Faculty of Computer Science and Engineering
Ho Chi Minh City University of Technology

Chapter 3
Data Regression

Assoc. prof. TRAN MINH QUANG
quangtran@hcmut.edu.vn
http://researchmap.jp/quang

2020/4/21

CONTENT

1. Introduction
2. Linear regression Hồi qui tuyến tính
3. Non-Linear regression
4. Applications
5. Problems with regression
6. Summary

REFERENCES

[1] Jiawei Han, Micheline Kamber, and Jian Pei, "Data Mining: Concepts and Techniques", 3rd Edition, Morgan Kaufmann Publishers, 2012.

[2] David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", MIT Press, 2001.

[3] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[4] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[5] ZhaoHui Tang, Jamie MacLennan, "Data Mining with SQL Server 2005", Wiley Publishing, 2005.

[6] Oracle, "Data Mining Concepts", B28129-01, 2008.

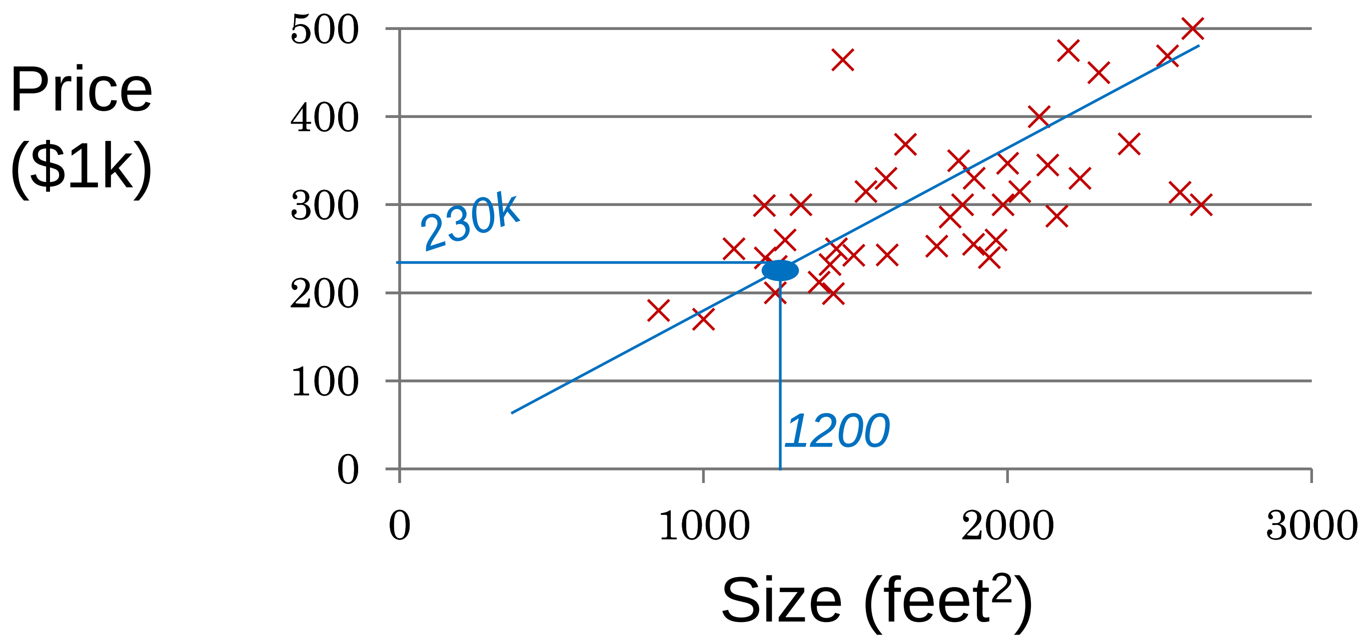
[7] Oracle, "Data Mining Application Developer's Guide", B28131-01, 2008.

[8] Ian H.Witten, Eibe Frank, "Data mining : practical machine learning tools and techniques", 2nd Edition, Elsevier Inc, 2005.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Oded Maimon, Lior Rokach, "Data Mining and Knowledge Discovery Handbook", 2nd Edition, Springer Science + Business Media, LLC 2005, 2010.

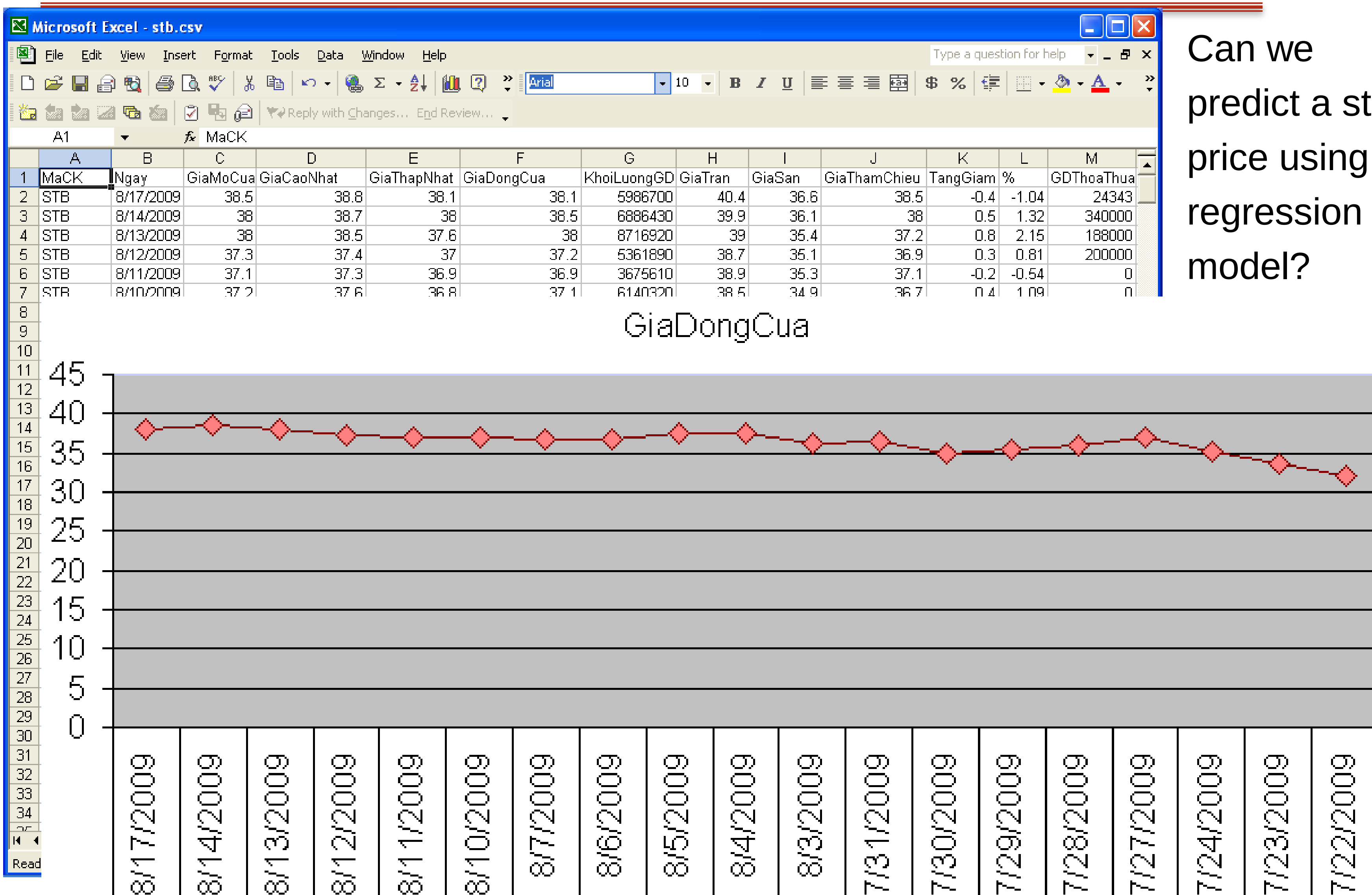
1. INTRODUCTION



+ Can we model the house price distribution based on their sizes ?

+ Can we predict a house price based on its size?

1. INTRODUCTION



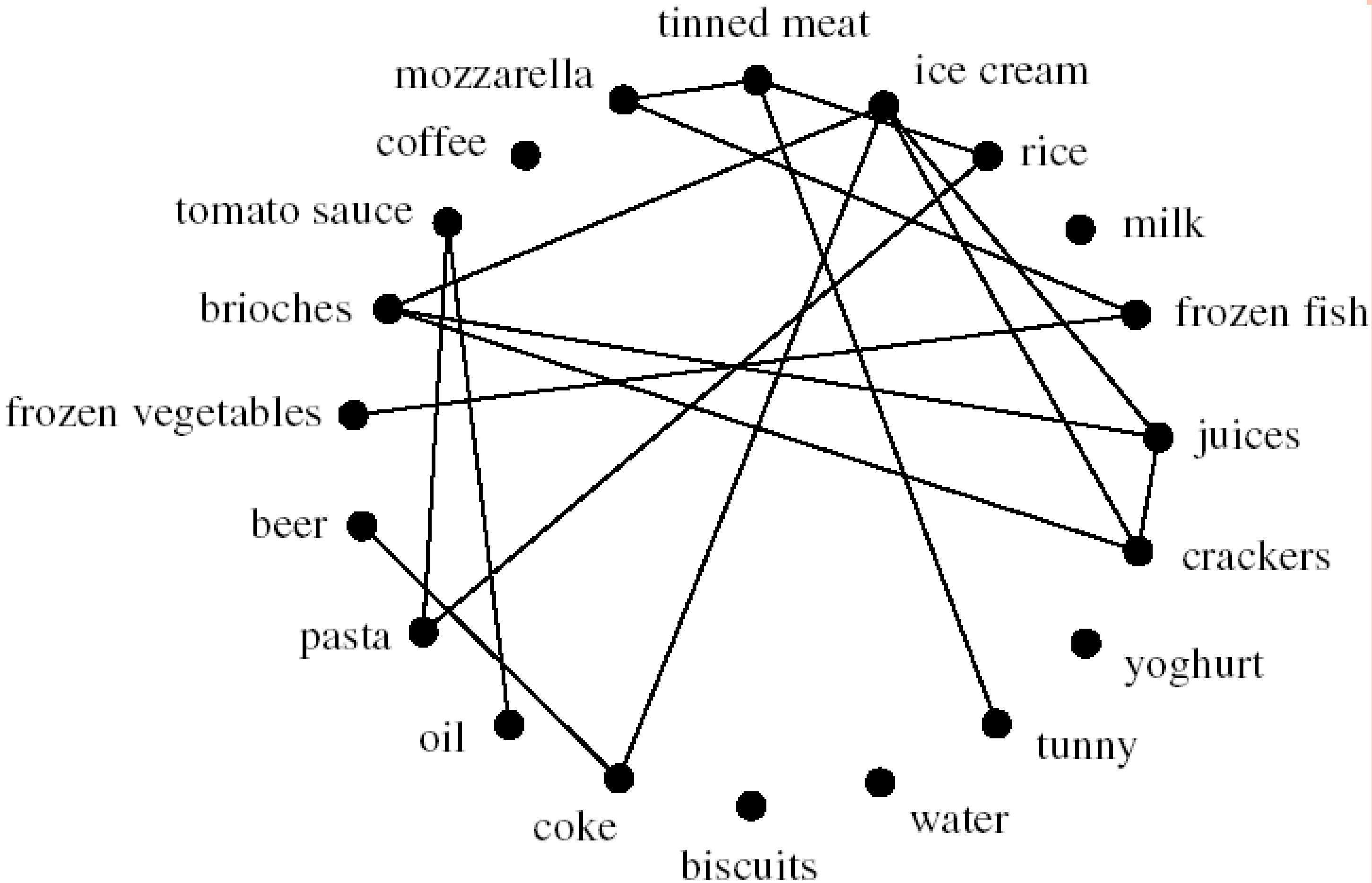
MaCK	Ngay	GiaMoCua	GiaCaoNhat	GiaThapNhat	GiaDongCua	KhoiLuongGD	GiaTran	GiaSan	GiaThamChieu	TangGiam	%	GDThoaThua
STB	8/17/2009	38.5	38.8	38.1	38.1	5986700	40.4	36.6	38.5	-0.4	-1.04	24343
STB	8/14/2009	38	38.7	38	38.5	6886430	39.9	36.1	38	0.5	1.32	340000
STB	8/13/2009	38	38.5	37.6	38	8716920	39	35.4	37.2	0.8	2.15	188000
STB	8/12/2009	37.3	37.4	37	37.2	5361890	38.7	35.1	36.9	0.3	0.81	200000
STB	8/11/2009	37.1	37.3	36.9	36.9	3675610	38.9	35.3	37.1	-0.2	-0.54	0
STB	8/10/2009	37.2	37.6	36.8	37.1	6140320	38.5	34.9	36.7	0.4	1.09	0

Can we predict a stock price using regression model?

1. Introduction

The market basket analysis problem

→ Can we find out association rules between products in transactions?



1. Introduction

- Analyzing the factors that impact on the quality of e-banking services (based on surveys from users)
  - Easy to use (+0.209)
  - Fast response (+0.261)
  - The ability to link with other billing services (+0.199)
  - Feelings of individuality (+0.15)
  - Privacy and security issues (-0.25)
  - …

1. INTRODUCTION

- Regression
  - J. Han et al (2001, 2006): Regression is a statistic mechanism that allows predicting real/numeric and continuous values
  - Wiki (2009): Regression analysis is a statistic mechanism that allows estimating the correlation between independent variables
  - R. D. Snee (1977): Regression is a statistic mechanism in data analytics and building models from experiments, it allows prediction, control, and learning the rules to which data is generated.
  - **Regression: Numeric data prediction (real-valued output)**
  - **Classification: "prediction" for discrete values**

1. INTRODUCTION

- Regression model: Describe the relationship between a set of predictors/independent variables and one or some responses/dependent variables

- Regression equation

$$\mathbf{Y} = f(\mathbf{X}, \theta)$$

X: a set of predictors/independent variables; describes the changes of responses/dependent variables Y

Y: responses/dependent variables; Describes the interesting facts/events

θ: Regression coefficients; Describes the relative effects of X on Y

1. Introduction

- Categories:
  - Linear v.s nonlinear
    - ✓ Linear in parameters: Linear association between parameters that affect Y
    - ✓ Nonlinear in parameters: Non-linear association between parameters that affect Y
  - Single variable v.s multiple variables
    - ✓ Single: $X = (X_1)$ v.s. Multiple: $X = (X_1, X_2, \ldots, X_k)$
  - Parametric v.s nonparametric and semiparametric
  - Symmetric v.s asymmetric
    - ✓ Symmetric: descriptive regression models (e.g., log-linear models)
    - ✓ Asymmetric: predictive regression models (e.g., generalized linear models)

1. Introduction

- Parametric, nonparametric, and semiparametric
  - Parametric: regression models with finite parameters
  - Nonparametric: regression models with infinite parameters
  - Semiparametric: regression models with finite interesting parameters

Regression model	Description
Parametric	$Y = \theta_0 + \theta_1 * X$
Nonparametric	$Y = \theta_0 + f(X)$
Semiparametric	$Y = \theta_0 + \theta_1 * X_1 + f(X_2)$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2. LINEAR REGRESSION

- Single variable (Univariate)
- Multiple variables (Multivariate)

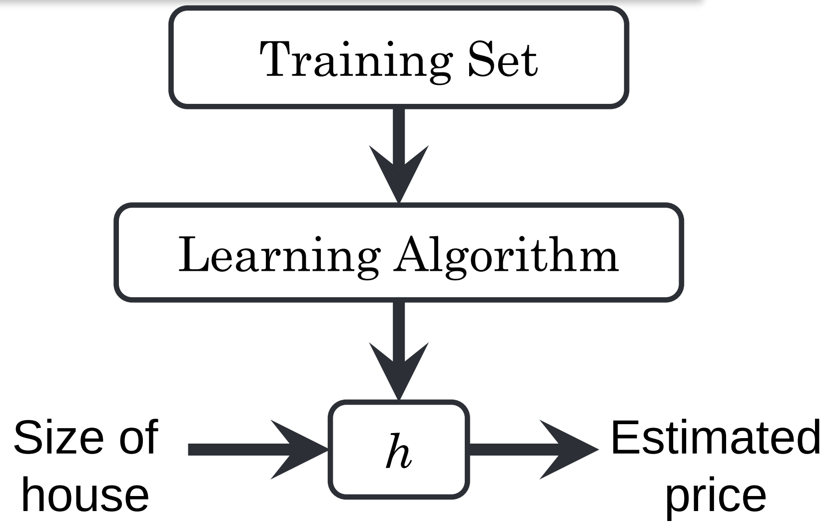
2.1. Univariate Linear Regression

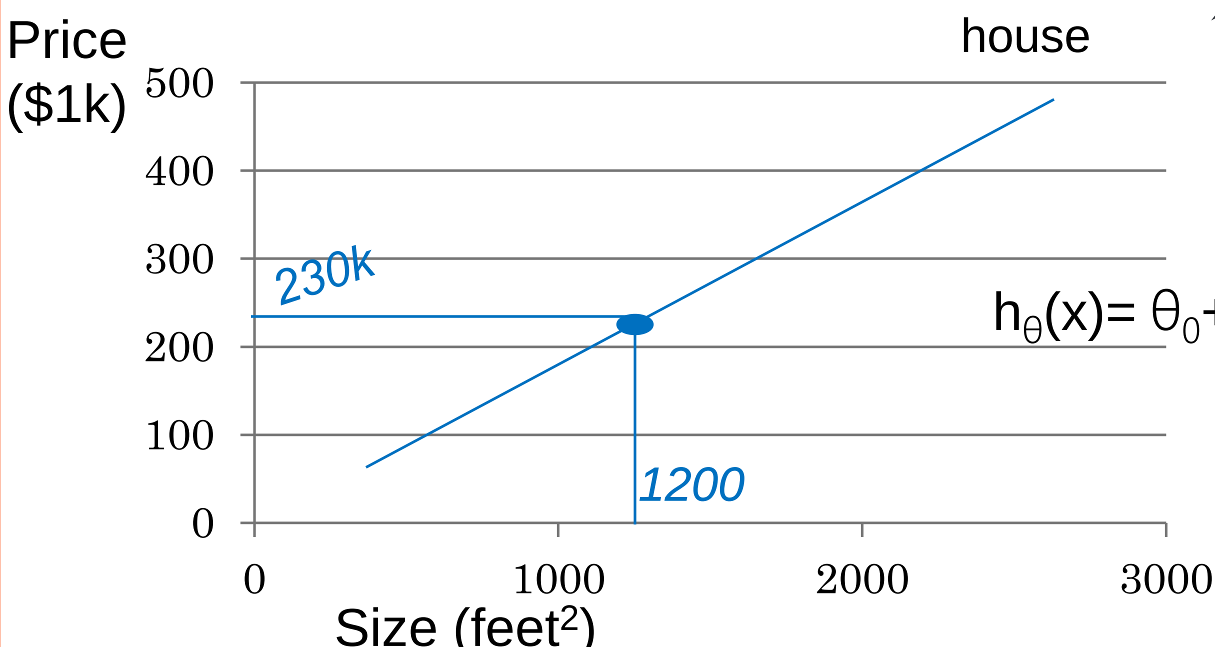
- Notations
  - N: size of training examples
  - x: input variable/feature
  - y: output/target variable
  - $(x^{(i)}, y^{(i)})$: i^{th} learning sample
  - $(x^{(1)}, y^{(1)}) = (2100, 450)$

Size feet² (x)	Price ($1k) (y)
2100	450
1416	232
1534	315
852	178
…	…

2.1. Univariate Linear Regression

- *Hypothesis (h)*
- $y = h(x)$; h is a mapping from x to y
- How to model h?





