

Khoa Khoa Học & Kỹ Thuật Máy Tính
Trường Đại Học Bách Khoa Tp. Hồ Chí Minh

Chương 3

Hồi Qui Dữ Liệu

TRAN MINH QUANG

quangtran@hcmut.edu.vn

<http://www.cse.hcmut.edu.vn/staff/Staff/quangtran>

<http://researchmap.jp/quang>

1

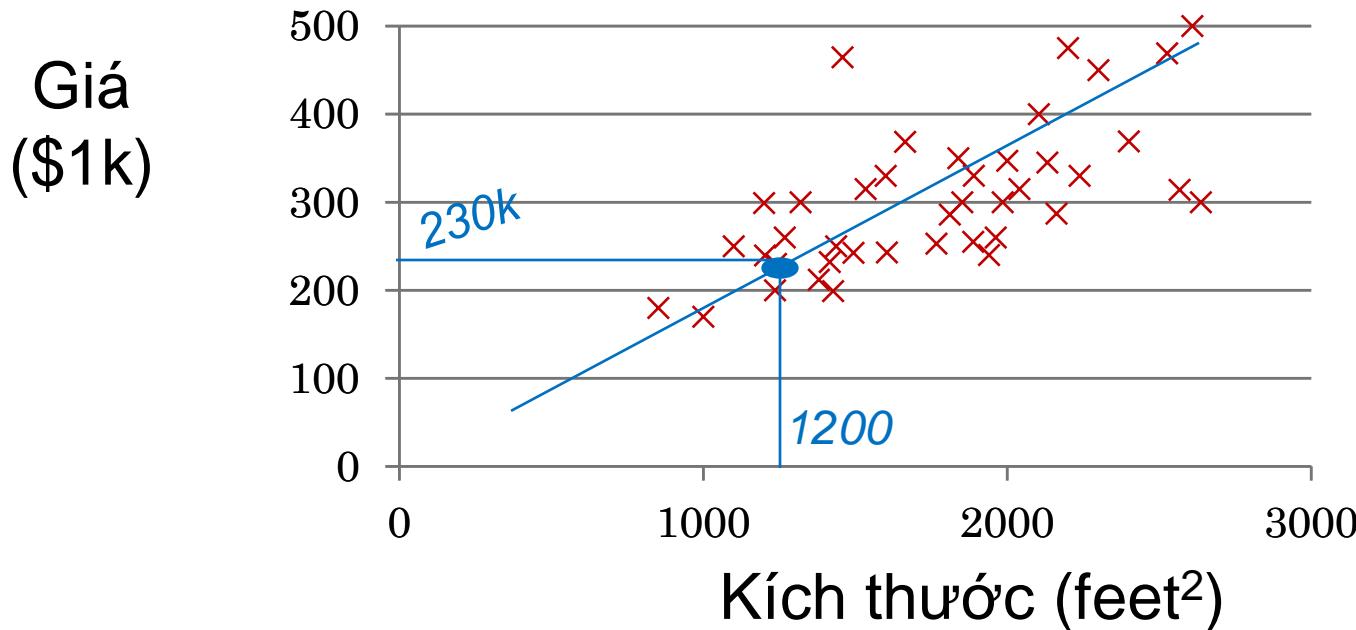
NỘI DUNG

1. Tổng quan về hồi qui
2. Hồi qui tuyển tính
3. Hồi qui phi tuyển
4. Ứng dụng
5. Các vấn đề với hồi qui
6. Tóm tắt

TÀI LIỆU THAM KHẢO

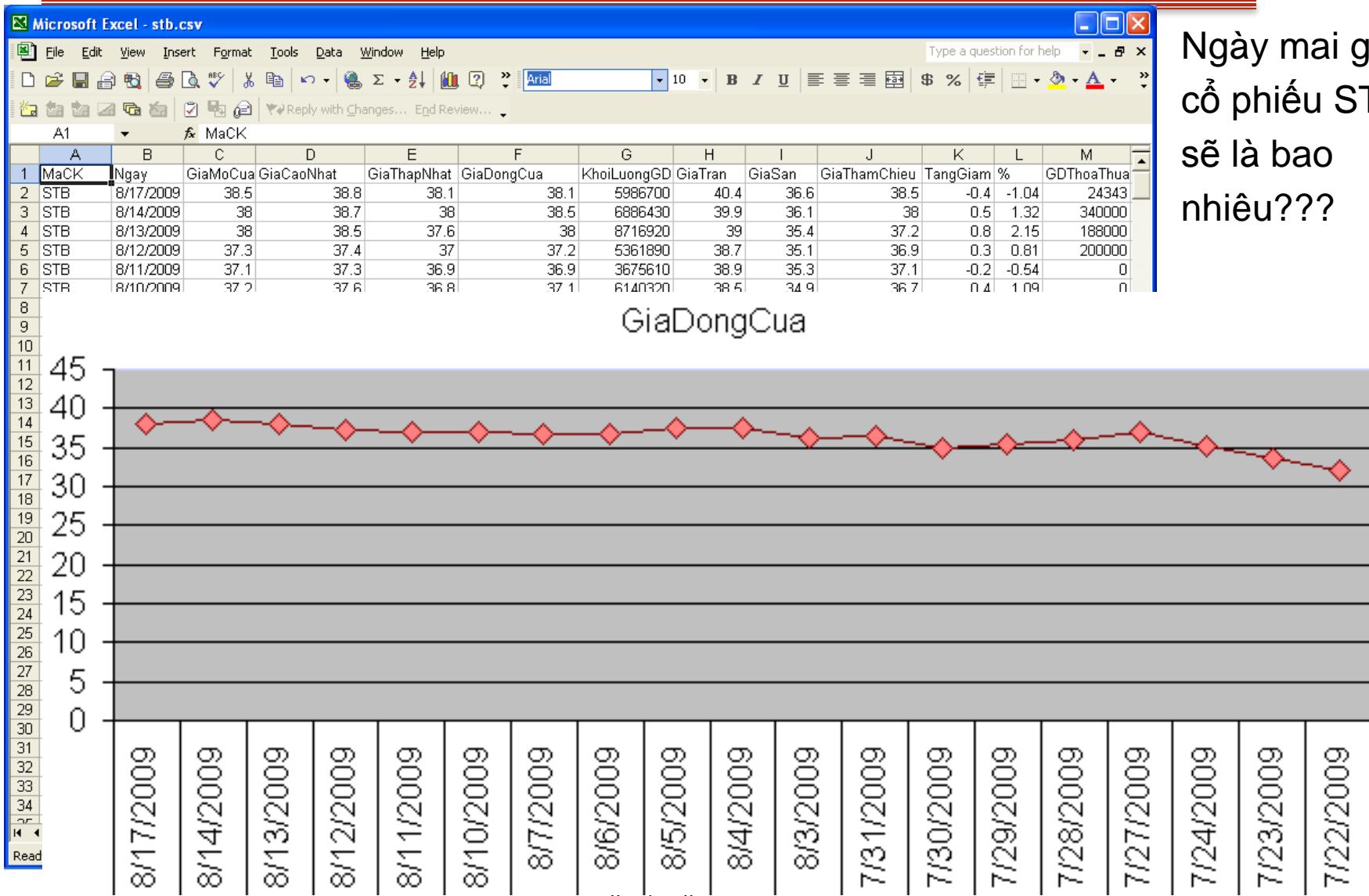
- [1] Jiawei Han, Micheline Kamber, and Jian Pei, “Data Mining: Concepts and Techniques”, 3rd Edition, Morgan Kaufmann Publishers, 2012.
- [2] David Hand, Heikki Mannila, Padhraic Smyth, “Principles of Data Mining”, MIT Press, 2001.
- [3] David L. Olson, Dursun Delen, “Advanced Data Mining Techniques”, 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. TỔNG QUAN VỀ HỒI QUI



- + Mô hình phân bố dữ liệu của giá nhà theo kích thước ?
- + Có thể dự định giá của căn nhà dựa vào kích thước hay không?

