \varepsilon$ OR $n > X$

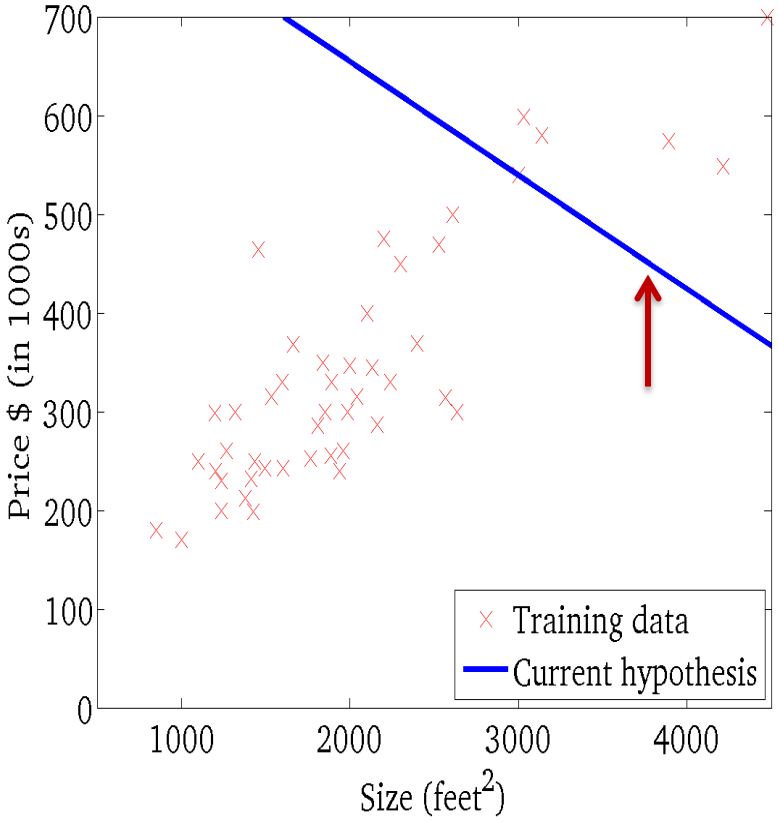
Stop Criteria

- If $J(\theta_0, \theta_1)$ decreases by less than a threshold ε (e.g., 10^{-3}) in one iteration
  - Each iteration, we use **all** m training examples to update θ_0, θ_1
  - Use updated θ_0, θ_1 to recalculate $J(\theta_0, \theta_1)$ and find the decrease
- Or, after X (e.g., 5000) number of iterations

Gradient Descent in Action

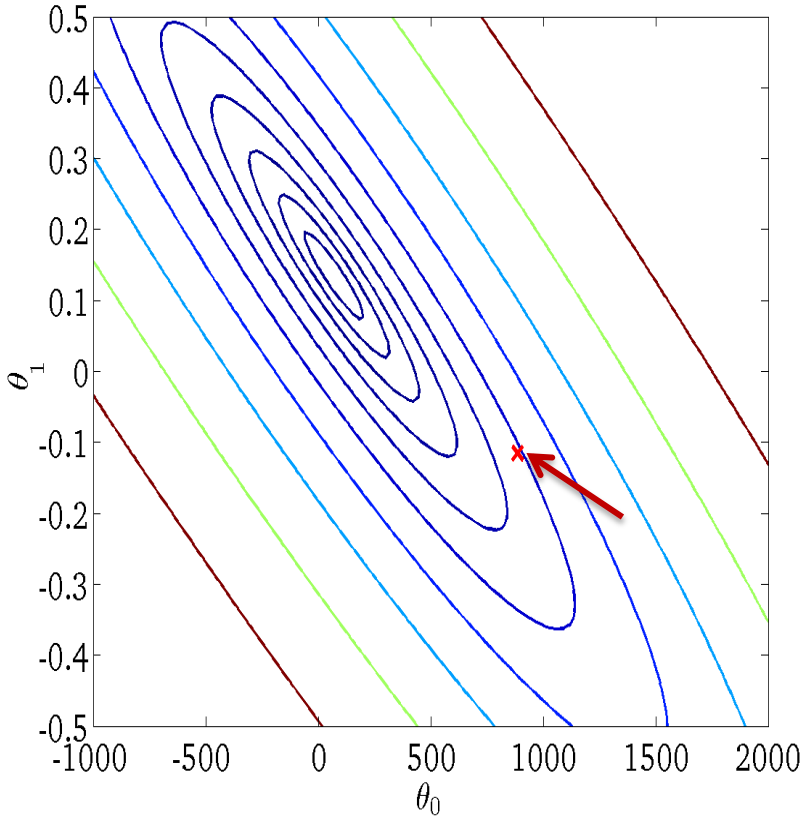
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

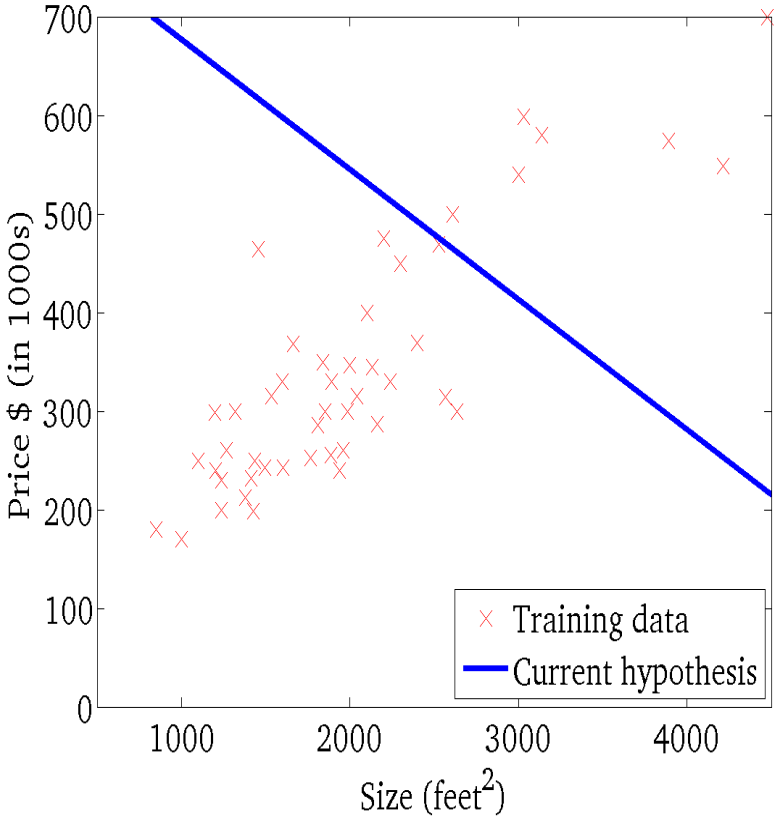
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