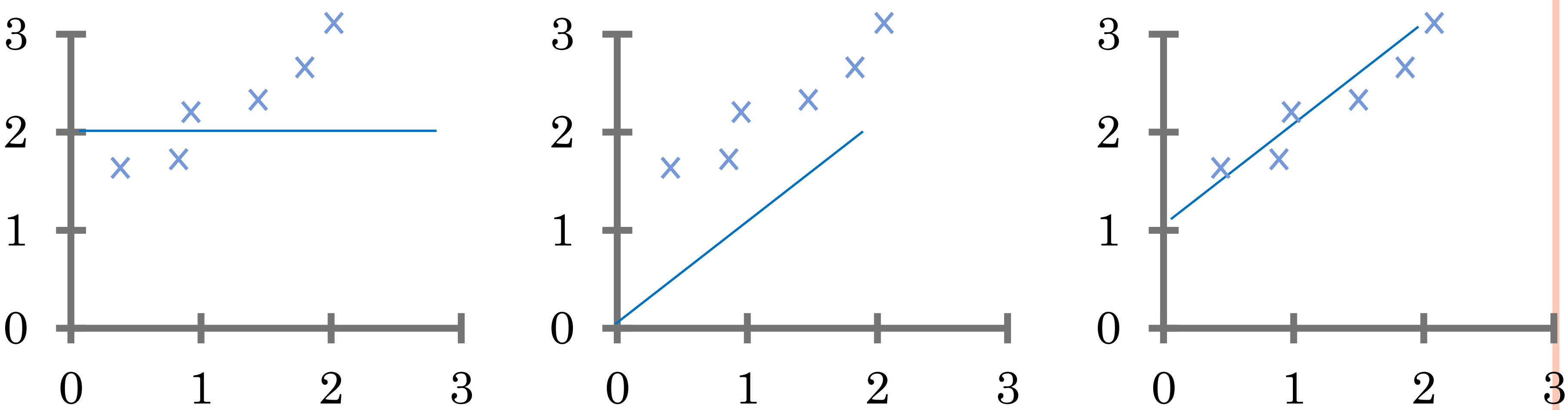
$h_\theta(x) = \theta_0 + \theta_1 x$

univariate

2.1. UNIVARIATE LINEAR REGRESSION

- *Hypothesis (h)*: $h_\theta(x) = \theta_0 + \theta_1 x$ => identify θ_i ?
- Method: "try_and_error", evaluate the ability of the regression line in describing sample data.

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



$\theta_0 = 2$
$\theta_1 = 0$

$\theta_0 = 0$
$\theta_1 = 0.5$

$\theta_0 = 1$
$\theta_1 = 0.5$

2.1. UNIVARIATE LINEAR REGRESSION

- Chose (θ_0, θ_1) so that $h_\theta(x^{(i)}) \simeq y^{(i)}$; $i=1\ldots N$
  - residual/prediction error

$$e = h_\theta(x^{(i)}) - y^{(i)}$$

  - MSE

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

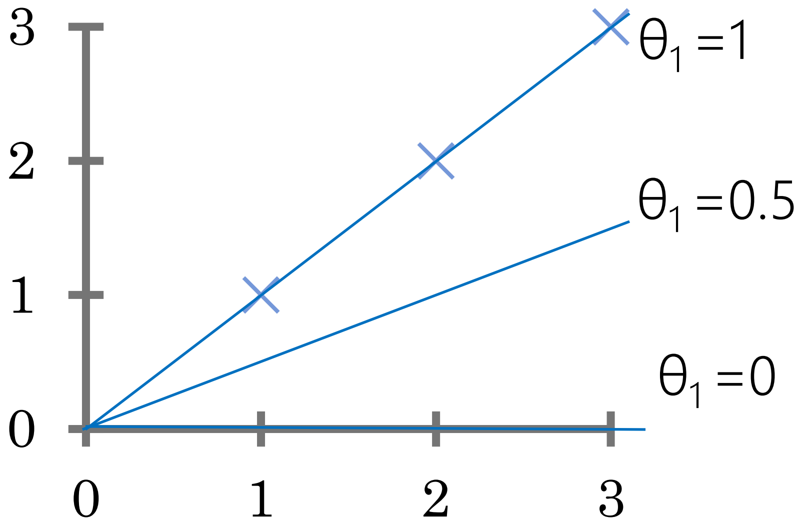
- Cost function $J(\theta_0, \theta_1)$ => ***minimize***

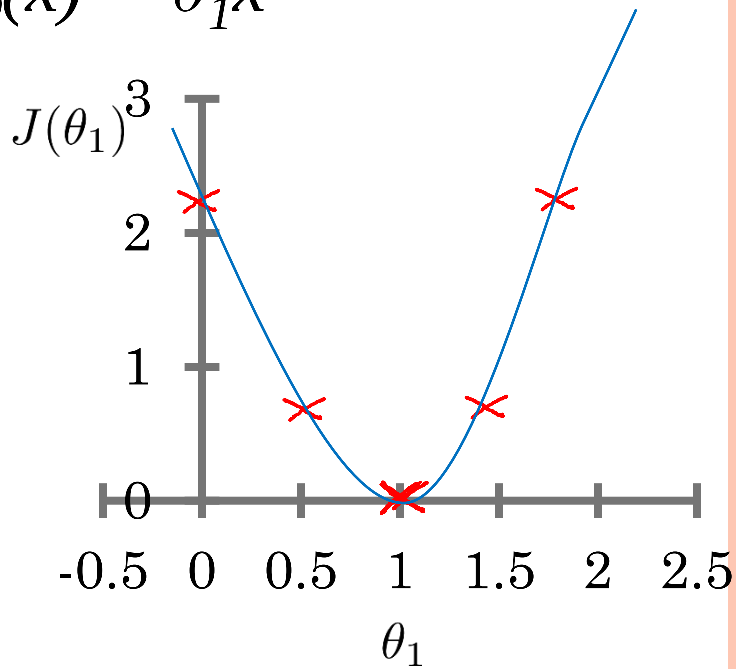
$$J(\theta_0, \theta_1) = \frac{1}{2N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2.1. Univariate linear regression

- Examine a simple case: $\theta_0=0$, $h_\theta(x) = \theta_1 x$



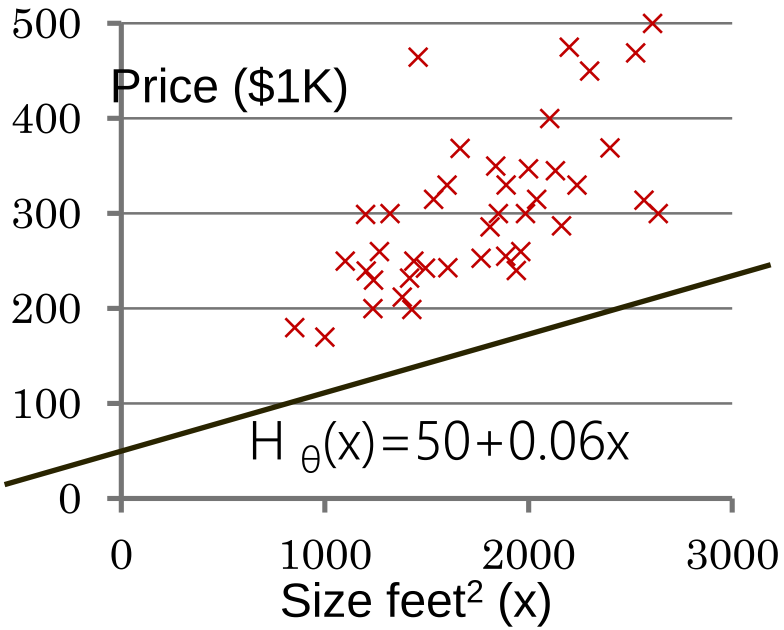


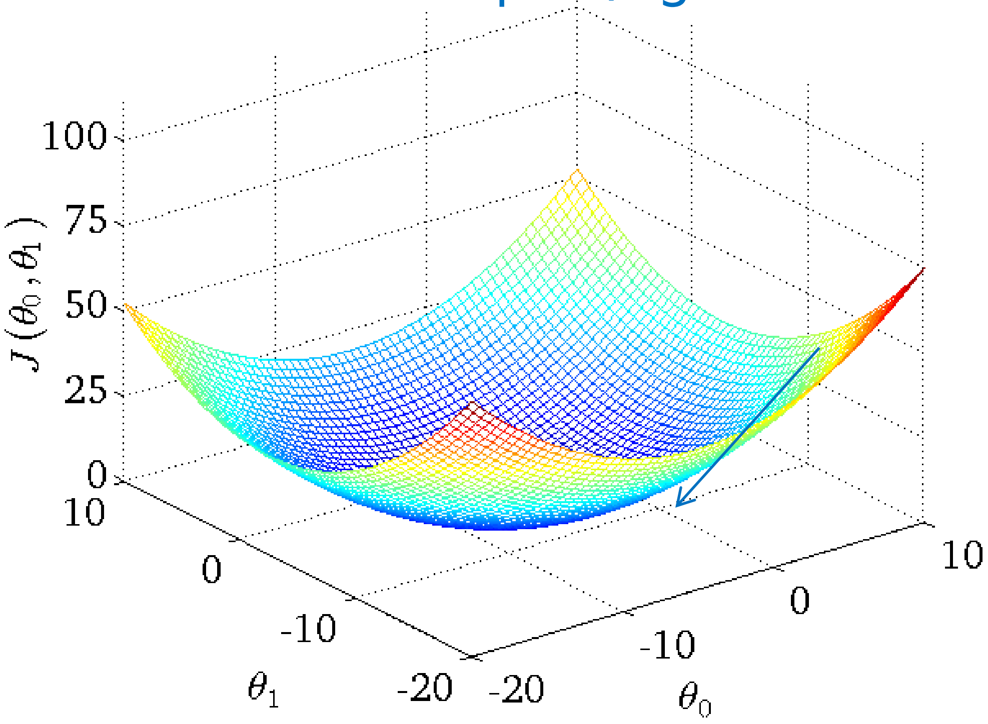
- $\theta_1 = 1 \Rightarrow J(\theta_1) = 0$; $\theta_1 = 0.5 \Rightarrow J(\theta_1) = 0.58$; $\theta_1 = 0 \Rightarrow J(\theta_1) = 2.3$

2.1. UNIVARIATE LINEAR REGRESSION

- An example with $h_\theta(x) = \theta_0+\theta_1 x$

Contour plots/figures

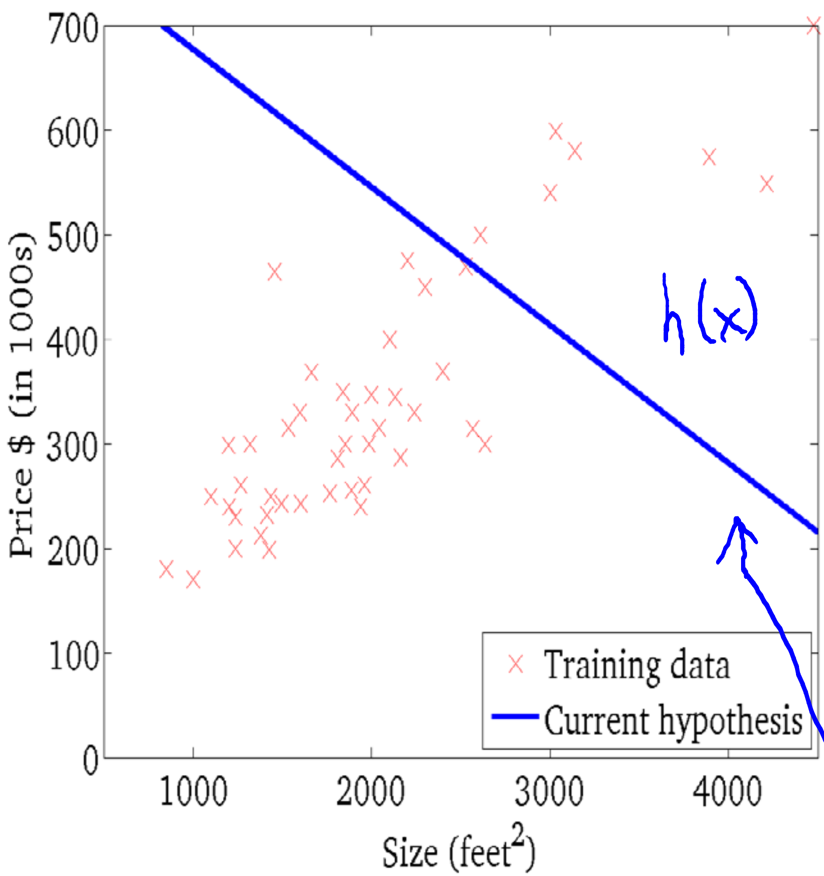




Source: Andrew Ng

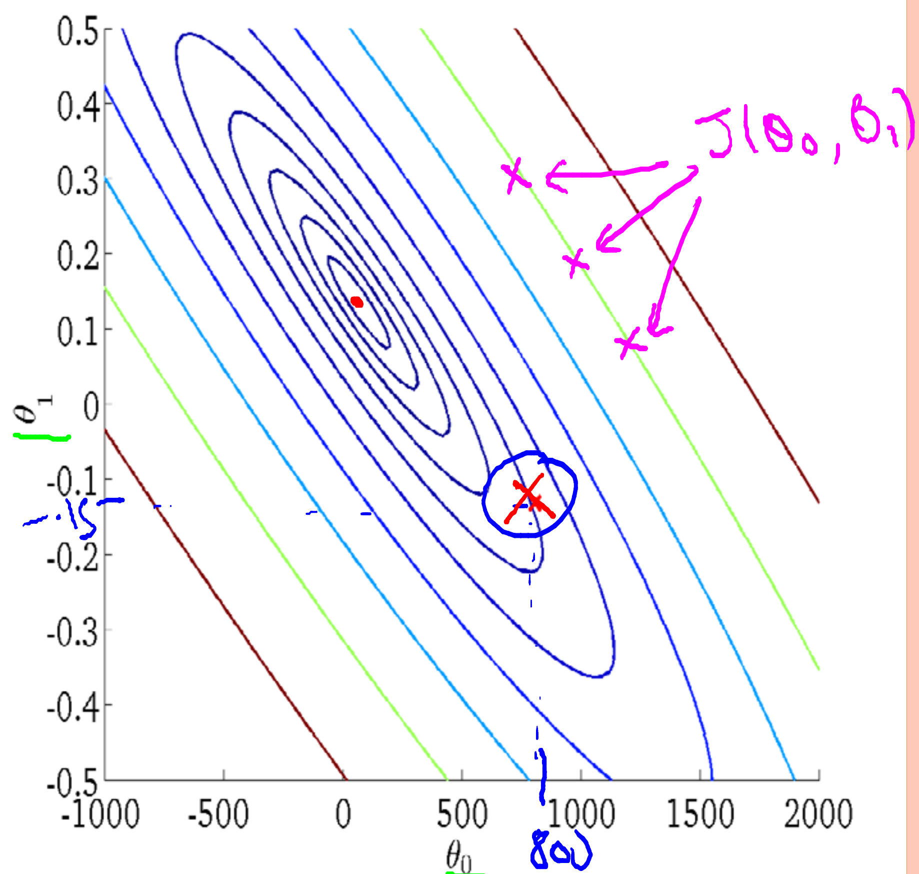
$h_\theta(x)$

(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)