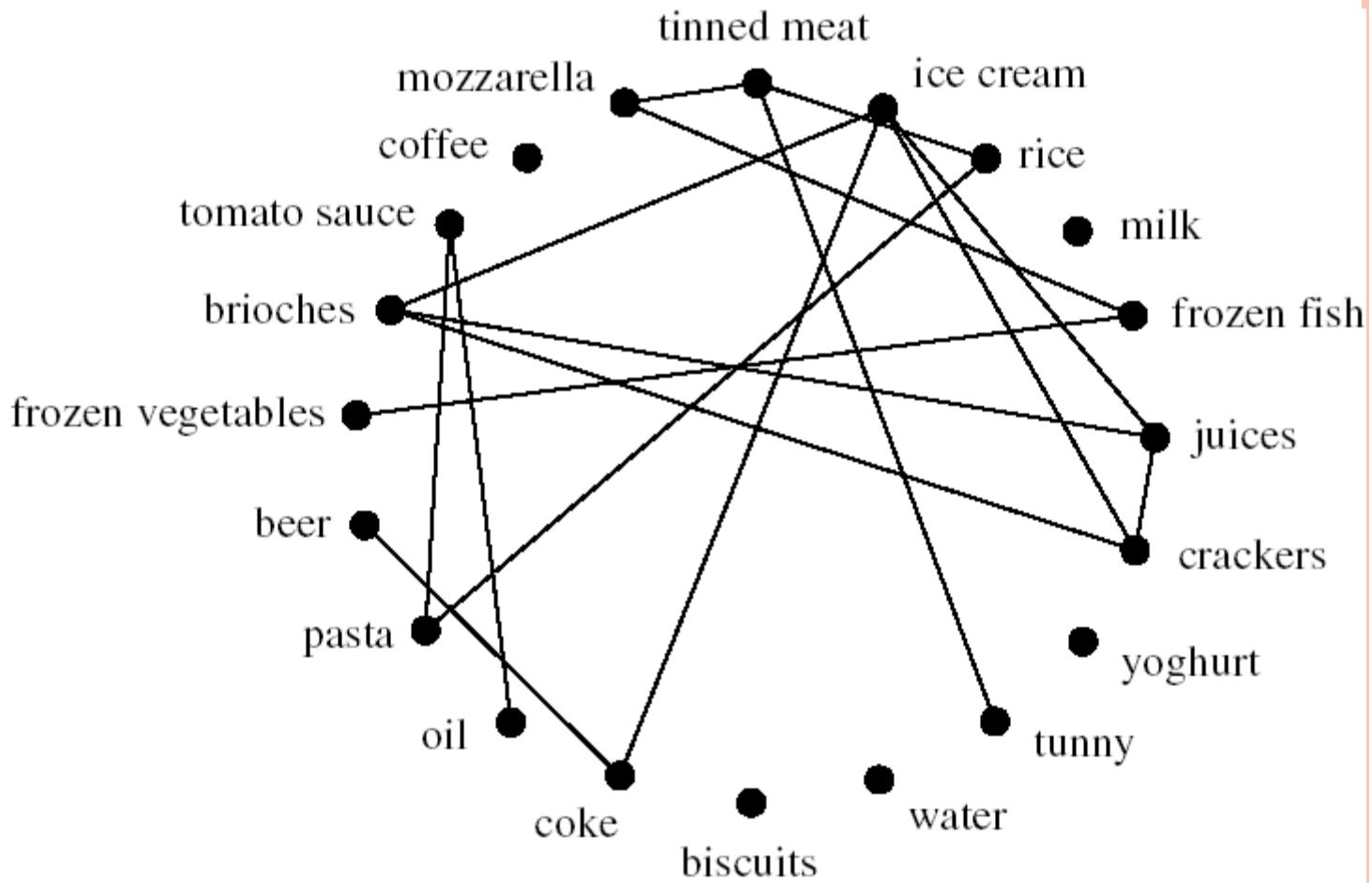
1. TỔNG QUAN VỀ HỒI QUI



1. TỔNG QUAN VỀ HÒI QUI

Bài toán phân tích giỏ hàng thị trường (market basket analysis)

→ sự kết hợp giữa các mặt hàng?



1. TỔNG QUAN VỀ HỒI QUI

- Khảo sát các yếu tố tác động chất lượng dịch vụ ngân hàng điện tử
 - Dễ sử dụng (+0.209)
 - Đáp ứng nhanh (+0.261)
 - Khả năng liên kết với các dịch vụ thanh toán (+0.199)
 - Cảm nhận về tính cá nhân (+0.15)
 - Cảm nhận về rủi ro (-0.25)
 - ...