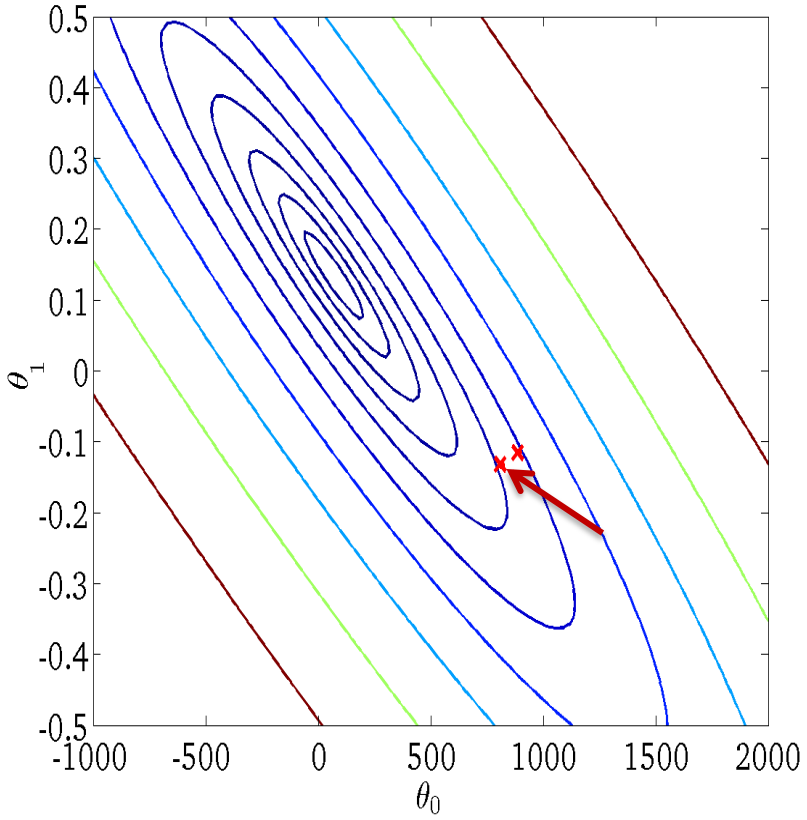
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

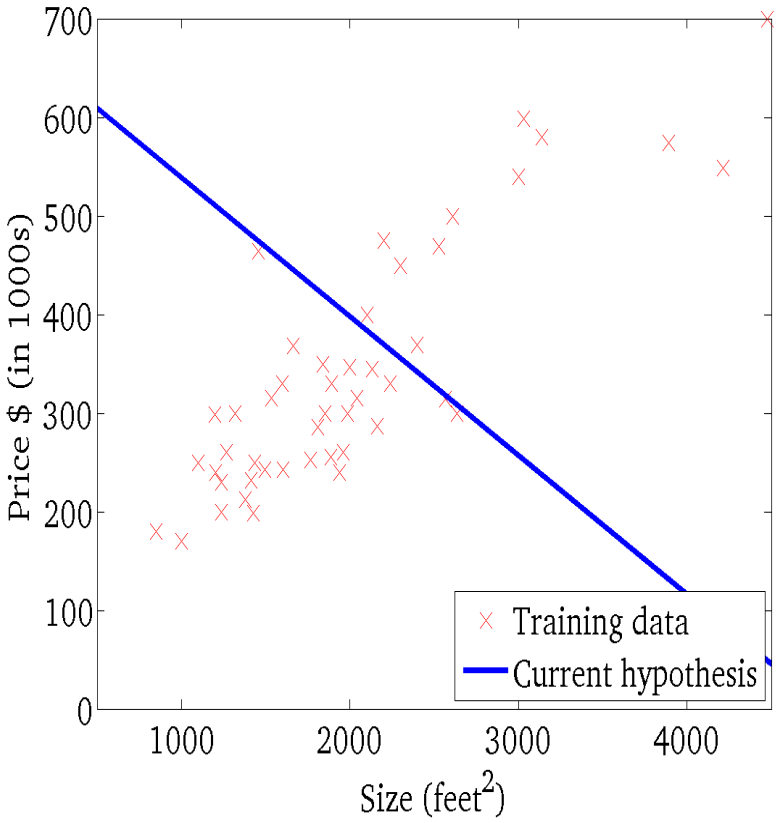
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

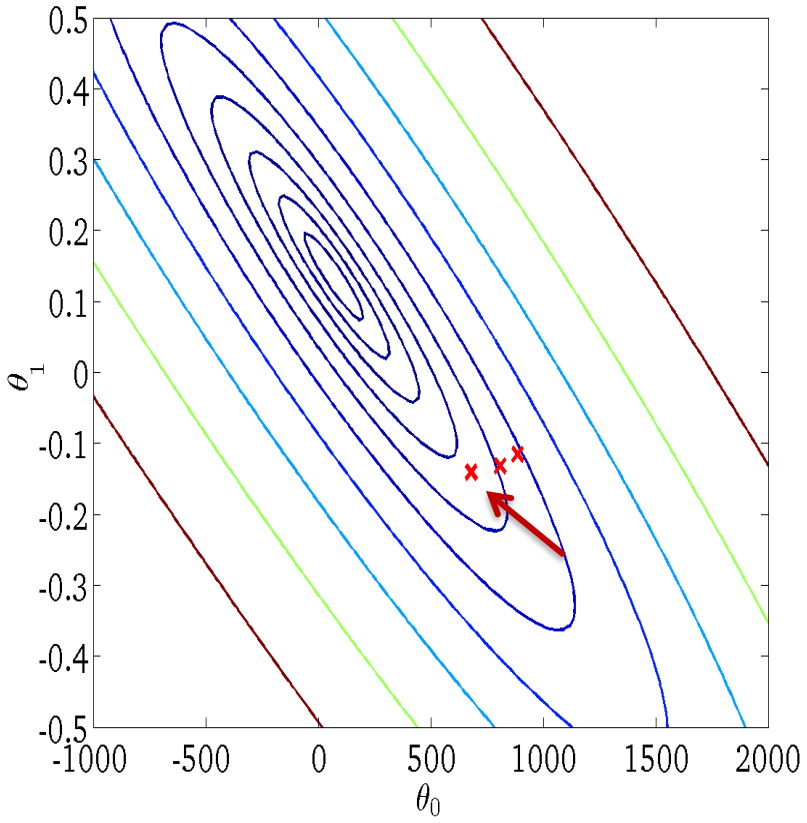
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