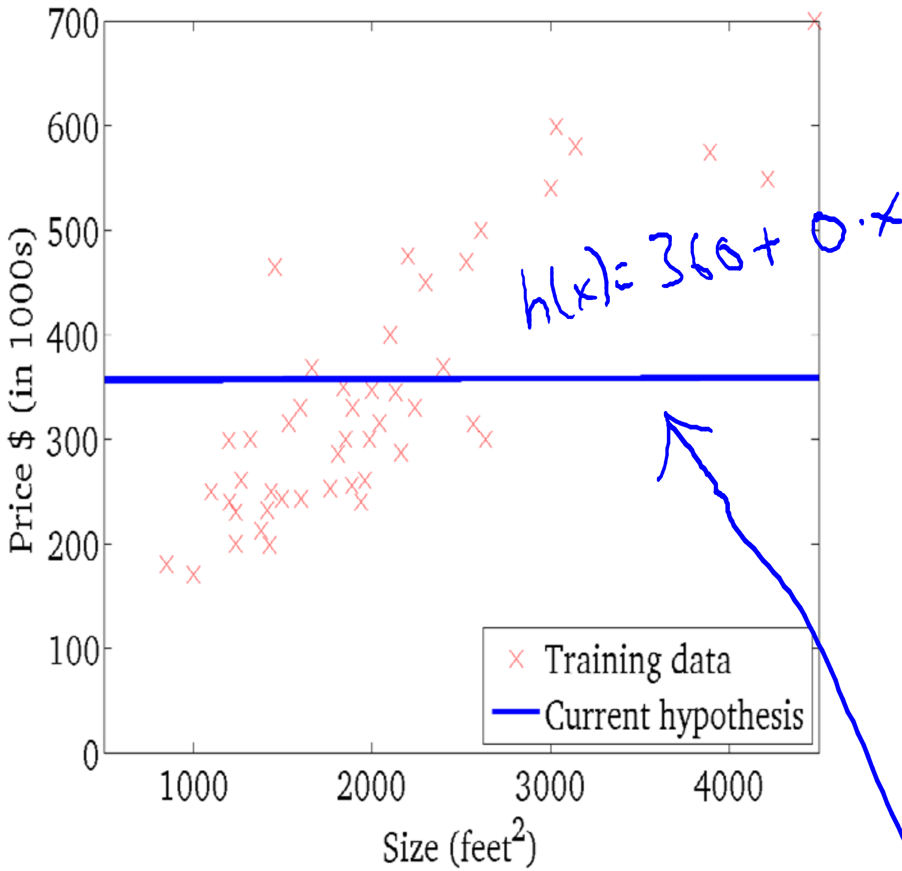


θ_0, θ_1

Source: Andrew Ng

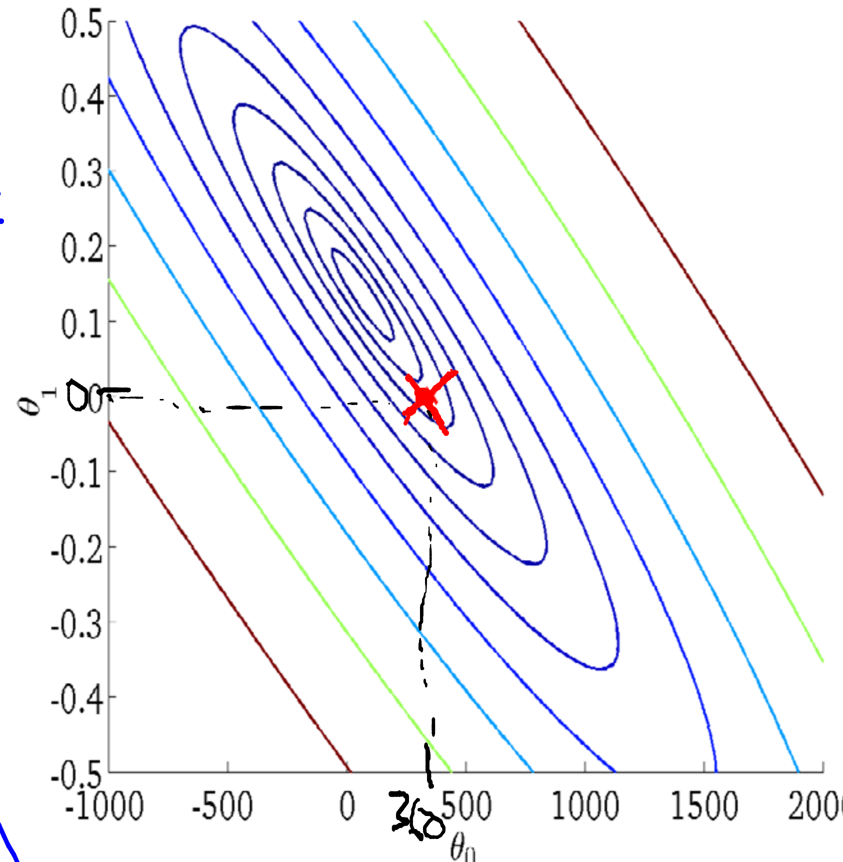
$h_\theta(x)$

(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)



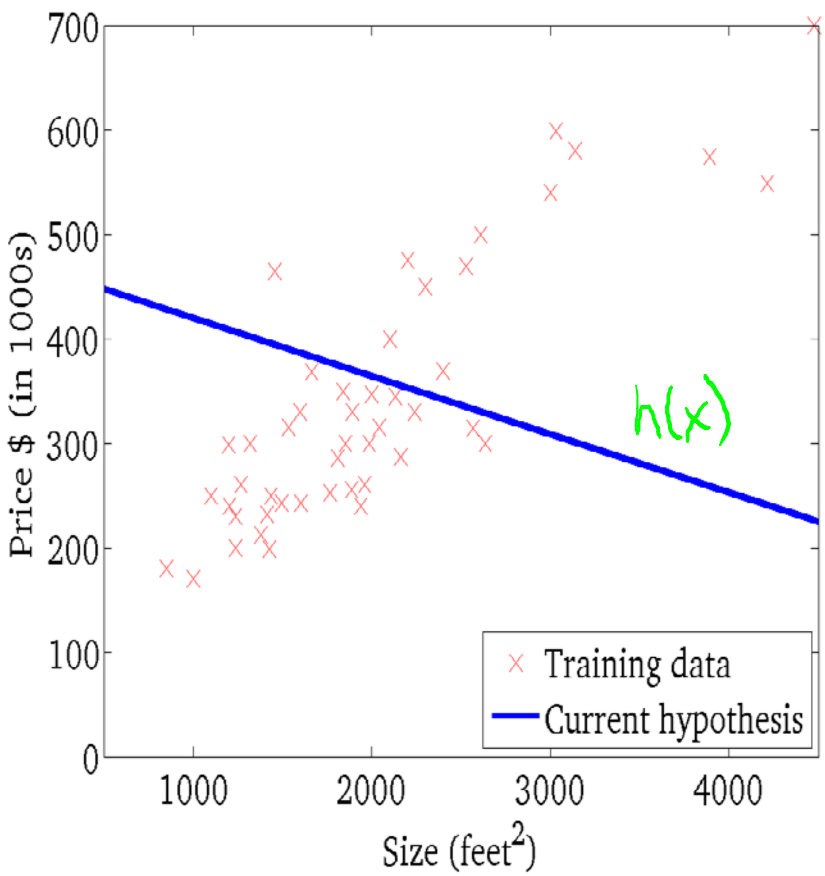
$\theta_0 = 360$

$\theta_1 = 0$

Source: Andrew Ng

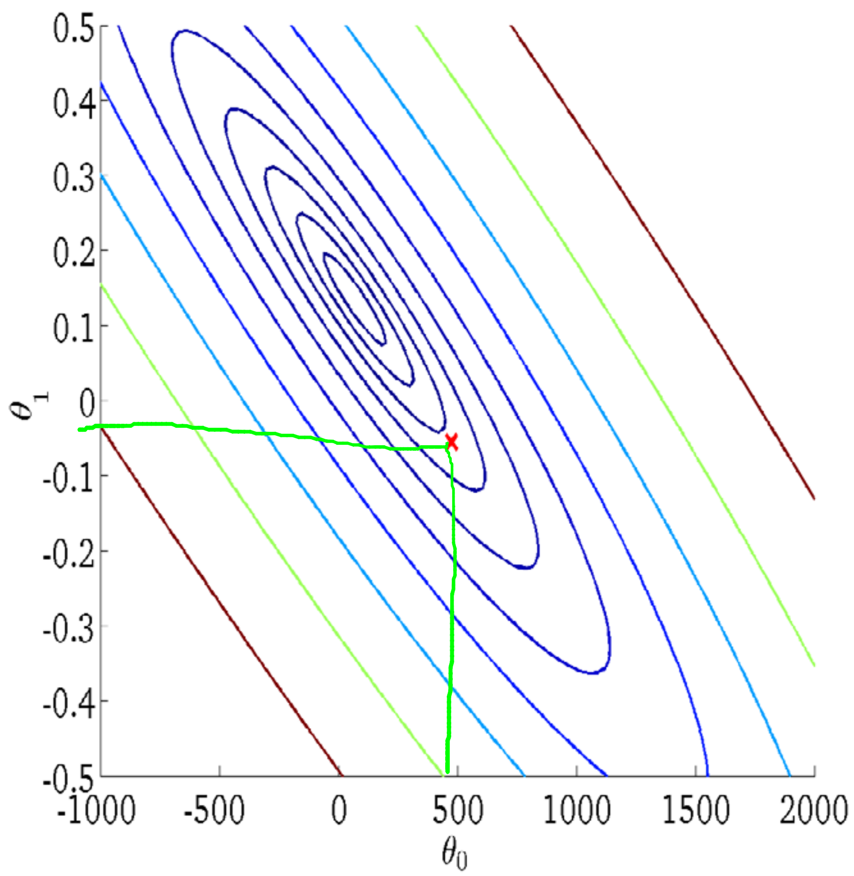
$$h_\theta(x)$$

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



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

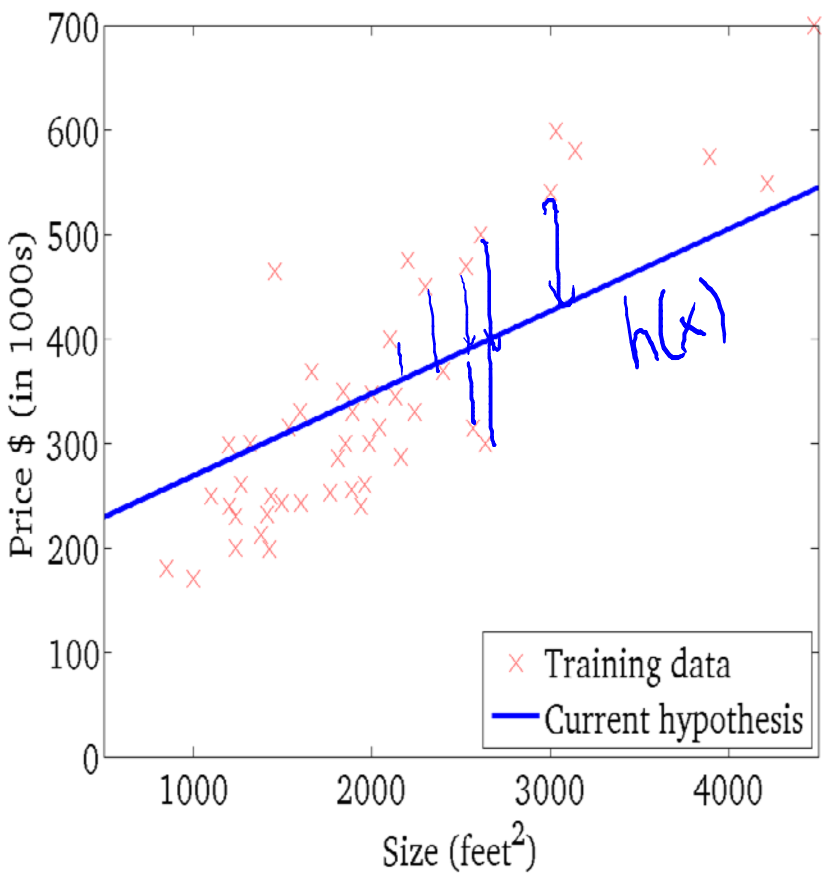
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

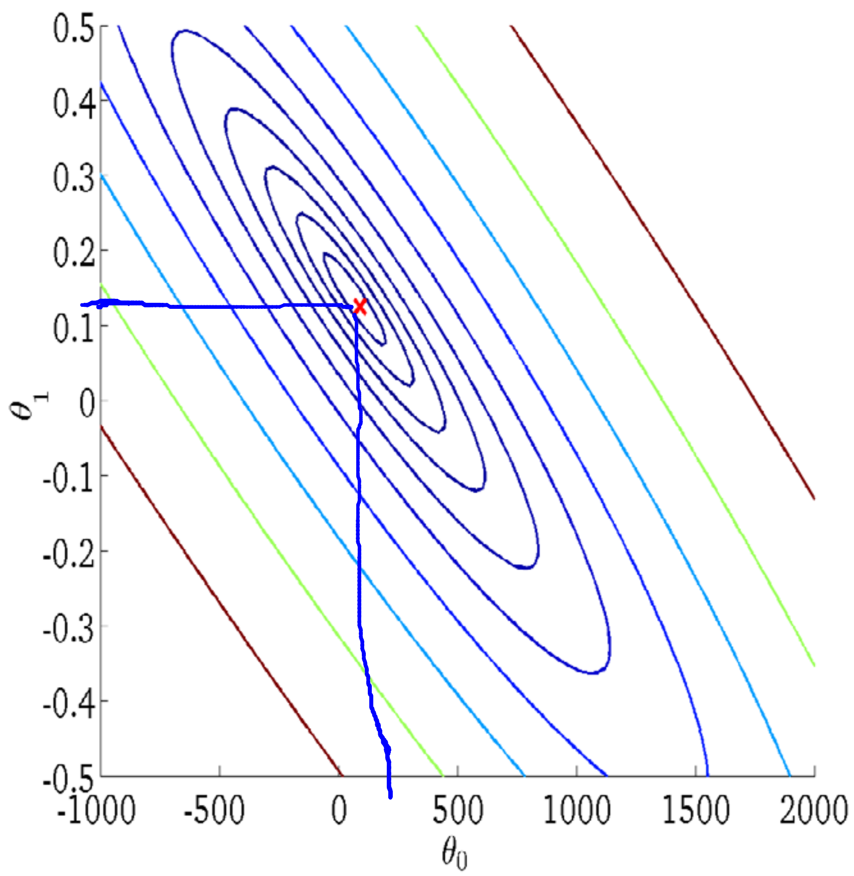
$$h_\theta(x)$$

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



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

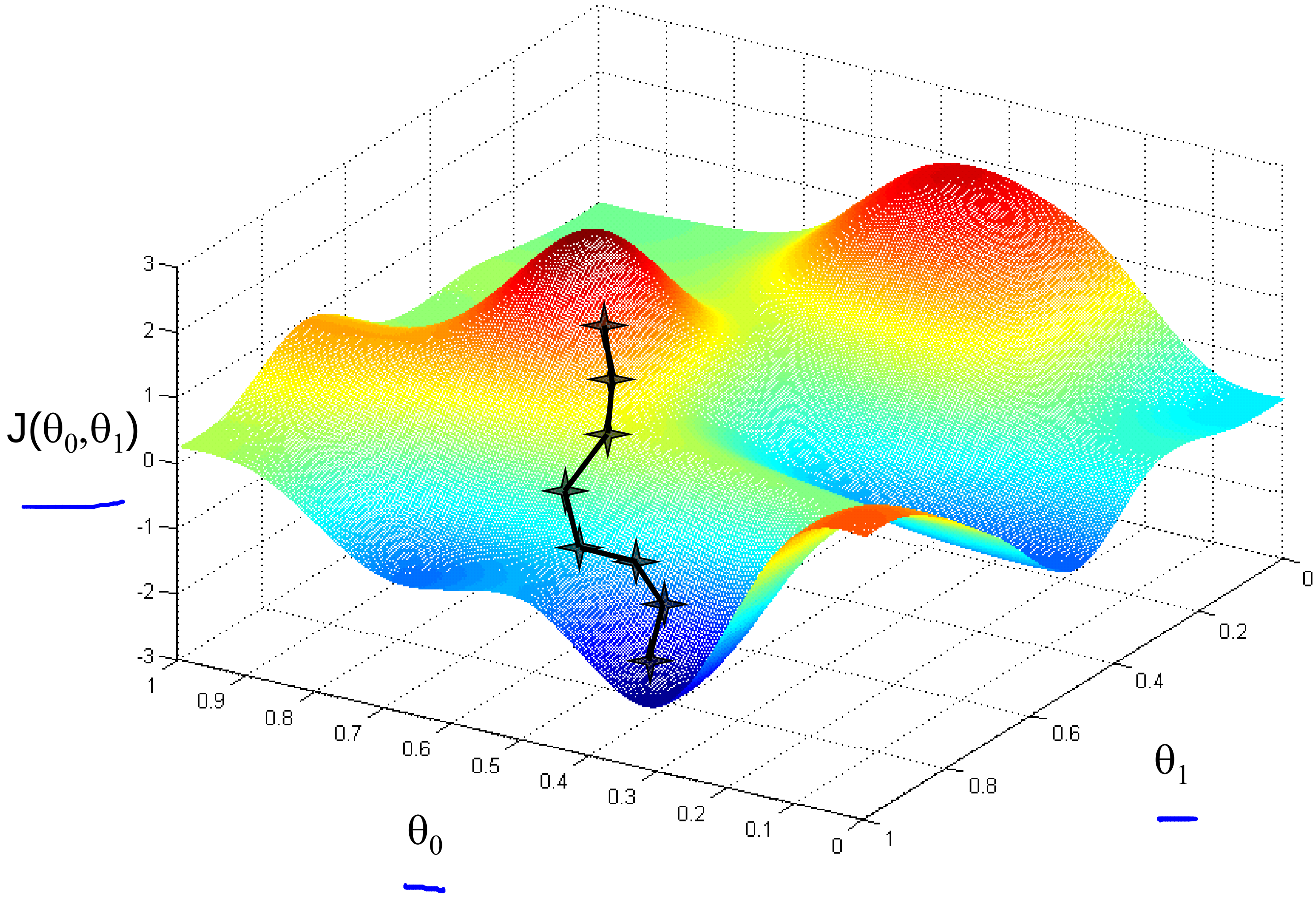
(function of the parameters θ_0, θ_1)