1. TỔNG QUAN VỀ HỒI QUI

○ Định nghĩa - Hồi qui (regression)

- J. Han et al (2001, 2006): Hồi qui là kỹ thuật thống kê cho phép dự đoán các trị (số) liên tục
- Wiki (2009): Hồi qui (Phân tích hồi qui – regression analysis) là kỹ thuật thống kê cho phép ước lượng các mối liên kết giữa các biến
- R. D. Snee (1977): Hồi qui là kỹ thuật thống kê trong lĩnh vực phân tích dữ liệu và xây dựng các mô hình từ thực nghiệm, cho phép mô hình hồi qui vừa được khám phá được dùng cho mục đích dự báo (prediction), điều khiển (control), hay học (learn) cơ chế đã tạo ra dữ liệu

⇒ Hồi qui (regression): **dự đoán giá trị số (real-valued output)**

⇒ Phân lớp (classification): **dự đoán giá trị rời rạc (discrete values)**

1. TỔNG QUAN VỀ HỒI QUI

- Mô hình hồi qui (regression model): mô tả mối liên kết (relationship) giữa một tập các biến dự báo/độc lập (predictor/independent variables) và một hay nhiều biến đáp ứng/phụ thuộc (responses/dependent variables)
- Phương trình hồi quy $\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\Theta})$

\mathbf{X} : các biến dự báo/độc lập; giải thích sự biến đổi của các đáp ứng \mathbf{Y}

\mathbf{Y} : các biến đáp ứng/phụ thuộc; mô tả các hiện tượng được quan tâm

$\boldsymbol{\Theta}$: các hệ số hồi qui (regression coefficients); mô tả sự ảnh hưởng tương đối của \mathbf{X} đối với \mathbf{Y}

1. TỔNG QUAN VỀ HỒI QUI

○ Phân loại

- tuyến tính (linear) và phi tuyến (nonlinear)
 - ✓ Linear in parameters: kết hợp tuyến tính các thông số tạo nên Y
 - ✓ Nonlinear in parameters: kết hợp phi tuyến các thông số tạo nên Y
- đơn biến (single variable) và đa biến (multiple variables)
 - ✓ Single: $X = (X_1)$ Vs. Multiple: $X = (X_1, X_2, \dots, X_k)$
- có thông số (parametric), phi thông số (nonparametric), và thông số kết hợp (semiparametric)
- đối称 (symmetric) và bất đối称 (asymmetric)
 - ✓ Symmetric: mô hình hồi qui có tính mô tả (descriptive) (eg. log-linear models)
 - ✓ Asymmetric: mô hình hồi qui có tính dự báo (predictive) (eg. generalized linear models)

1. TỔNG QUAN VỀ HỒI QUI

- Hồi qui có thông số (parametric), phi thông số (nonparametric), và thông số kết hợp (semiparametric)
 - Parametric: mô hình hồi qui với hữu hạn thông số
 - Nonparametric: mô hình hồi qui với vô hạn thông số
 - Semiparametric: mô hình hồi qui với hữu hạn thông số được quan tâm

Mô hình hồi qui	Dạng mô tả
Parametric	$Y = \theta_0 + \theta_1 * X$
Nonparametric	$Y = \theta_0 + f(X)$
Semiparametric	$Y = \theta_0 + \theta_1 * X_1 + f(X_2)$

2. HỒI QUI TUYẾN TÍNH

- Hồi qui tuyến tính đơn biến (Univariate)
- Hồi qui tuyến tính đa biến (Multivariate)