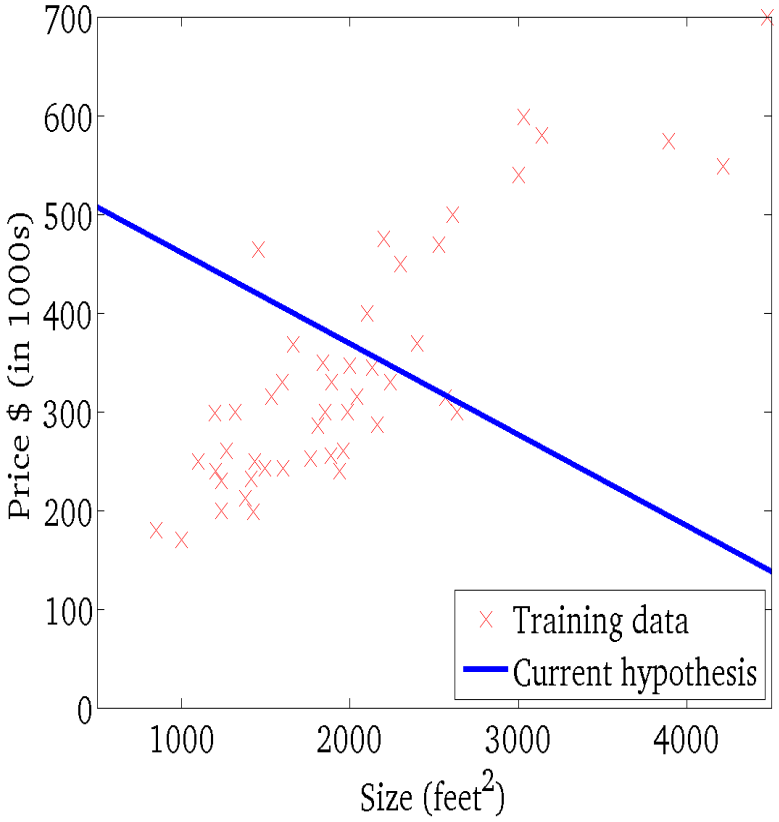
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

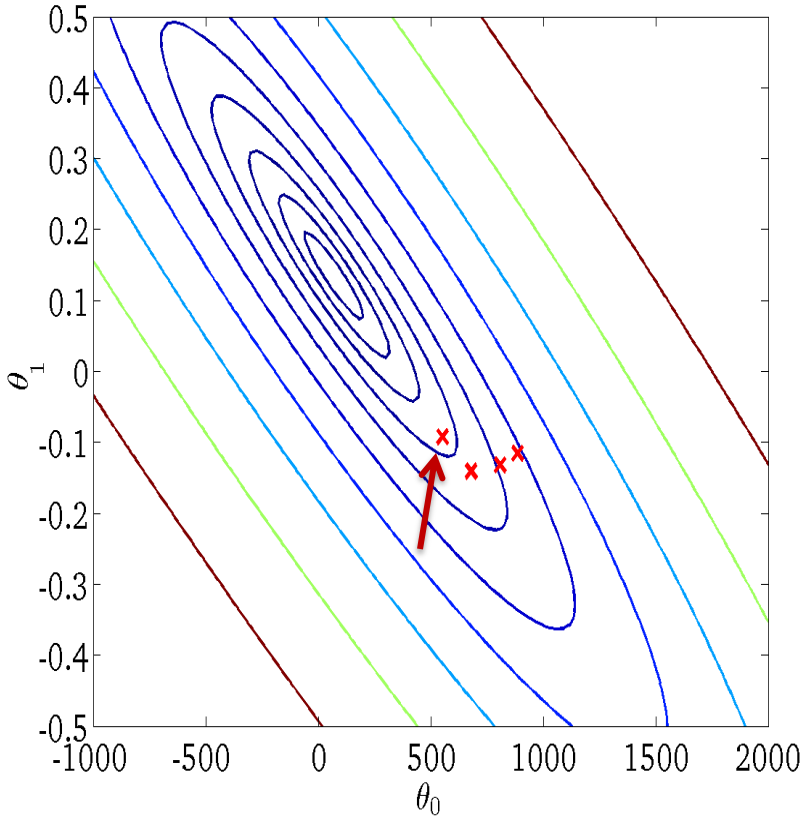
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

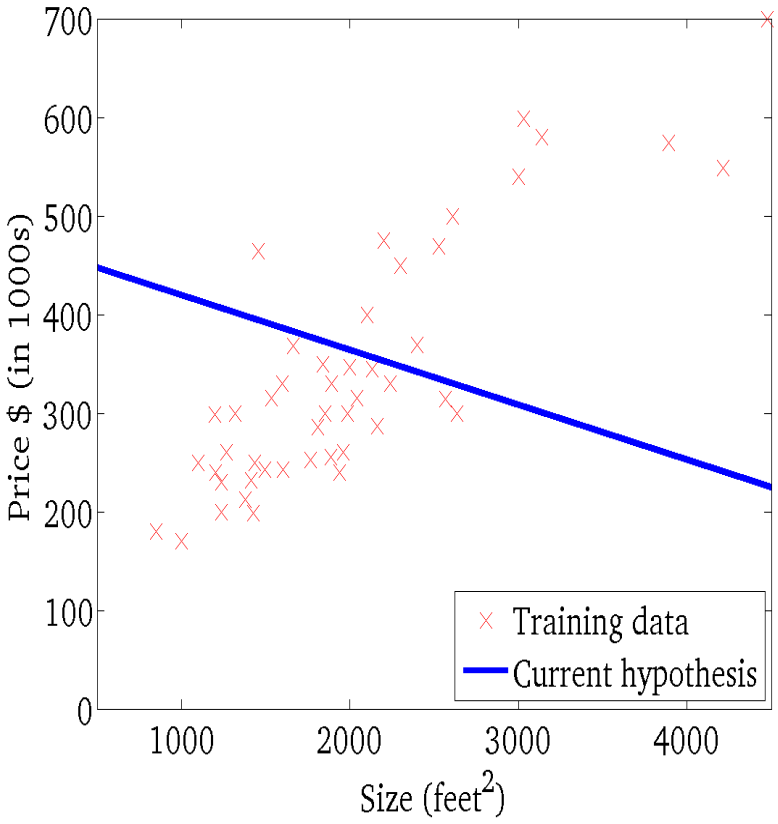
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