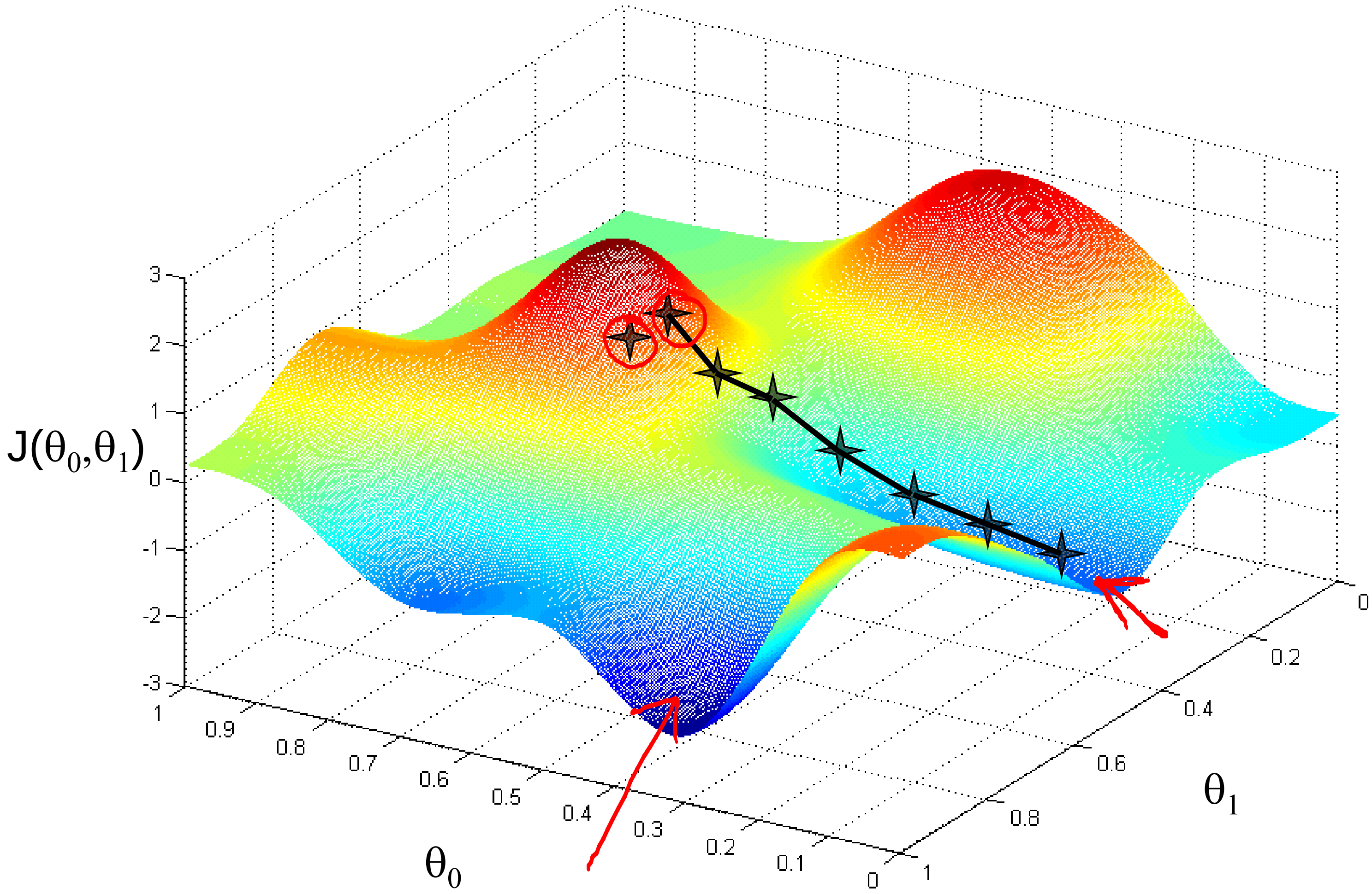
Source: Andrew Ng

2.1. Univariate linear regression

- Gradient descent method => find our the point that minimize J(θ_0,θ_1)
- Method:
  i. Initiate with a random parameter (θ_0,θ_1), ex. (θ_0=0, θ_1=0)
  ii. Change (θ_0,θ_1) to reduce J(θ_0,θ_1)
  iii. Iterate step ii until J(θ_0,θ_1) is (or we believe/accept that it is) minimum



Source: Andrew Ng



Source: Andrew Ng

2.1. UNIVARIATE LINEAR REGRESSION

- Gradient descent algorithm

Repeat until convergence{

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$$ // for j=0 and j=1, simultaneously

}

Learning rate

Correct : Simultaneously update

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\theta_1 := \text{temp1}$$

Wrong:

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$
$$\theta_0 := \text{temp0}$$
$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$
$$\theta_1 := \text{temp1}$$

2.1. UNIVARIATE LINEAR REGRESSION

- Gradient descent algorithm: minimize J(θ_0,θ_1)

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

$$\frac{\partial}{\partial\theta_0}J(\theta_0,\theta_1)$$

Repeat until convergence{

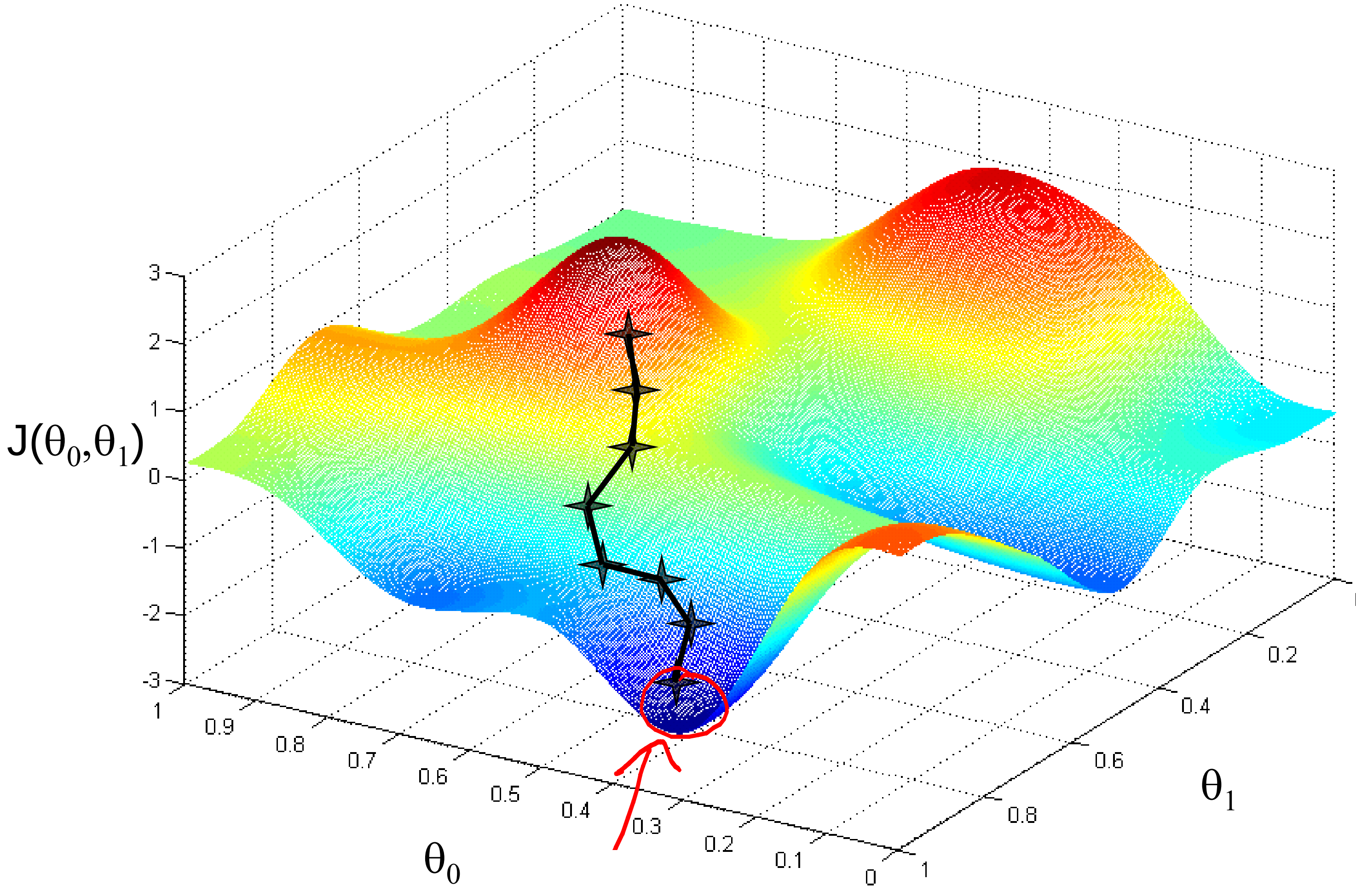
$$\theta_0=\theta_0-\boxed{\alpha\frac{1}{N}\sum_{i=1}^{N}(h_\theta(x^{(i)})-y^{(i)})}$$

$$\theta_1=\theta_1-\boxed{\alpha\frac{1}{N}\sum_{i=1}^{N}(h_\theta(x^{(i)})-y^{(i)})x^{(i)}}$$

Update θ_0 and θ_1 simultaneously

}

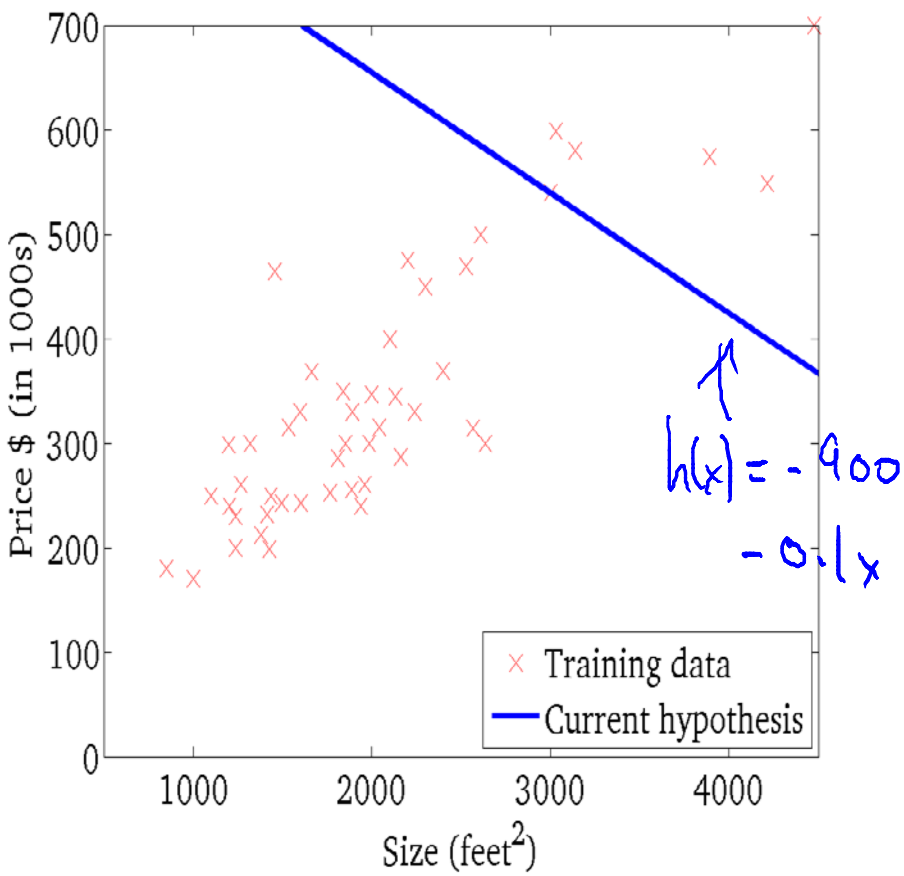
$$\frac{\partial}{\partial\theta_1}J(\theta_0,\theta_1)$$



Source: Andrew Ng

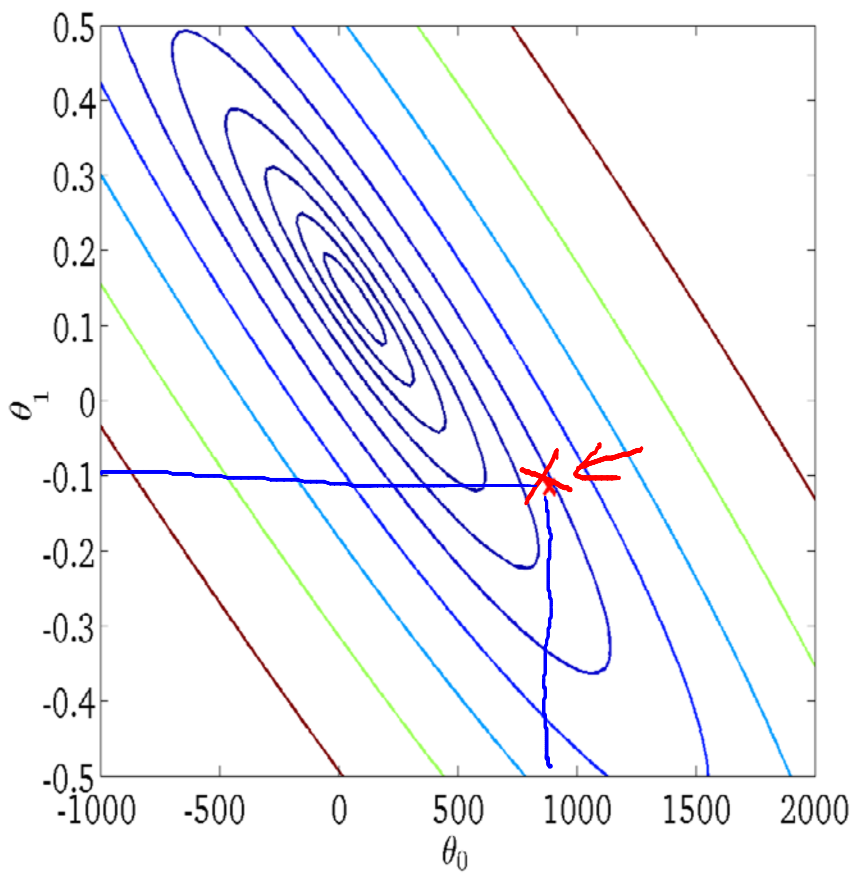
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

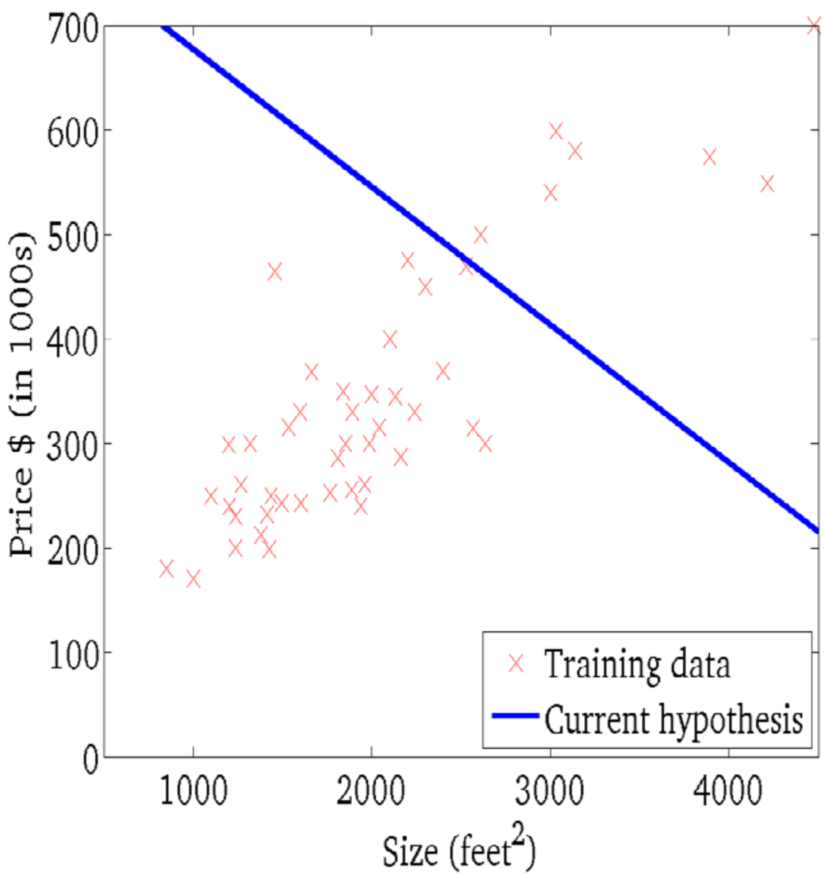
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