2.1. HỎI QUI TUYẾN TÍNH ĐƠN BIỀN

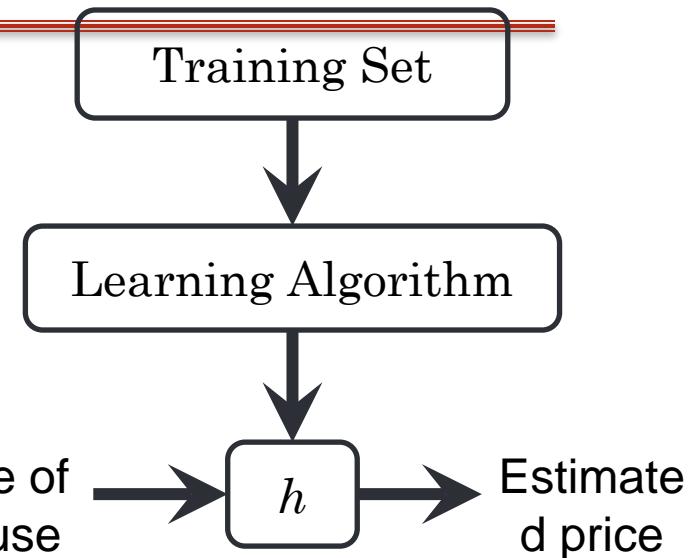
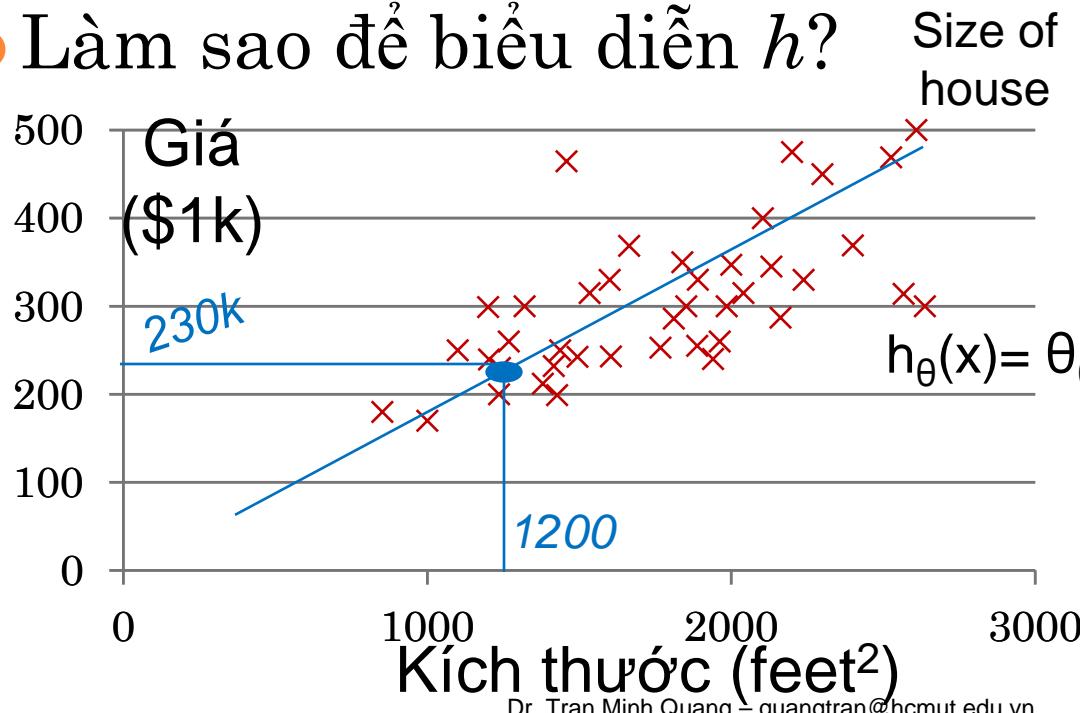
○ Ký hiệu

- ✓ N: số lượng mẫu học (size of training examples)
- ✓ x: biến đầu vào (input variable/feature)
- ✓ y: biến đích (output/target variable)
- ✓ $(x^{(i)}, y^{(i)})$: mẫu học thứ i
- ✓ $(x^{(1)}, y^{(1)}) = (2100, 450)$

Kích thước feet ² (x)	Giá (\$1k) (y)
2100	450
1416	232
1534	315
852	178
...	...

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

- *Hypothesis* (h): mô hình dự đoán giả định
- $y = h(x)$; h là một phép ánh xạ từ x sang y
- Làm sao để biểu diễn h ?