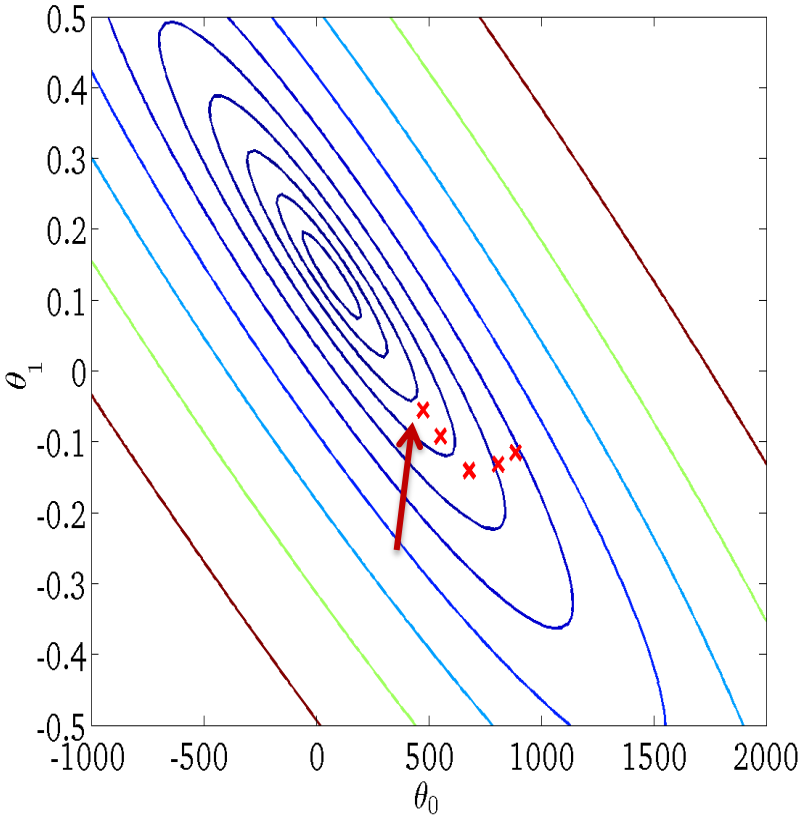
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

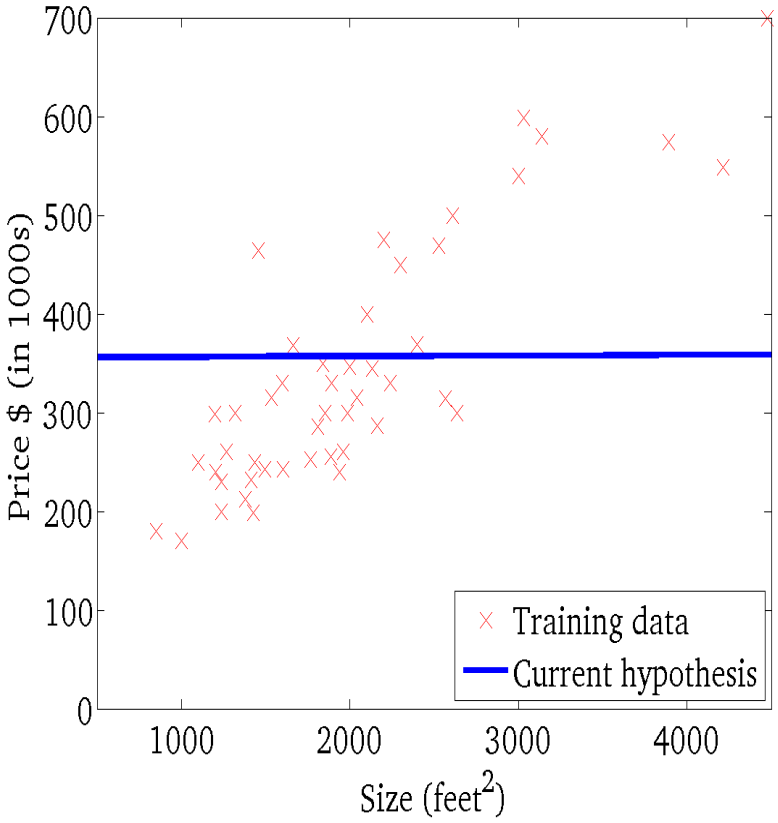
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

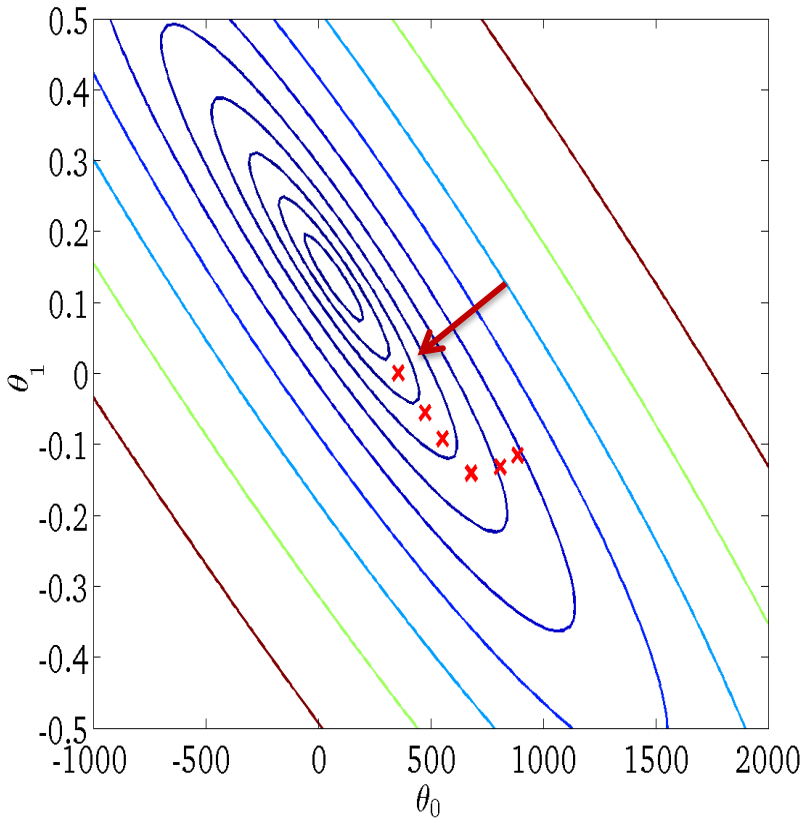
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