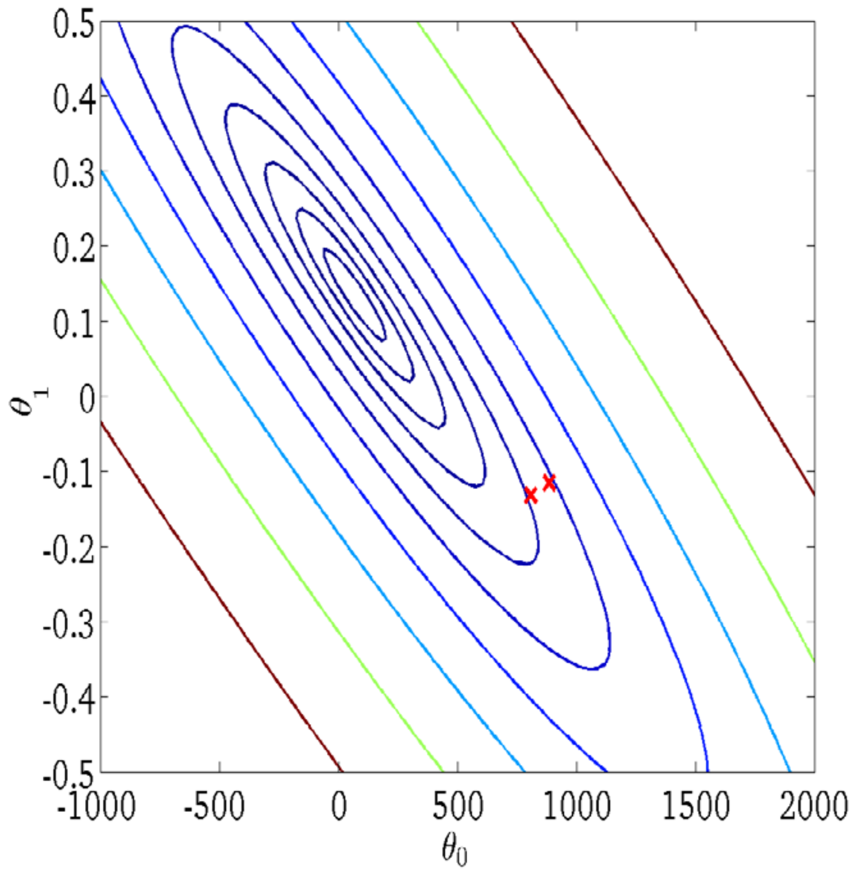
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

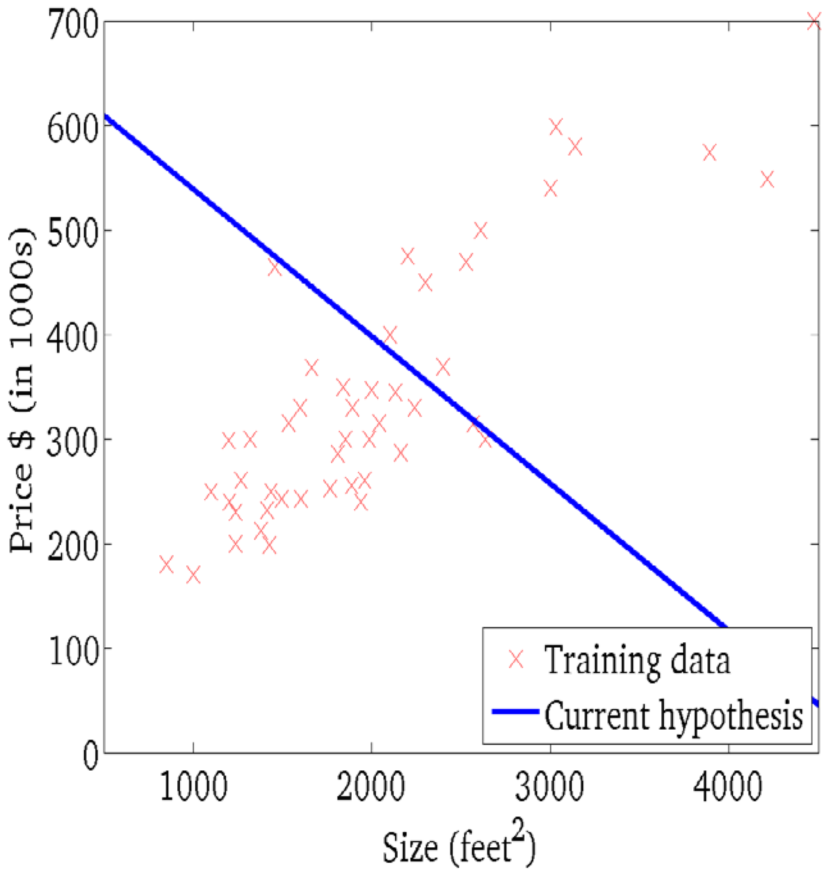
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

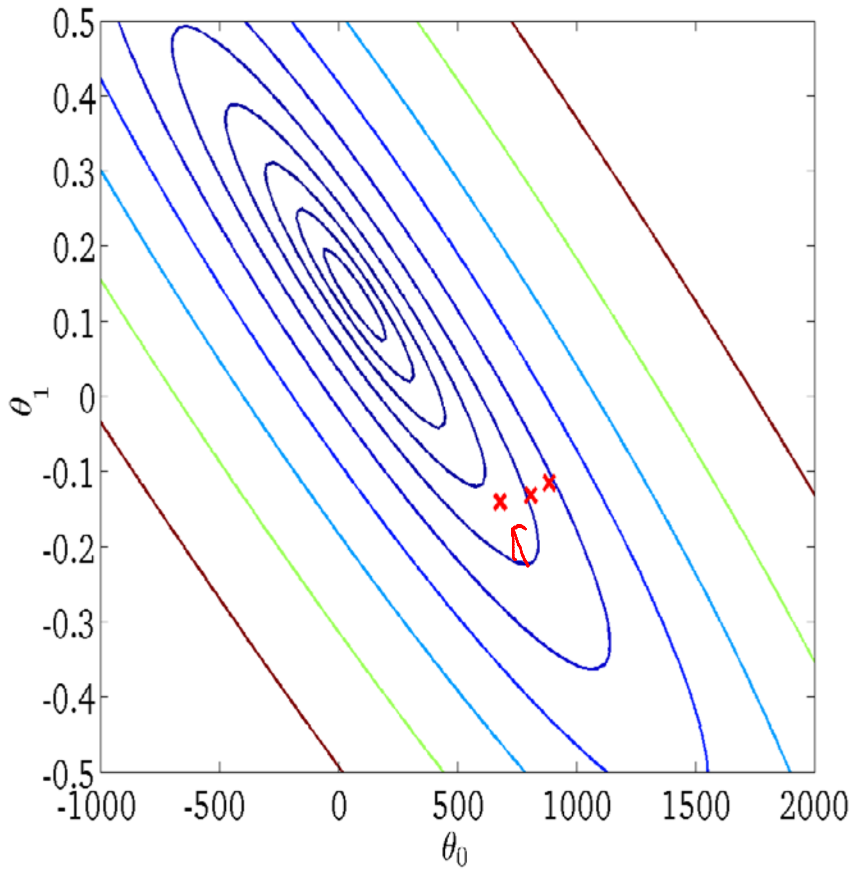
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

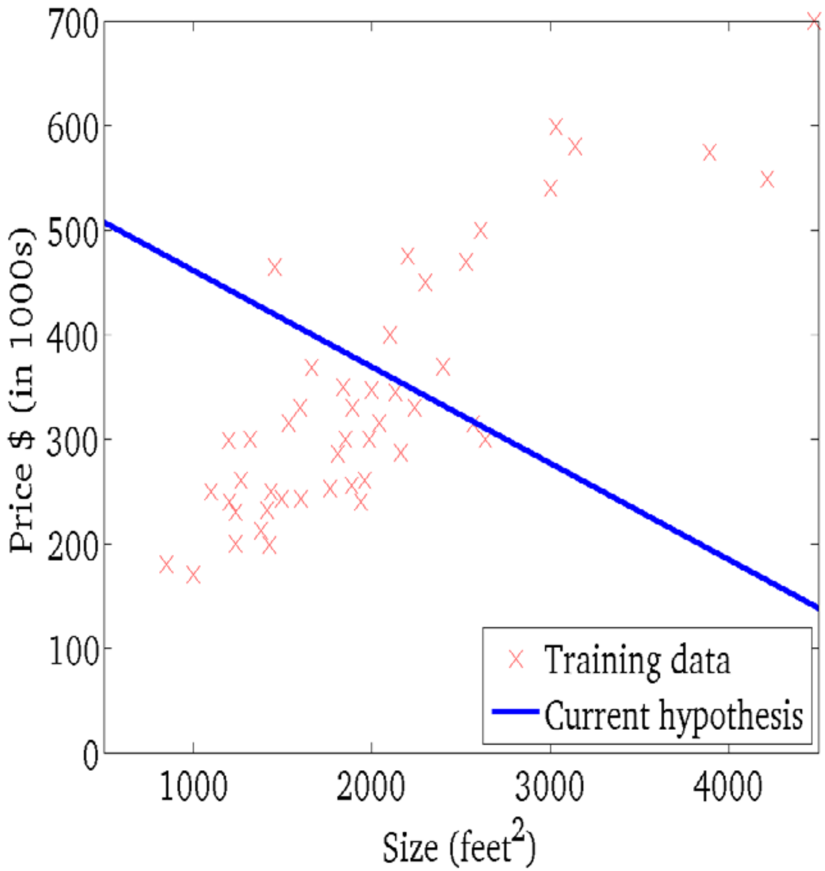
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

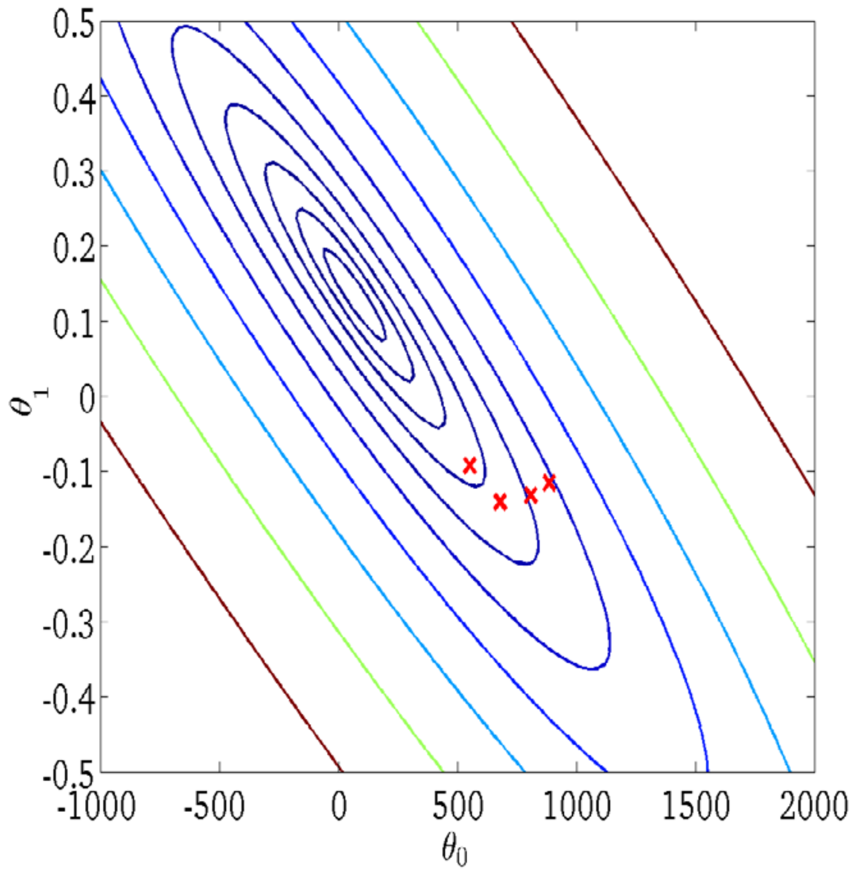
$$h_\theta(x)$$

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



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

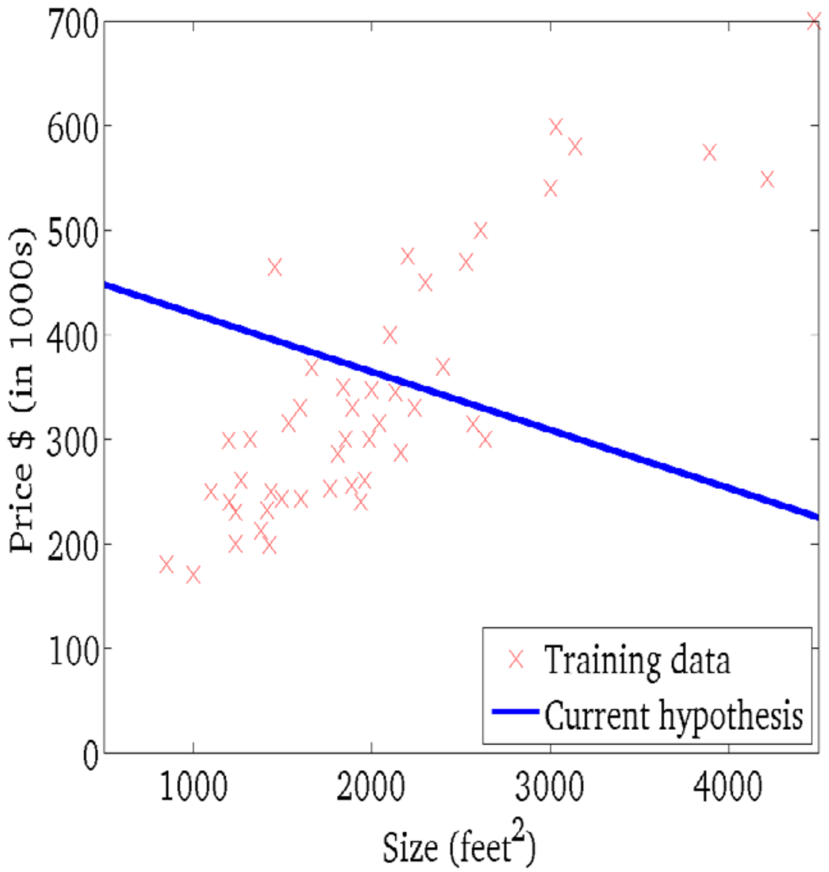
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