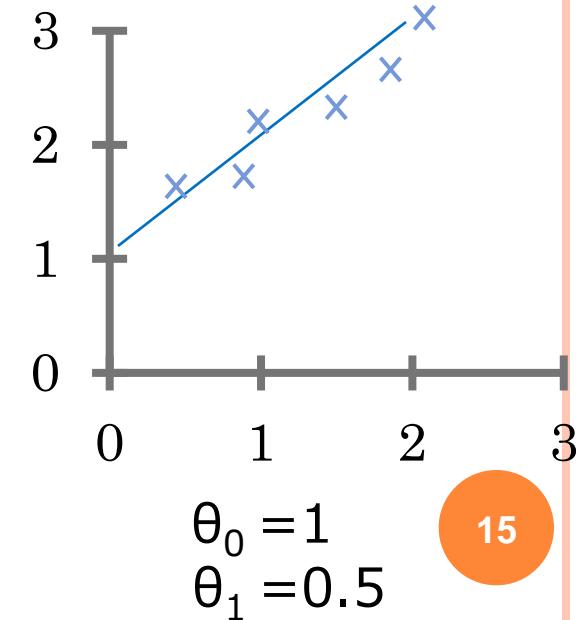
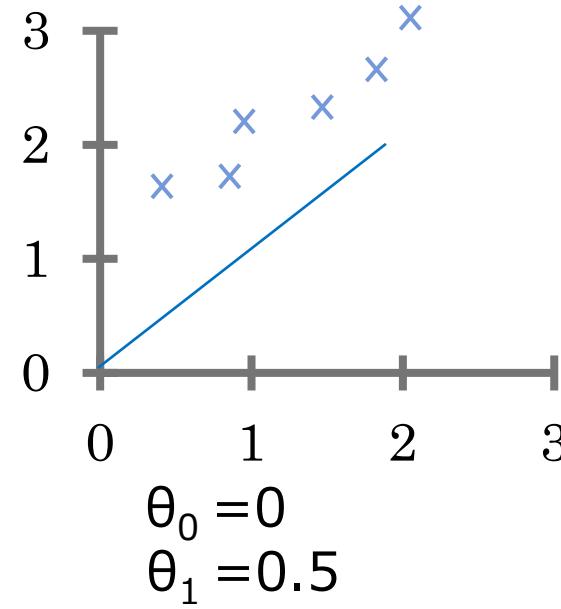
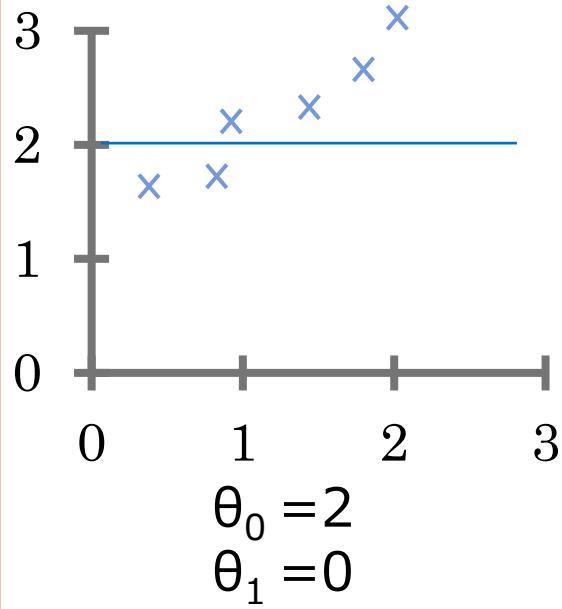


Một biến (univariate)

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Hypothesis (h): $h_{\theta}(x) = \theta_0 + \theta_1 x \Rightarrow$ Xác định θ_i ?
- Phương pháp “thử và sai”, kiểm tra khả năng mô tả của đường hồi qui cho tập dữ liệu mẫu?

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



2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Chọn (θ_0, θ_1) sao cho $h_{\theta}(x^{(i)}) \approx y^{(i)}$; $i=1\dots N$
 - Sai số/thặng dư (residual/prediction error)