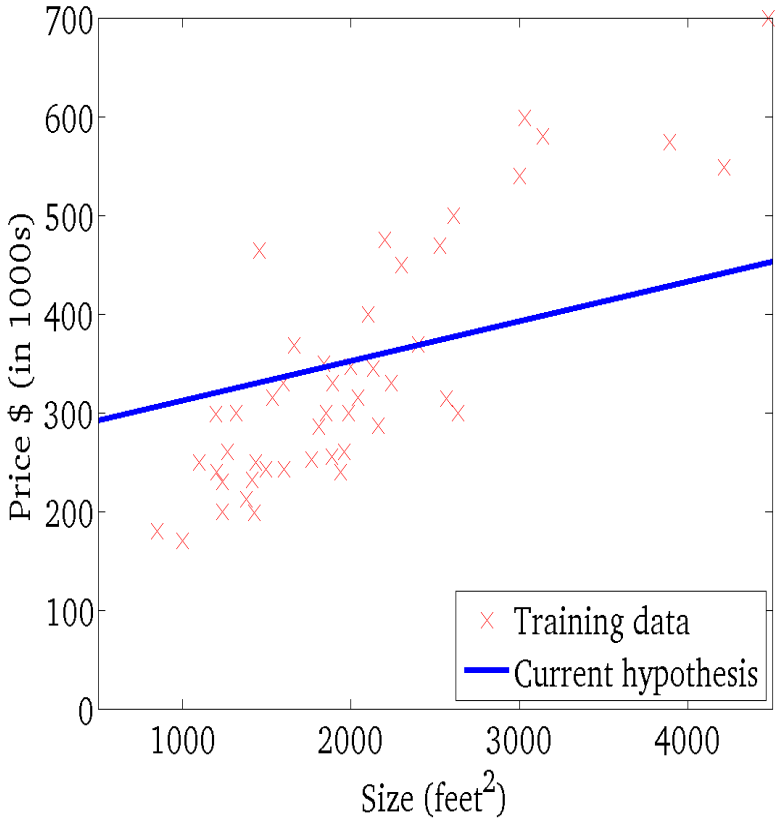
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

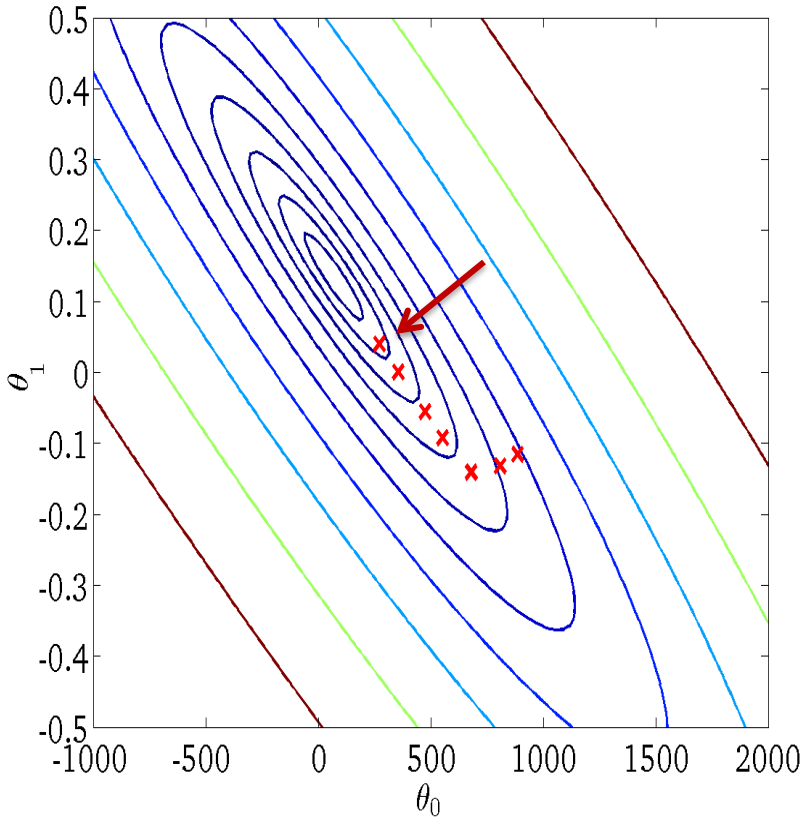
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

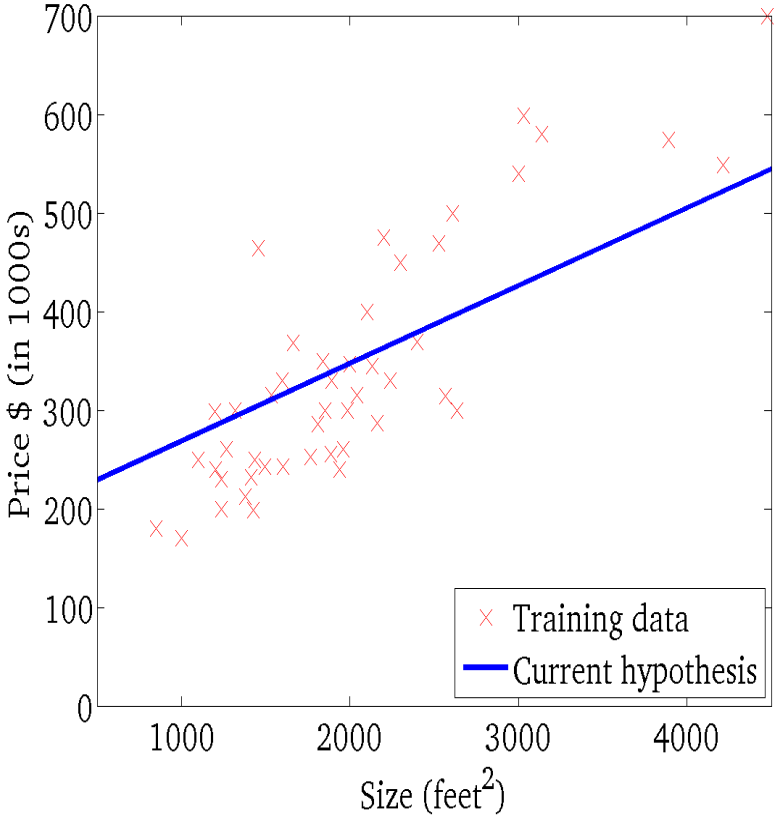
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