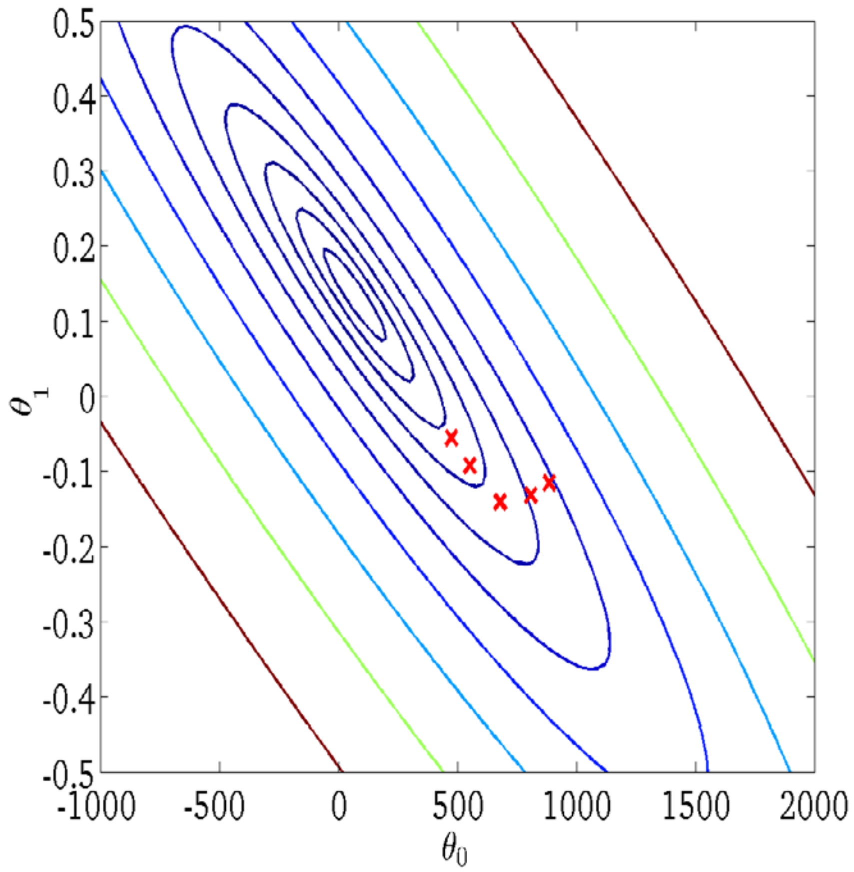
$$h_\theta(x)$$

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



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

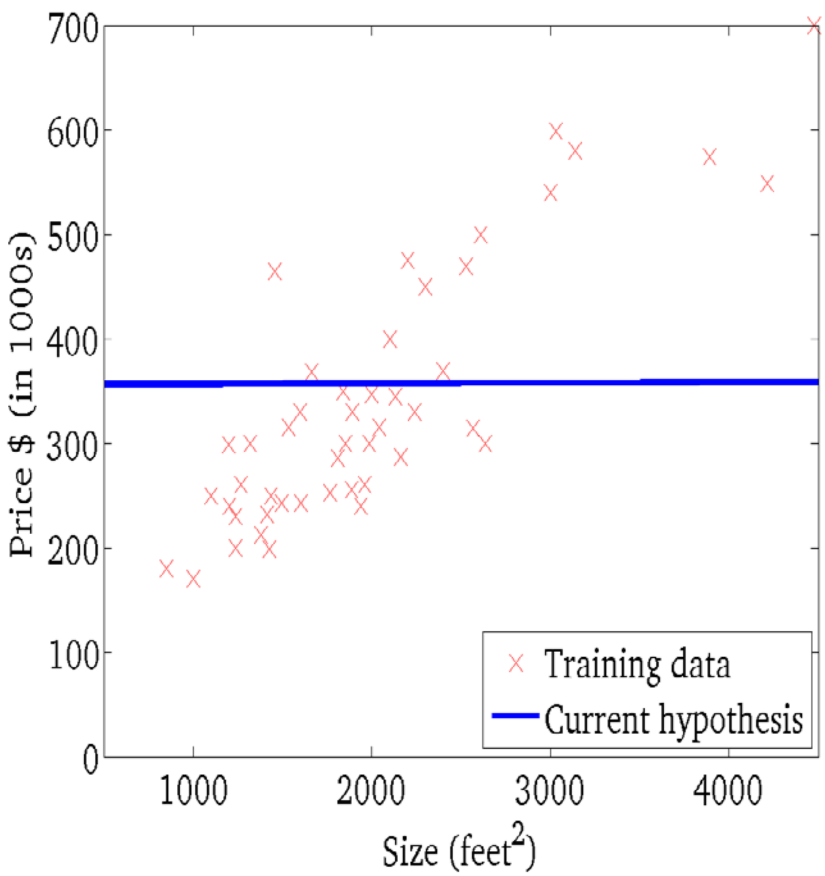
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

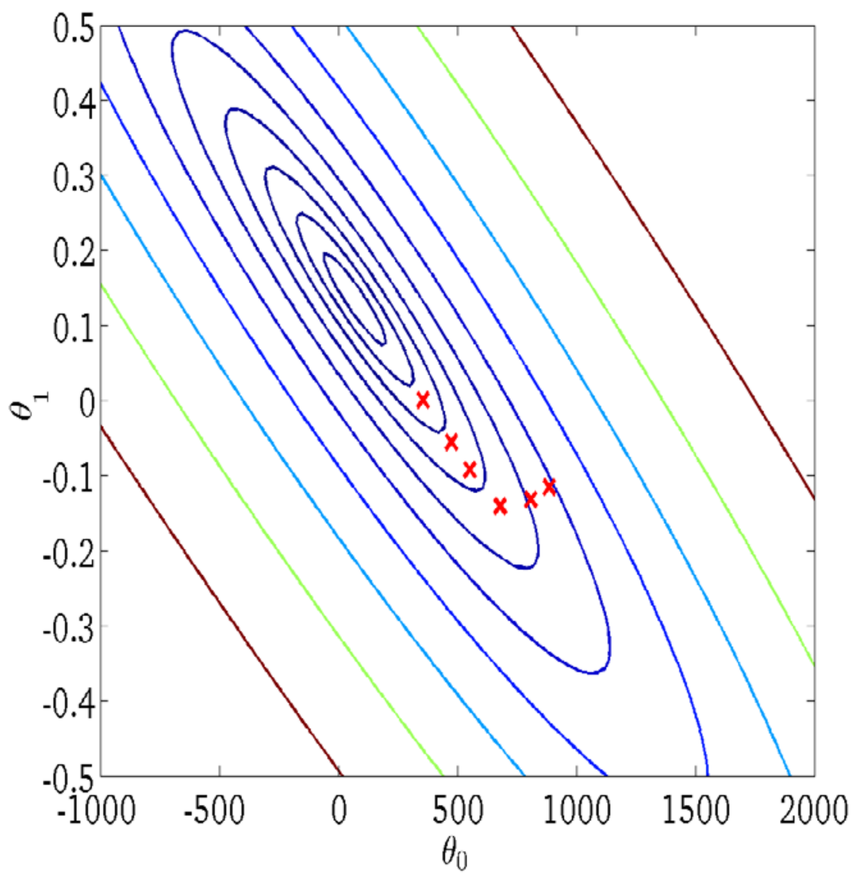
$$h_\theta(x)$$

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



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

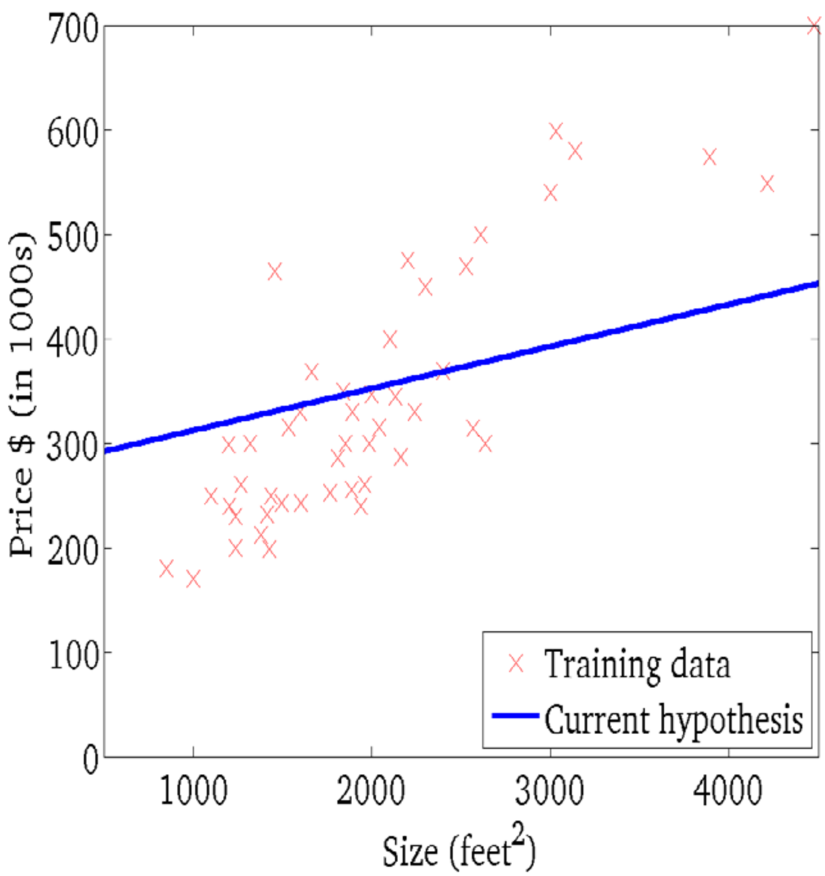
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

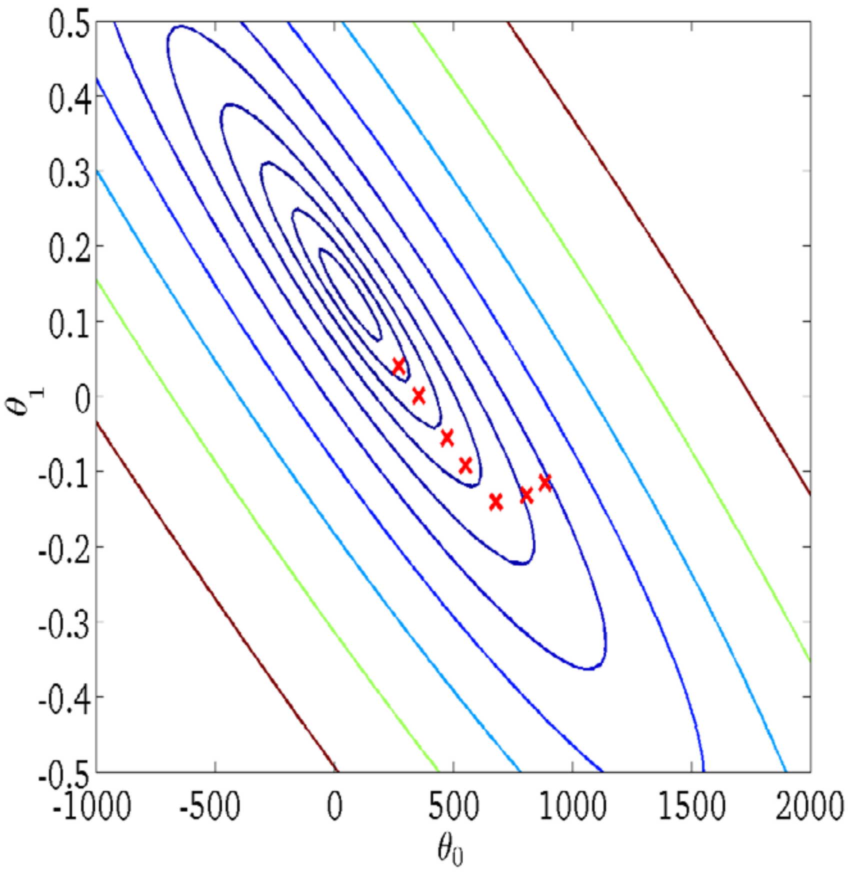
$h_\theta(x)$

(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)



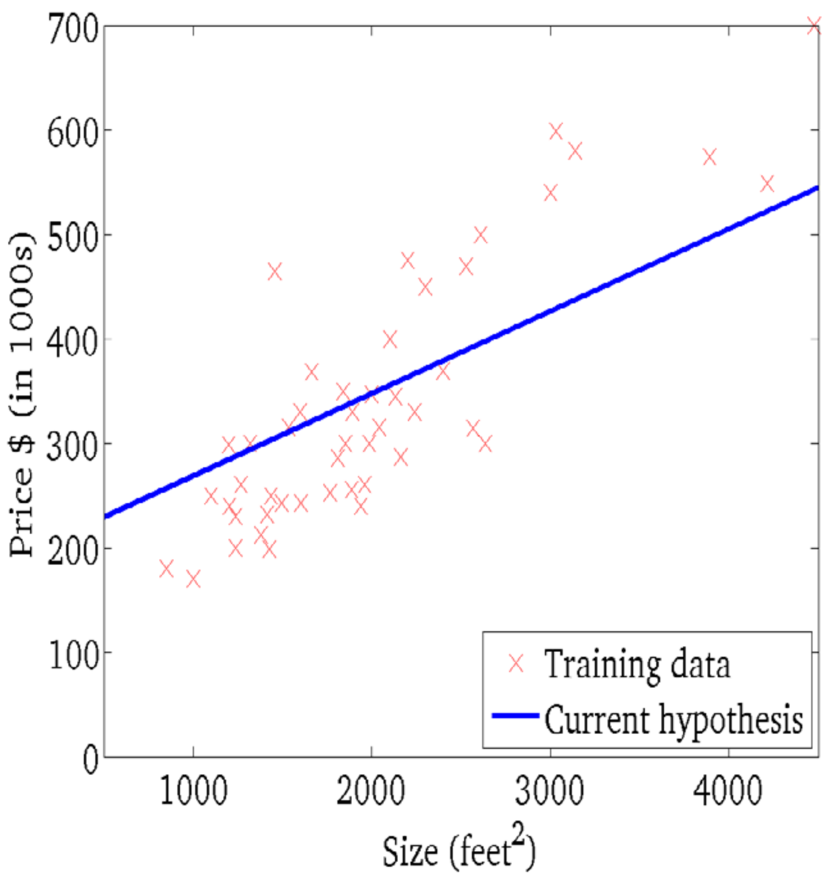
Source: Andrew Ng

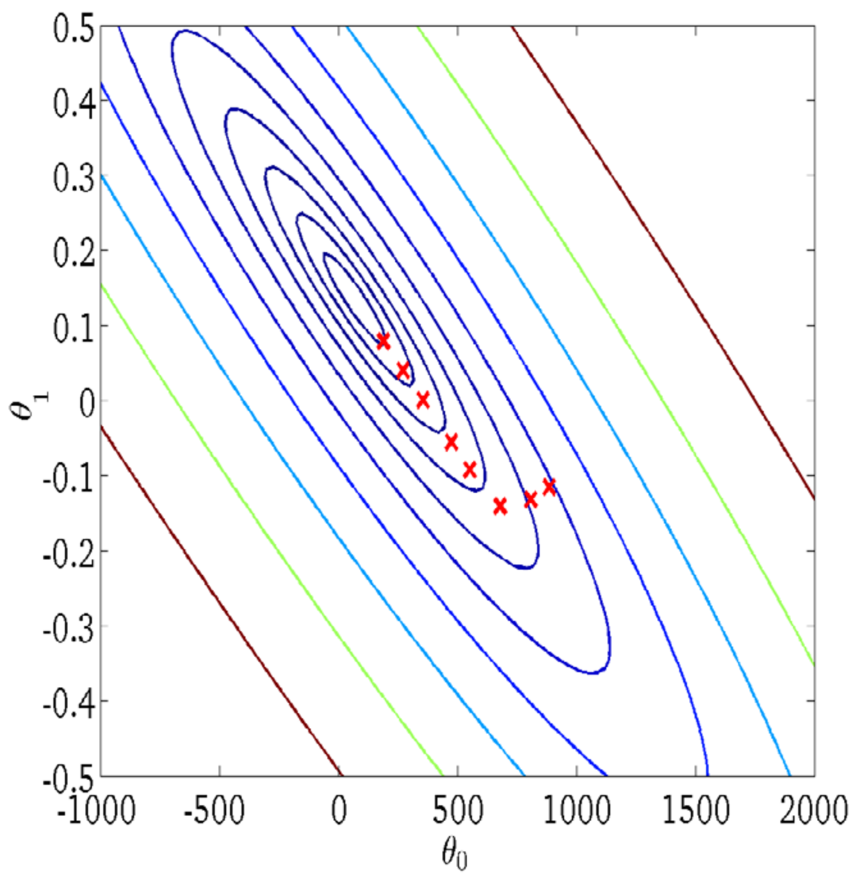
$h_\theta(x)$

(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)

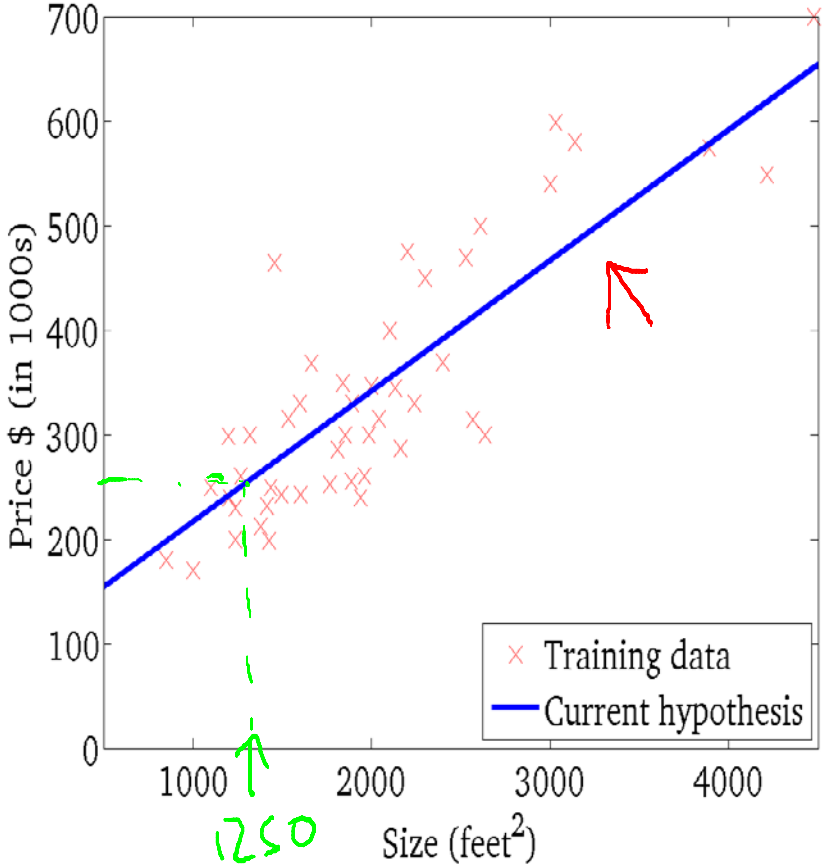




Source: Andrew Ng

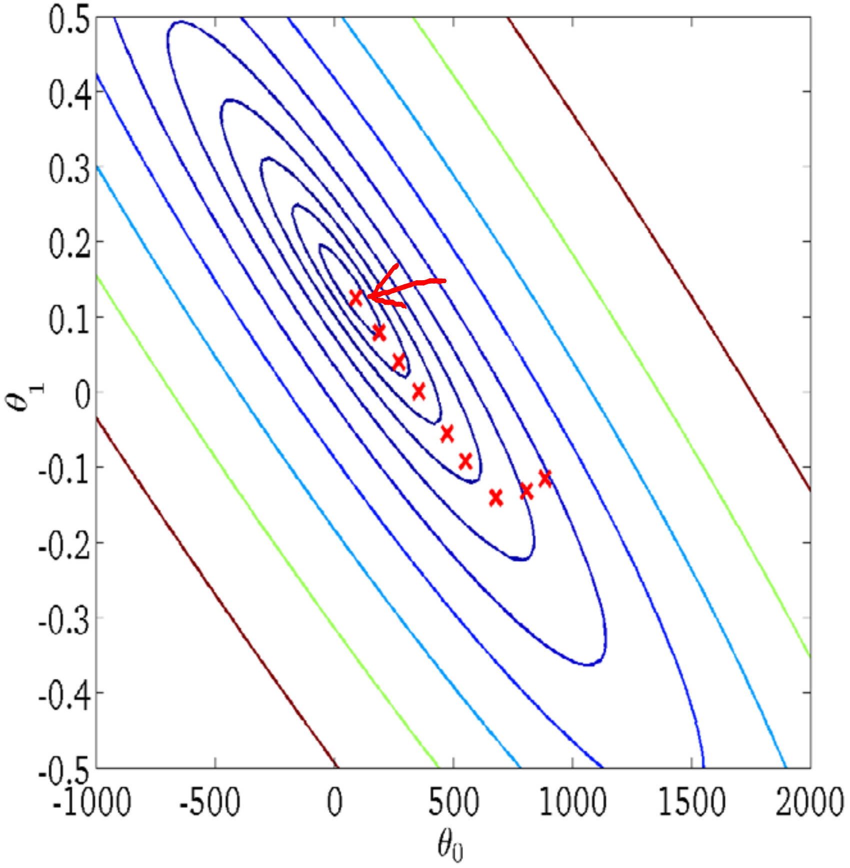
$h_\theta(x)$

(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)