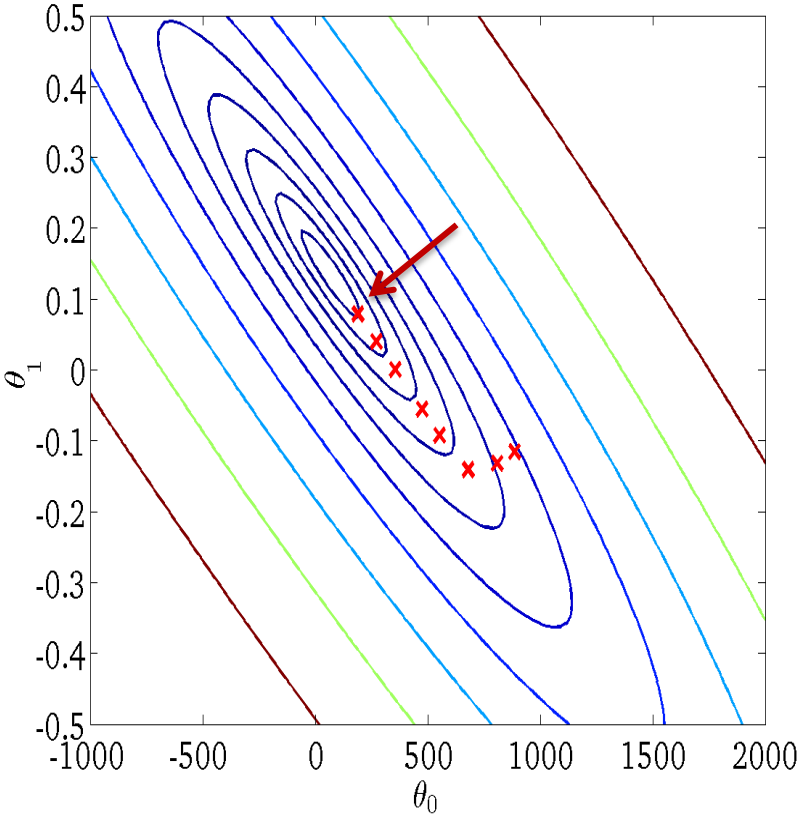
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

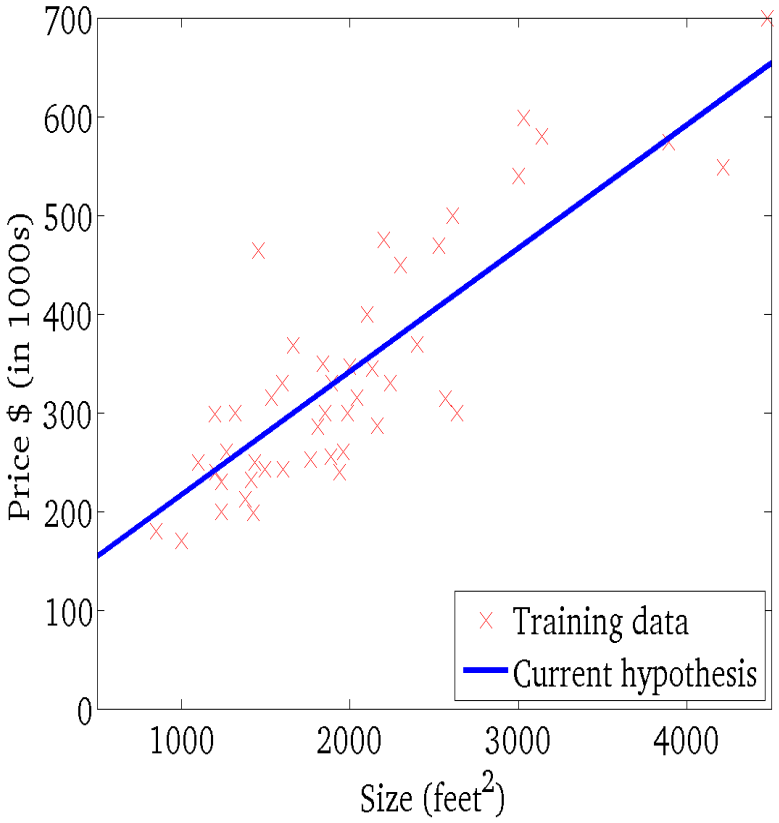
(function of the parameters θ_0, θ_1)



Gradient Descent in Action

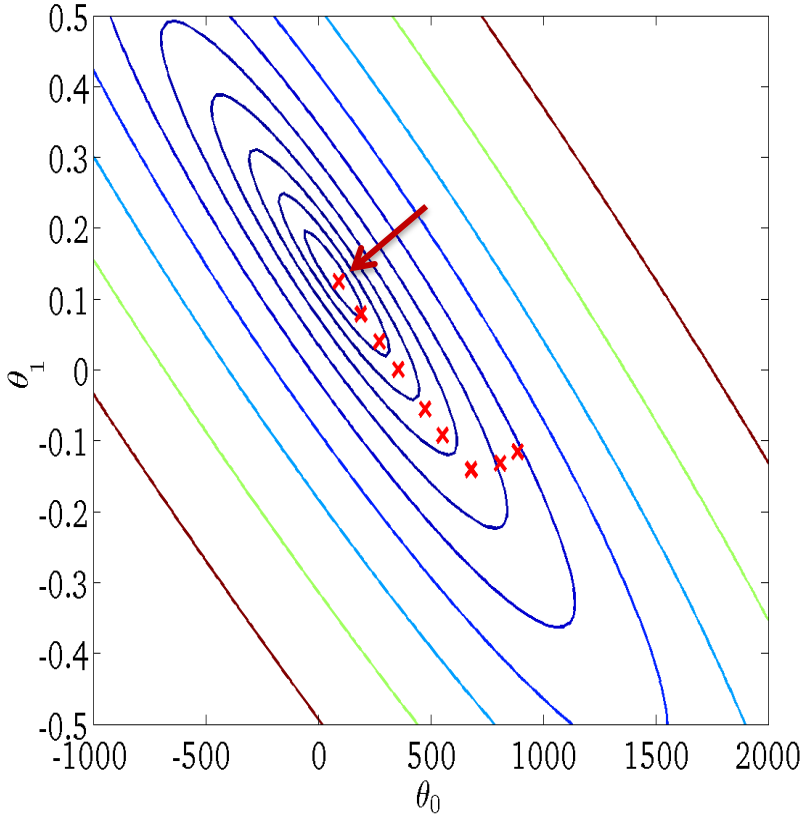
$h_\theta(x)$

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

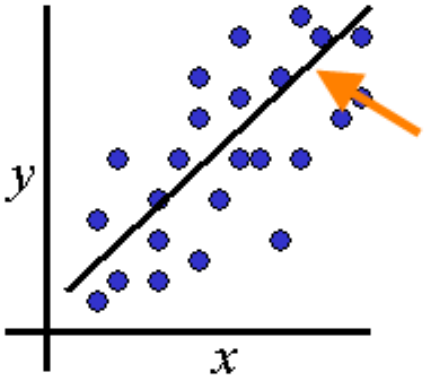


$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Summary: Univariate Linear Regression



Target: Find best fitting line $h_\theta(x) = \theta_0 + \theta_1 x$

minimize $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$

Squared Error: (Estimated – Actual)^2

Repeat until convergence {

$$\theta_0 = \theta_0 - \alpha \cdot \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0}$$

$$\theta_1 = \theta_1 - \alpha \cdot \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1}$$

}

Derivative (or Gradient)

Learning Rate (or Step Size)

Gradient Descent Algorithm

Multivariate Linear Regression

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots$$

$n = 1$

GPA (x)	1st Salary (y)
3.5	20000
2.1	10000
…	…

m

$n = 3$

GPA (x_1)	# of Exchanges (x_2)	Age (x_3)	1st Salary (y)
3.5	1	23	20000
2.1	2	22	10000
…	…	…	…

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

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$

$$J(\theta_0, \theta_1, \cdots) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Example with Training data

The training data contain some example measurements of the profit gained by opening an outlet in the cities with the population ranging between 30,000 and 100,000. The y-values are the profit measured in USD, and the x-values are the populations of the city. Each city population and profit tuple constitutes one training example in training dataset.

City Population (10^4) x	Profit (10^4) y
6.4862	6.5987
5.5277	9.1302
8.5186	13.662
7.0032	11.854

Initialize Gradient Descent

Use training dataset to develop a linear regression model and solve it by using gradient descent algorithm. Find the values of θ_0, θ_1, and cost function J in the first two iterations. ($\theta_0 = 0, \theta_1 = 0, \alpha = 0.01$)

Initial setting:

$\theta_0[0] = \theta_1[0] = 0$

Iteration 1

- Update θ_0:

$$\frac{\partial J}{\partial \theta_0}[0] = \frac{1}{4}\sum_{i=1}^{4}\left(\theta_0[0] + \theta_1[0]x^{(i)} - y^{(i)}\right)$$

$$= 1/4\left[-y^{(1)} - y^{(2)} - y^{(3)} - y^{(4)}\right]$$

$$= 1/4\left[-6.5987 - 9.1302 - 13.662 - 11.854\right] = -10.311$$

$$\theta_0[1] = \theta_0[0] - 0.01\frac{\partial J}{\partial \theta_0}[0] = 0.10311$$

- Update θ_1:

$$\frac{\partial J}{\partial \theta_1}[0] = \frac{1}{4}\sum_{i=1}^{4}\left(\theta_0[0] + \theta_1[0]x^{(i)} - y^{(i)}\right)x^{(i)}$$

$$= 1/4\left[-y^{(1)}x^{(1)} - y^{(2)}x^{(2)} - y^{(3)}x^{(3)} - y^{(4)}\,x^{(4)}\right]$$

$$= 1/4\left[-6.5987\times 6.4862 - 9.1302\times 5.5277 - 13.662\times 8.5186\right.$$

$$\left. - 11.854\times 7.0032\right] = -73.167$$

$$\theta_1[1] = \theta_1[0] - 0.01\frac{\partial J}{\partial \theta_1}[0] = 0.73167$$

Cost Update

- Update $J(\theta_0, \theta_1)$:

$$J(\theta_0[0], \theta_1[0]) = \frac{1}{8}\sum_{i=1}^{4}\left(\theta_0[0] + \theta_1[0]x^{(i)} - y^{(i)}\right)^2$$

$$= 1/8\left[\left(y^{(1)}\right)^2 + \left(y^{(2)}\right)^2 + \left(y^{(3)}\right)^2 + \left(y^{(4)}\right)^2\right]$$

$$= 1/8\left[(6.5987)^2 + (9.1302)^2 + (13.662)^2 + (11.854)^2\right] = 56.759$$

$$J(\theta_0[1], \theta_1[1]) = \frac{1}{8}\sum_{i=1}^{4}\left(\theta_0[1] + \theta_1[1]x^{(i)} - y^{(i)}\right)^2$$

$$= 1/8\left[(0.10311+0.73167\times6.4862-6.5987)^2\right.$$
$$+ (0.10311+0.73167\times5.5277-9.1302)^2$$
$$+ (0.10311+0.73167\times8.5186-13.662)^2$$
$$\left.+ (0.10311+0.73167\times7.0032-11.854)^2\right] = 15.864$$

Iteration 2

- Update θ_0:

$$\frac{\partial J}{\partial \theta_0}[1] = \frac{1}{4}\sum_{i=1}^{4}\left(\theta_0[1] + \theta_1[1]x^{(i)} - y^{(i)}\right) = \cdots$$

$$\theta_0[2] = \theta_0[1] - 0.01\,\frac{\partial J}{\partial \theta_0}[1] = \cdots$$

- Update θ_1:

$$\frac{\partial J}{\partial \theta_1}[1] = \frac{1}{4}\sum_{i=1}^{4}\left(\theta_0[1] + \theta_1[1]x^{(i)} - y^{(i)}\right)x^{(i)} = \cdots$$

$$\theta_1[2] = \theta_1[1] - 0.01\,\frac{\partial J}{\partial \theta_1}[1] = \cdots$$

- Update $J(\theta_0, \theta_1)$:

$$J(\theta_0[2], \theta_1[2]) = \frac{1}{8}\sum_{i=1}^{4}\left(\theta_0[2] + \theta_1[2]x^{(i)} - y^{(i)}\right)^2 = \cdots$$

Linear Regression by Linear Algebra

- Minimizing function:

$$\min_{\theta_0,\theta_1,\cdots} J(\theta_0,\theta_1,\cdots)$$

$$= \min_{\theta_0,\theta_1,\cdots} \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

$$= \min_{\theta_0,\theta_1,\cdots} \frac{1}{2m}\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots - y^{(i)}\right)^2$$

- Necessary Condition: $\frac{\partial J(\theta_0,\theta_1,\cdots)}{\partial \theta_i} = 0, 0 \le i \le n$

Normal Equation

$$
\begin{cases}
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_0} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) = 0 \\
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_1} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) x_1^{(i)} = 0 \\
\cdots \\
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_n} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) x_n^{(i)} = 0
\end{cases}
$$

Normal Equation for Univariate Case

- Notions:

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}, \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}, \overline{xy} = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}y^{(i)}, \overline{x^2} = \frac{1}{m}\sum_{i=1}^{m} (x^{(i)})^2$$

- Equations:

$$\begin{cases} \frac{1}{m}\sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) = \theta_0 + \bar{x}\theta_1 - \bar{y} = 0 \\ \frac{1}{m}\sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)x^{(i)} = \theta_0\bar{x} + \overline{x^2}\theta_1 - \overline{xy} = 0 \end{cases}$$

- Solutions:

$$\begin{cases} \theta_0 = \frac{\bar{y} * \overline{x^2} - \bar{x} * \overline{xy}}{\overline{x^2} - (\bar{x})^2} \\ \theta_1 = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - (\bar{x})^2} \end{cases}$$

Same Example

The training data contain some example measurements of the profit gained by opening an outlet in the cities with the population ranging between 30,000 and 100,000. The y-values are the profit measured in USD, and the x-values are the populations of the city. Each city population and profit tuple constitutes one training example in training dataset.