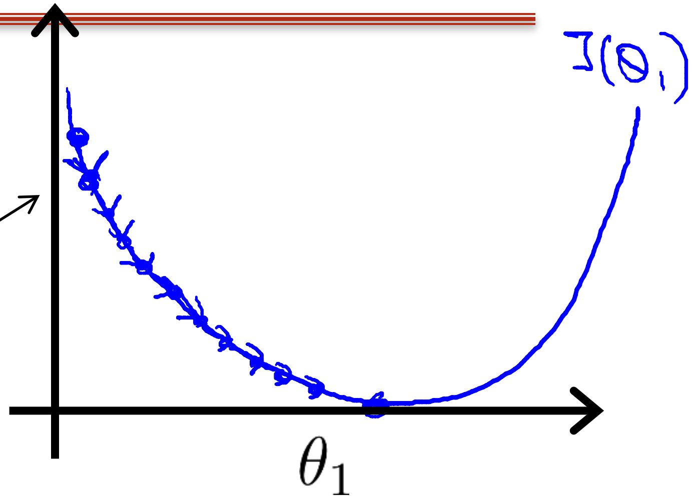
Source: Andrew Ng

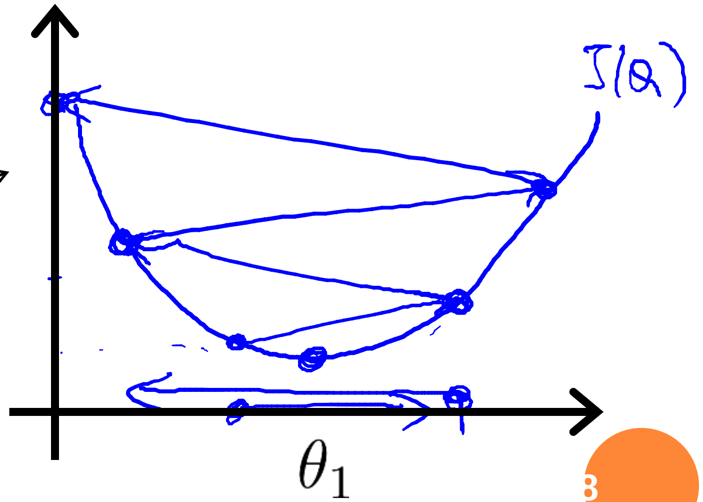
2.1. UNIVARIATE LINEAR REGRESSION

- Learning rate α

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

  - ✓ Too small: Learn slowly
  - ✓ Too big: Difficult to converge to the point where $J(\theta_0,\theta_1)$ is minimum





2.1. UNIVARIATE LINEAR REGRESSION

- Summary: Given N samples, a univariate regression model is described as follow (e_i describes the change of Y which is not explainable from X)
  - Straight line form

$$y_i = h_\theta(x_i) = \theta_0 + \theta_1 x_i + e_i, i = 1,2,\ldots, N$$

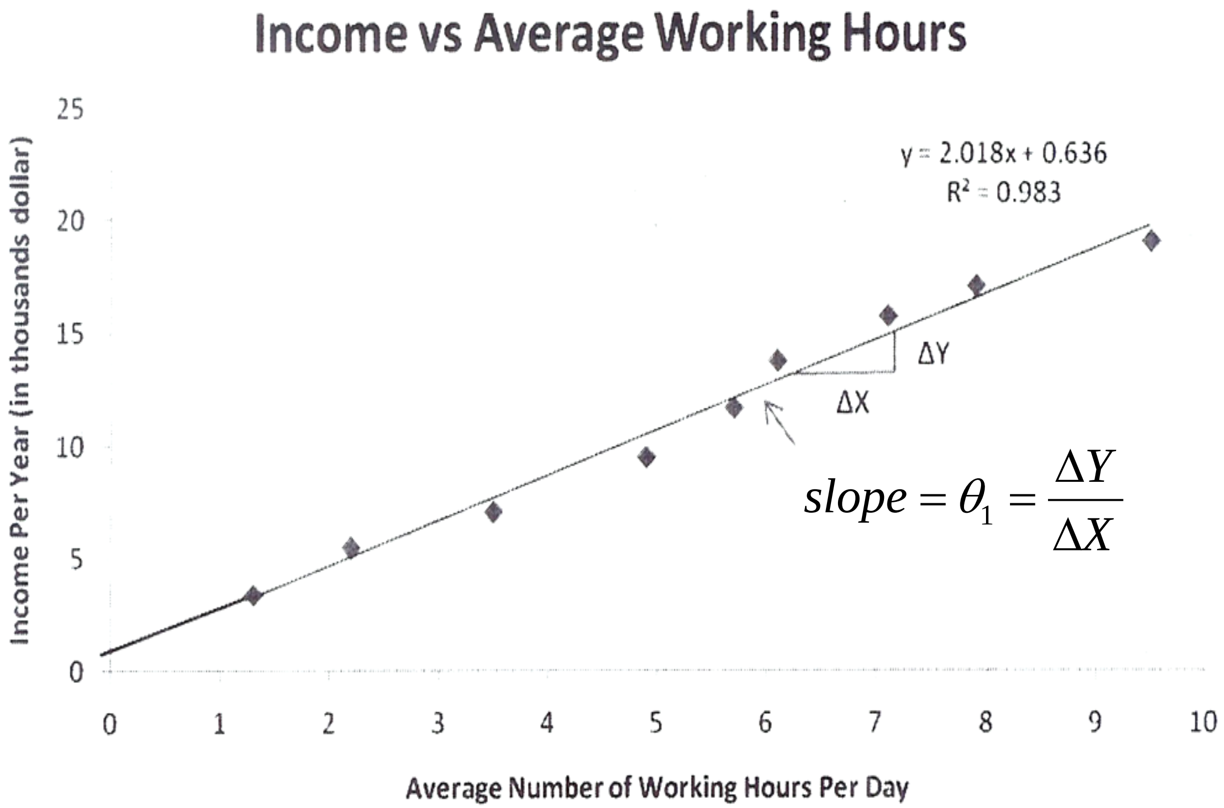
  - Parabol form

$$y_i = h_\theta(x_i) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + e_i, i = 1,2,\ldots, N$$

- Estimating (θ_0,θ_1) by *gradient descent* method or can be quickly estimated by:

$$\theta_1 = \frac{\sum_{i=1}^{N}(x^{(i)} - \overline{x})(y^{(i)} - \overline{y})}{\sum_{i=1}^{N}(x^{(i)} - \overline{x})^2} \qquad \theta_0 = \overline{y} - \theta_1 \overline{x}$$

2.1. Univariate linear regression



- $Y = \theta_0 + \theta_1 * X1 \rightarrow Y = 0.636 + 2.018 * X$
- The sign of θ_1 describes the effect direction (positive/negative) of X on Y.

2.1. UNIVARIATE LINEAR REGRESSION

Quantity Sold	Price($)
8500	2
4700	5
5800	3
7400	2
6200	5
7300	3
5600	4

y=quantySold=9323−823*price

2.1. UNIVARIATE LINEAR REGRESSION

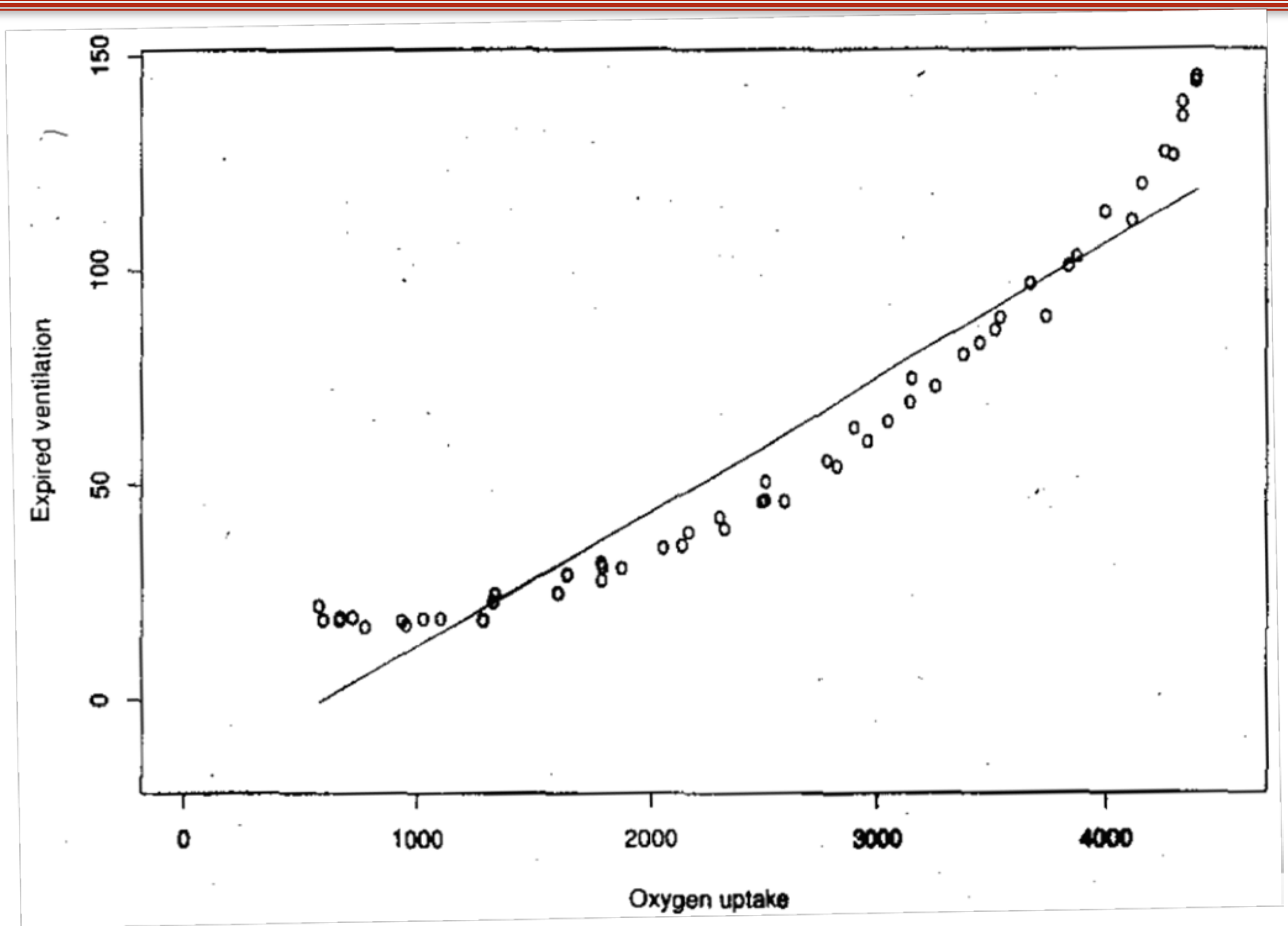


Figure 11.1, [2], pp. 372. Expired ventilation plotted against oxygen uptake in a series of trials, with fitted straight line: y = $\theta_0 + \theta_1$x.

2.1. Univariate linear regression

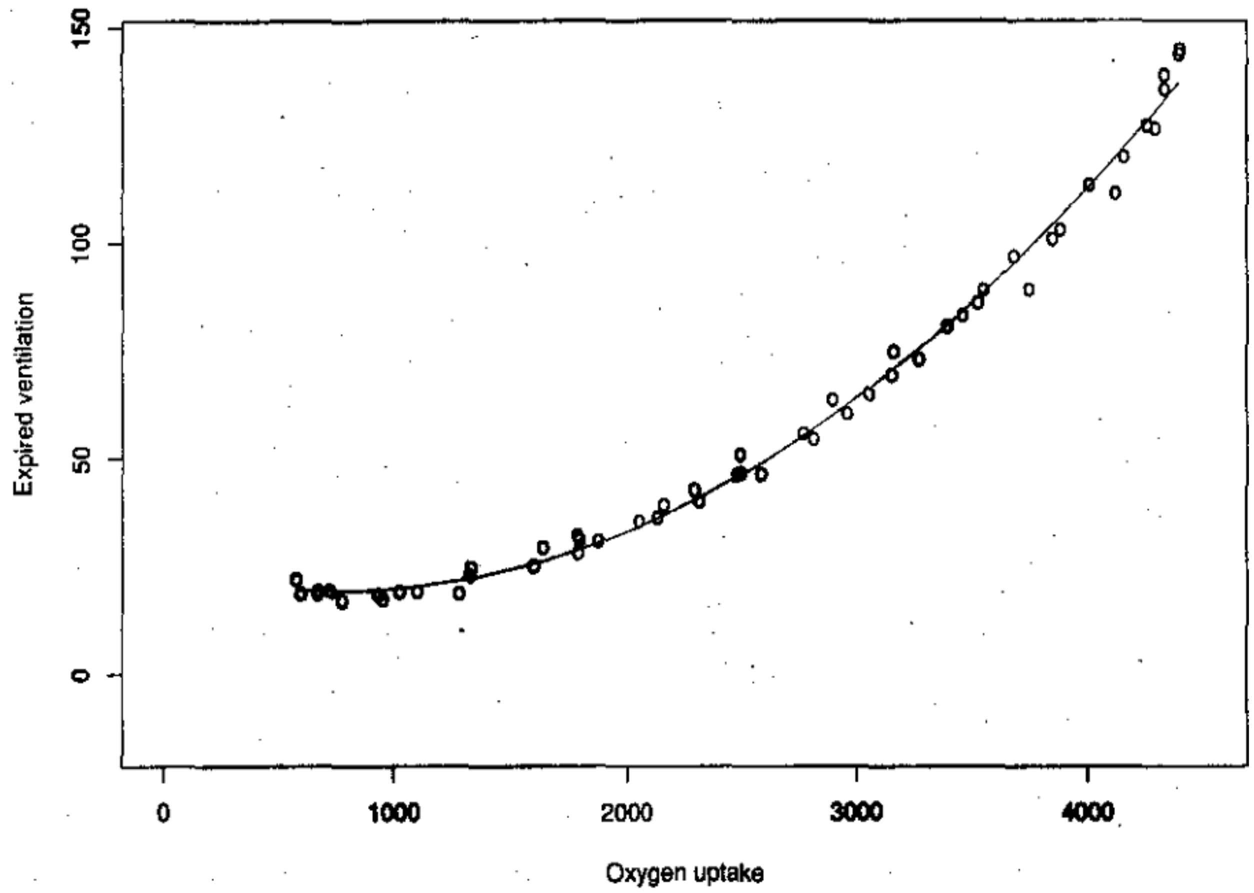


Figure 11.2, [2], pp. 373. The data from Figure 11.1 with a model that includes a term in x^2: $y = \theta_0 + \theta_1 x + \theta_2 x^2$.