City Population (10^4) x	Profit (10^4) y
6.4862	6.5987
5.5277	9.1302
8.5186	13.662
7.0032	11.854

Solution with Normal Equation

$$\bar{x} = \frac{1}{4}(6.4862+5.5277 + 8.5186 + 7.0032) = 6.8839$$

$$\bar{y} = \frac{1}{4}(6.5987+9.1302 + 13.662 + 11.854) = 10.311$$

$$\overline{xy} = \frac{1}{4}(6.4862\times6.5987+5.5277\times9.1302 + 8.5186\times13.662 + 7.0032\times11.854)$$
$$= 73.167$$

$$\overline{x^2} = \frac{1}{4}(6.4862\times6.4862+5.5277\times5.5277 + 8.5186\times8.5186 + 7.0032\times7.0032)$$
$$= 48.559$$

$$\begin{cases} \theta_0 + \bar{x}\theta_1 - \bar{y} = 0 = \theta_0 + 6.8839\theta_1 - 10.311 = 0 \\ \bar{x}\theta_0 + \overline{x^2}\theta_1 - \overline{xy} = 6.8839\theta_0 + 48.55\theta_1 - 73.167 = 0 \end{cases} \Rightarrow \begin{cases} \theta_0 = -2.5471 \\ \theta_1 = 1.8679 \end{cases}$$

$$J(\theta_0, \theta_1) = 1/8\,[(-2.5471+1.8679\times6.4862-6.5987)^2 + \cdots] = 1.5598$$

Basic Knowledge about Matrix

$$
\begin{array}{ccccccc}
a_{11}x_1 & + & a_{12}x_2 & + \cdots + & a_{1n}x_n & = b_1 \\
a_{21}x_1 & + & a_{22}x_2 & + \cdots + & a_{2n}x_n & = b_2 \\
\vdots & & \vdots & & \vdots & \vdots \\
a_{m1}x_1 & + & a_{m2}x_2 & + \cdots + & a_{mn}x_n & = b_m
\end{array}
$$

$$
A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}
$$

$$
A\mathbf{x} = \mathbf{b}
$$
$$
\mathbf{x} = A^{-1}\mathbf{b}
$$

Multivariate Linear Regression

$$
\begin{cases}
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_0} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) = 0 \\
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_1} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) x_1^{(i)} = 0 \\
\cdots \\
\dfrac{\partial J(\theta_0, \theta_1, \cdots)}{\partial \theta_n} = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}\left(\theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} - y^{(i)}\right) x_n^{(i)} = 0
\end{cases}
$$

$$
\left(\sum_{i=1}^{m} 1 \cdot 1\right)\theta_0 + \left(\sum_{i=1}^{m} x_1^{(i)} \cdot 1\right)\theta_1 + \left(\sum_{i=1}^{m} x_2^{(i)} \cdot 1\right)\theta_2 + \cdots + \left(\sum_{i=1}^{m} x_n^{(i)} \cdot 1\right)\theta_n = \sum_{i=1}^{m} y^{(i)} \cdot 1
$$

$$
\left(\sum_{i=1}^{m} 1 \cdot x_1^{(i)}\right)\theta_0 + \left(\sum_{i=1}^{m} x_1^{(i)} x_1^{(i)}\right)\theta_1 + \left(\sum_{i=1}^{m} x_2^{(i)} x_1^{(i)}\right)\theta_2 + \cdots + \left(\sum_{i=1}^{m} x_n^{(i)} x_1^{(i)}\right)\theta_n = \sum_{i=1}^{m} y^{(i)} x_1^{(i)}
$$

$$
\cdots
$$

$$
\left(\sum_{i=1}^{m} 1 \cdot x_n^{(i)}\right)\theta_0 + \left(\sum_{i=1}^{m} x_1^{(i)} x_n^{(i)}\right)\theta_1 + \left(\sum_{i=1}^{m} x_2^{(i)} x_n^{(i)}\right)\theta_2 + \cdots + \left(\sum_{i=1}^{m} x_n^{(i)} x_n^{(i)}\right)\theta_n = \sum_{i=1}^{m} y^{(i)} x_n^{(i)}
$$

Matrix Form of Normal Equation

$$A = \begin{bmatrix} \sum_{i=1}^{m} 1 \cdot 1 & \sum_{i=1}^{m} x_1^{(i)} \cdot 1 & \sum_{i=1}^{m} x_2^{(i)} \cdot 1 & \cdots & \sum_{i=1}^{m} x_n^{(i)} \cdot 1 \\ \sum_{i=1}^{m} 1 \cdot x_1^{(i)} & \sum_{i=1}^{m} x_1^{(i)} x_1^{(i)} & \sum_{i=1}^{m} x_2^{(i)} x_1^{(i)} & \cdots & \sum_{i=1}^{m} x_n^{(i)} x_1^{(i)} \\ & & \cdots & & \\ \sum_{i=1}^{m} 1 \cdot x_n^{(i)} & \sum_{i=1}^{m} x_1^{(i)} x_n^{(i)} & \sum_{i=1}^{m} x_2^{(i)} x_n^{(i)} & \cdots & \sum_{i=1}^{m} x_n^{(i)} x_n^{(i)} \end{bmatrix} \quad b = \begin{bmatrix} \sum_{i=1}^{m} y^{(i)} \cdot 1 \\ \sum_{i=1}^{m} y^{(i)} x_1^{(i)} \\ \cdots \\ \sum_{i=1}^{m} y^{(i)} x_n^{(i)} \end{bmatrix}$$

$$A\boldsymbol{\theta} = \mathbf{b}$$
$$\boldsymbol{\theta} = A^{-1}\mathbf{b}$$

Matrix Form of Normal Equation

- Matrix Form of solution θ in terms of matrics X and y ?

$$X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_j^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_j^{(2)} & \cdots & x_n^{(2)} \\ & & & \vdots & & \\ 1 & x_1^{(m)} & \cdots & x_j^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, \quad y = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Features

Target variables

$$(X^T X)\theta = X^T y$$
$$\theta = (X^T X)^{-1} X^T y$$