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2.2. Multivariate linear regression

- The house prices is affected by several variables/factors

Size (feet²)	Number of rooms	Floors	Age	Price($1K)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

n: number of input attributes (e.g., n = 4)

$x^{(i)}$: input (features) of the i^{th} training sample

$x_j^{(i)}$: value of attribute j in the training sample i^{th}

$y^{(i)}$: i^{th} output in the training dataset

E.x.,

$$x^{(1)} = \begin{bmatrix} 2014 \\ 5 \\ 1 \\ 45 \end{bmatrix}; x_3^{(1)} = 1$$

2.2. MULTIVARIATE LINEAR REGRESSION

- Hypothesis:

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

$$h_\theta(x) = 100 + 3x_1 + 2x_2 + 1.5x_3 - 2x_4$$

- Presenting in a matrix form ($x_0 = 1$)

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}; \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}; \theta^T = [\theta_0, \theta_1, \theta_2, \ldots, \theta_n]$$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n = \theta^T x$$

2.2. Multivariate linear regression

- Gradient descent:

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n = \theta^T x$$

- Coefficients $\theta(\theta_0,\ldots, \theta_n)$: an n+1 vector
- Minimize: $J(\theta) = J(\theta_0,\ldots, \theta_n)$

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Repeat until convergence{

$\theta_j := \theta_j - \alpha \dfrac{\partial}{\partial \theta_j} J(\theta)$ // simultaneously update for every j=0,...,n

}

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2.2. HỒI QUI TUYẾN TÍNH ĐA BIẾN

```
Repeat until convergence{
```

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$ // simultaneously update for every j=0,...,n

```
}
```

$$\theta_0 = \theta_0 - \alpha \frac{1}{N} \sum_{i=1}^{N} (h_\theta(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$x_0^{(i)}=1$

$$\theta_1 = \theta_1 - \alpha \frac{1}{N} \sum_{i=1}^{N} (h_\theta(x^{(i)}) - y^{(i)}) x_1^{(i)}$$

$$\theta_2 = \theta_2 - \alpha \frac{1}{N} \sum_{i=1}^{N} (h_\theta(x^{(i)}) - y^{(i)}) x_2^{(i)}$$

....

2.2. MULTIVARIATE LINEAR REGRESSION

- Feature scaling: to assure that all input features are in the same scale
  - E.g., x_1 = size (0 - 2000 feet^2)

    x_2 = number of rooms (1 - 5)

  => They are not in the same scale. The convergence speed is affected because of this imbalance scaling.

2.2. Multivariate linear regression

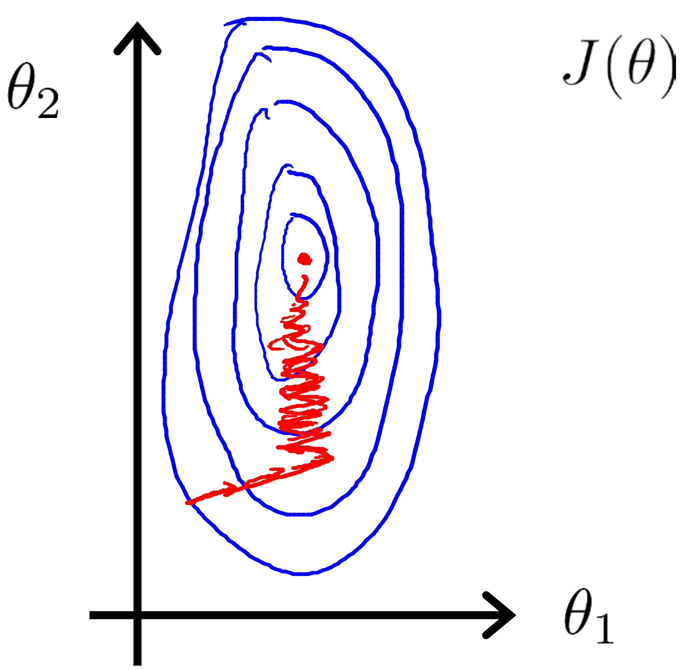
- Assure that features are in the same scale

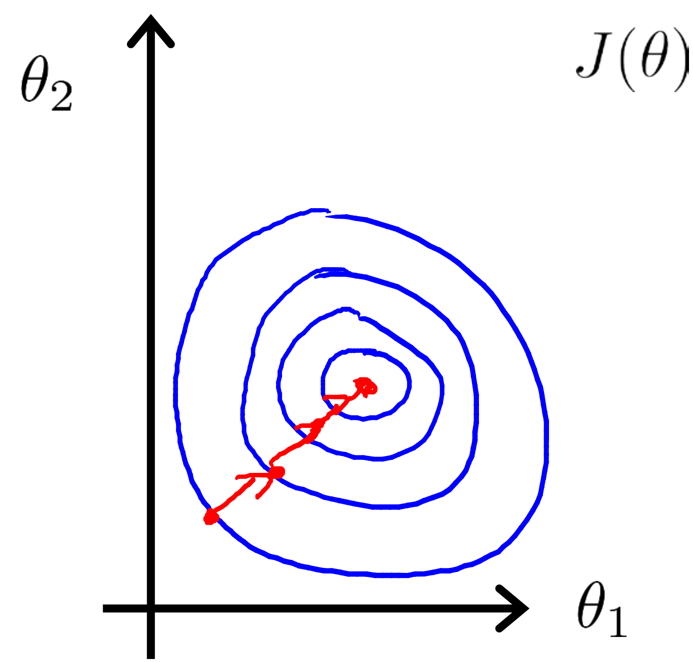
E.g. x_1 = size (0-2000 feet^2)

x_2 = number of bedrooms (1-5)

Scaled: x_1= size/2000

x_2= number of bedrooms /5





Source: Andrew Ng

2.2. Multivariate linear regression

- Feature scaling
  - Normalize all feature to a range of [-1, 1]
  - Ex., $x_0 = 1$; $0 \le x_1 \le 3$; $-2 \le x_2 \le 0.5$ => **OK**

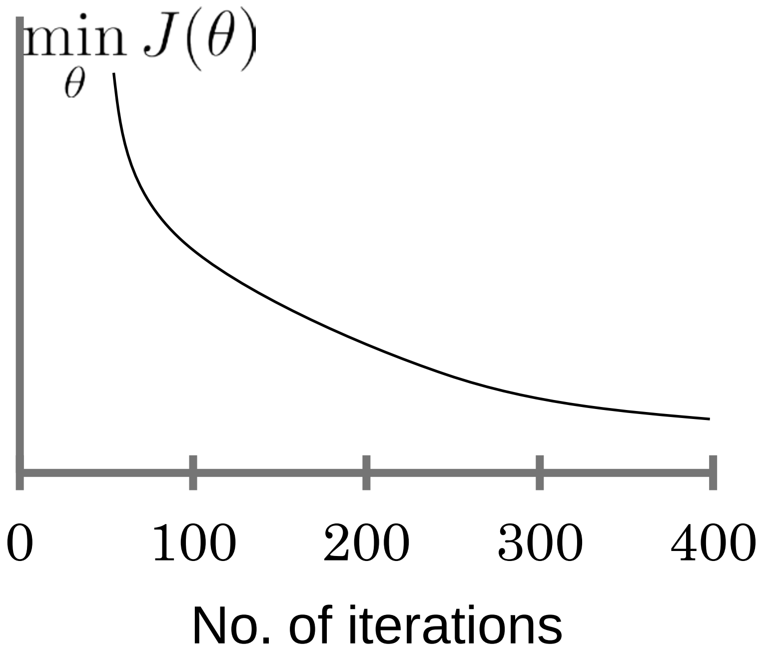
    $-100 \le x_3 \le 100$; $-0.0001 \le x_4 \le 0.0001$ => Normalize

- Apply normalization methods in chapter 2
  - Ex.

$$v' = \frac{v - \bar{v}}{\sigma v} \qquad v' = \frac{v - \bar{v}}{V_{\max} - V_{\min}}$$

2.2. MULTIVARIATE LINEAR REGRESSION

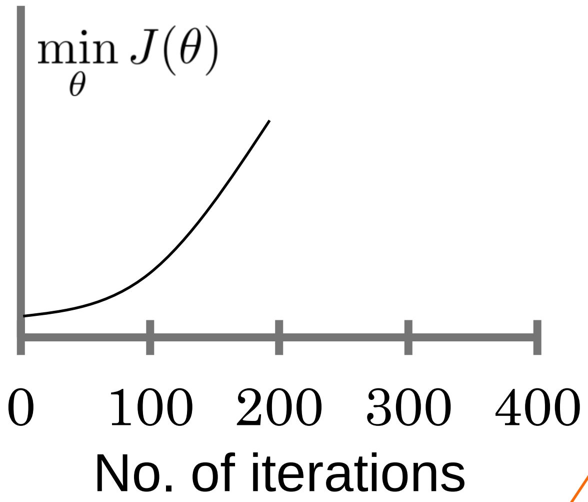
- Validate the Gradient descent algorithm
  - $J(\theta)$ must decrease after each iteration
  - We can plot $J(\theta)$ by θ for intuitively check the convergence ability of the algorithm

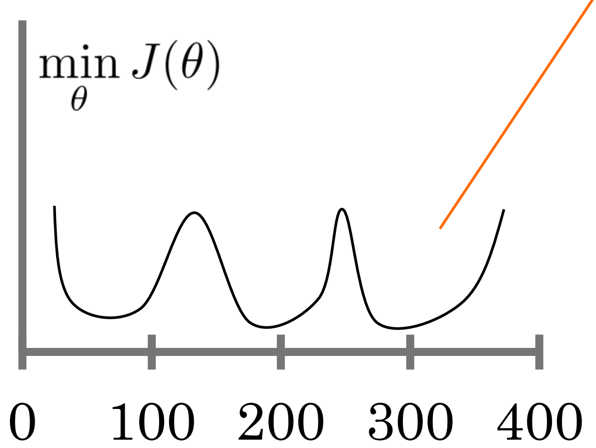


Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

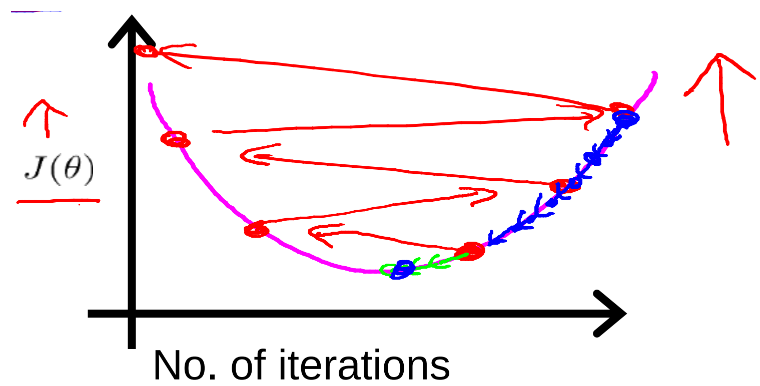
2.2. Multivariate Linear Regression

- Un-converged gradient descent





Should reduce α



-α too small: slow convergence

-α too large: J(θ) may not reduce at each iteration => the algorithm may no converge

- try α: 0.001, 0.003, 0.01, 0.03, 0.1, 0.3,…

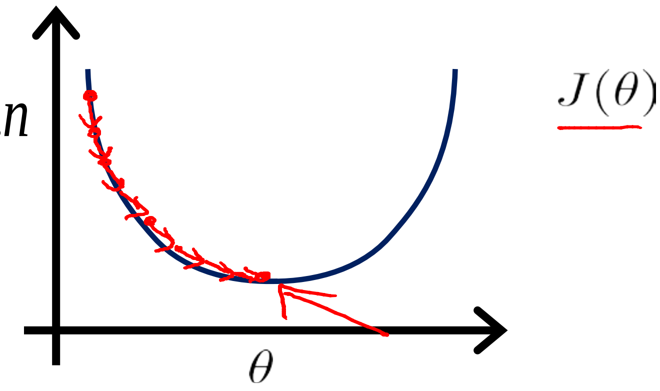
2.2. Multivariate linear regression

- Use normal equation to identify θ

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j^{(i)} = 0; \forall j = 1..n$$

$$\theta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

Gradient Descent



Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2.2. Multivariate linear regression

Examples: N = 4, X: an Nx(n+1) matrix; y: an Nx1 matrix

x_0	size (feet²) x_1	No. of rooms x_2	Floors x_3	Years x_4	Price ($1K) y
1	2104	5	1	45	460
1	1416	3	2	40	232
1	1534	3	2	30	315
1	852	2	1	36	178

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix} \qquad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

$$\theta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$$

2.2. MULTIVARIATE LINEAR REGRESSION

- Given a training dataset: N examples, n features

<u>Gradient descent</u>

- must select α
- needs a large number of iterations
- workable event with a large n (e.g., $n = 10^6$)

<u>Normal equation</u>

- don't need to select α
- don't need iterations
- Must compute $(X^TX)^{-1}$
- May not work when n is large (when $n = 10^4$ then gradient descent should be used)

2.2. MULTIVARIATE LINEAR REGRESSION

- Note: the non-invertible issue in the normal equation method, i.e., (X^TX) is not invertible
- Resolve:
  - Check the linear dependence of variables. Ex., the size in meter (x_1) and the size in feet (x_2) => remove dependent variables
  - Too much features ($n > N$). Ex., $n = 20$, $N = 10$ => reduce the number of features; find surrogate features; correct more data samples,…

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

2.2. MULTIVARIATE LINEAR REGRESSION

- Another example:

Quantity Sold	Price($)	Advertising ($)
8500	2	2800
4700	5	200
5800	3	400
7400	2	500
6200	5	3200
7300	3	1800
5600	4	900

2.2. MULTIVARIATE LINEAR REGRESSION

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.980681
R Square	**0.96174**
Adjusted R Square	0.942604
Standard Error	310.5239
Observations	7

y = Quantity Sold = 8536.214 -835.722 * Price + 0.592 * Advertising

ANOVA

	df	SS	MS	F	Significance F
Regression	2	9694300	4847150	50.26854	**0.0014641**
Residual	4	385700.4	96425.11		
Total	6	10080000			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	**8536.21**	386.9117	22.06243	2.5E-05	7461.974654	9610.453	7461.975	9610.453
Price($)	**-835.72**	99.65304	-8.38632	0.001106	-1112.40356	-559.041	-1112.4	-559.041
Advertising ($)	**0.59223**	0.104347	5.675579	0.004755	0.302515325	0.881942	0.302515	0.881942

3. Non-linear regression

- **Y** = f(**X**, θ)
  - Y is a non-linear function in terms of relationship between parameters θ.
  - Ex: Exponential, logarithmic function, Gauss, …

$$f(x,\theta) = \frac{\theta_1 x}{\theta_2 + x}$$

- Identify optimal θ: Optimization algorithms
  - Local optimization
  - Global optimization (using sum of squared residuals/errors)

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

4. Applications

- Data mining
  - Data preprocessing: Smoothing, noise removal,…
  - Mining tasks: numerical-values prediction, descriptive analysis
- Apply in many domains: biology, agriculture, social issues, economy, business, finances, insurance, e-commerce, marketing, security, science, robotics, control systems, automation,…

5. Issues in Regression

- Assumptions
  - Data distribution: the relationship between predictors and dependent variables
  - Independence of predictors
  - Continuous values of variables (both predictor & responses)
  - Errors: How to identify them?
- The amount of data processed is not large
- How to identify the regression model
- Advanced techniques for regression:
  - Artificial Neural Network (ANN)
  - Support Vector Machine (SVM)

5. ISSUES IN REGRESSION

- Evaluation of a regression model:
  - Collect new data to evaluate the prediction results
  - Use the existing data (as testing dataset) for evaluation
  - Data splitting
    - ✓ Training data: To build the model
    - ✓ Testing data → *validate/evaluate the model*
  - K-fold cross-validation
    - ✓ Iterate k times:
      - ✓ Training data: (k-1) portions of data
      - ✓ Test data: the k^{th} portion of data → *accuracy*
    - ✓ *Average(accuracy)* of k times

Dr. Tran Minh Quang – quangtran@hcmut.edu.vn

5. Issues in Regression

- Evaluation of the regession model:
  - Accuracy
    - ✓ Sum of squared errors (SSE)

    -> Overall measure of errors: **smaller is better**

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

    - ✓ Mean squared error (MSE): measure of the variability in the response variable left unexplained by the regression: **smaller is better**

$$MSE = \frac{SSE}{n-m-1}$$

(n: sample size, m: number of regression coefficients)

Daniel T. Larose, “Data mining methods and models”, John Wiley & Sons, Inc, 2006.

5. Issues in Regression

- Accuracy (Con't)
  - ✓ The standard error of the estimate (S)
  - ✓ Đánh giá sai số thông thường trong quá trình dự đoán, sự sai lệch giữa giá trị dự đoán và giá trị thực của biến đáp ứng
  - ✓ Measure the common error in the prediction process. It is the mean difference between the predicted and the actual values.
  - ✓ Presents the precision of the prediction generated by the regression model

$$S = \sqrt{MSE} = \sqrt{\frac{SSE}{n-m-1}}$$

Daniel T. Larose, "Data mining methods and models", John Wiley & Sons, Inc, 2006.

5. Issues in Regression

- Factors affect the success of building regression models
  - ✓ Proper problem formulation
  - ✓ Selection of important variables and model form
  - ✓ Good dataset (both in volume and quality)
  - ✓ The use of good coefficient estimation procedures (e.g., gradient descent)
  - ✓ Model validation techniques

Daniel T. Larose, “Data mining methods and models”, John Wiley & Sons, Inc, 2006.

6. SUMMARY

- Regression
  - A statistical technique, applied to continuous attributes/features
  - Simple yet useful, applicable in various domains
  - One of example showing the contribution of statistics in data mining
- Types: Linear/non-linear, Univariate/Multivariate, Parametric/Non-parametric/Semi-parametric, Symmetric/Assymetric

Q&A

quangtran@hcmut.edu.vn