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

- MSE

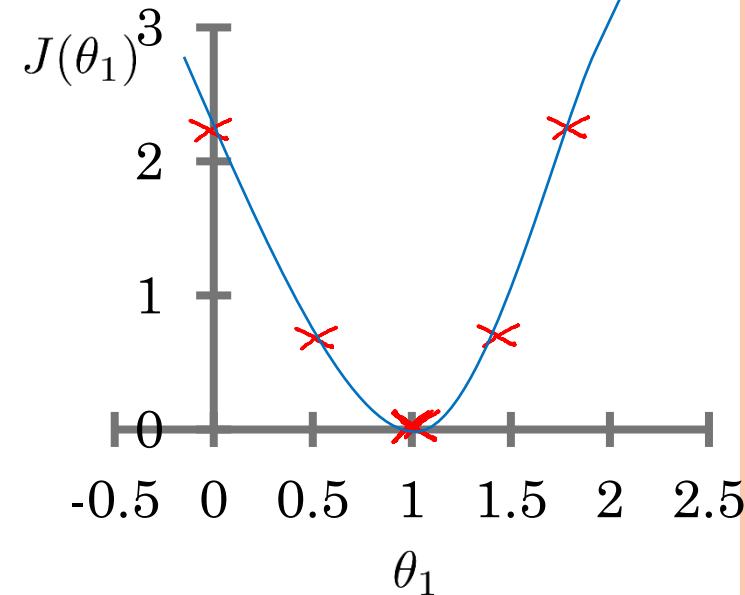
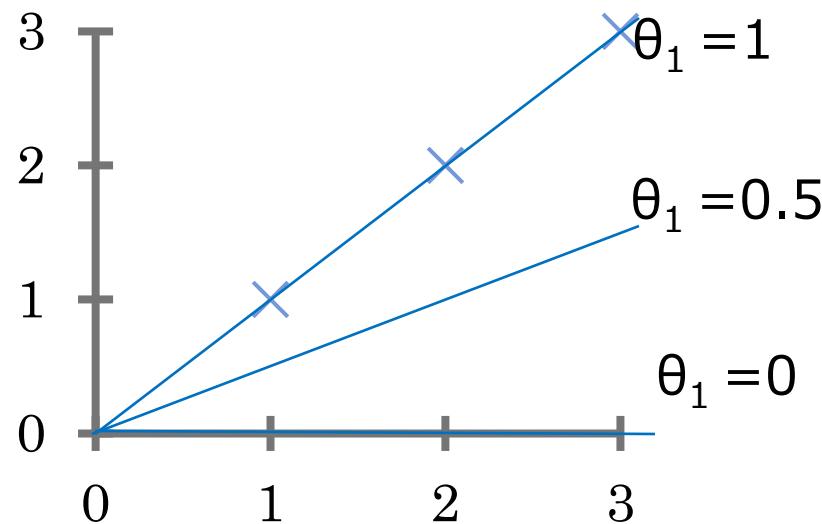
$$MSE = \frac{1}{N} \sum_{i=1}^N (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

- Cost function $J(\theta_0, \theta_1) \Rightarrow$ **cực tiểu hóa (minimize)**

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

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

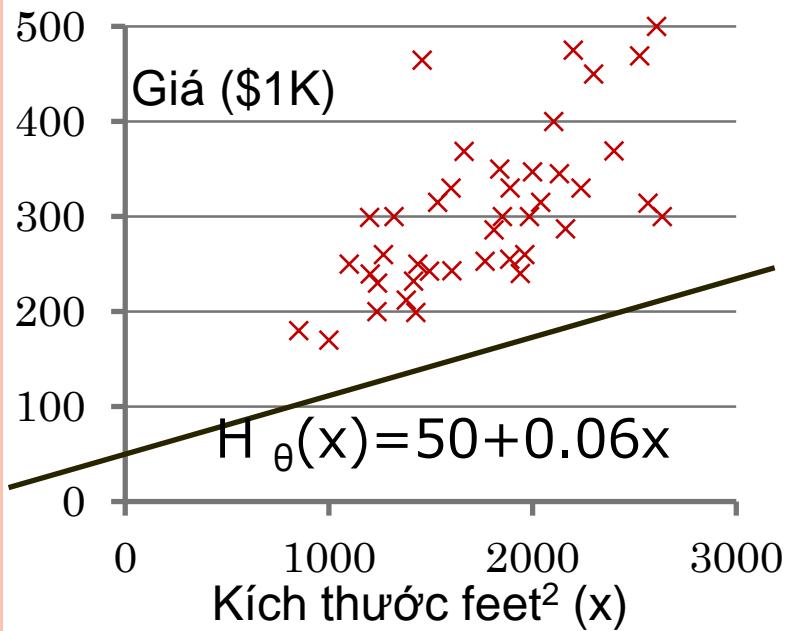
- Xét trường hợp đơn giản $\theta_0=0$, $h_\theta(x) = \theta_1 x$



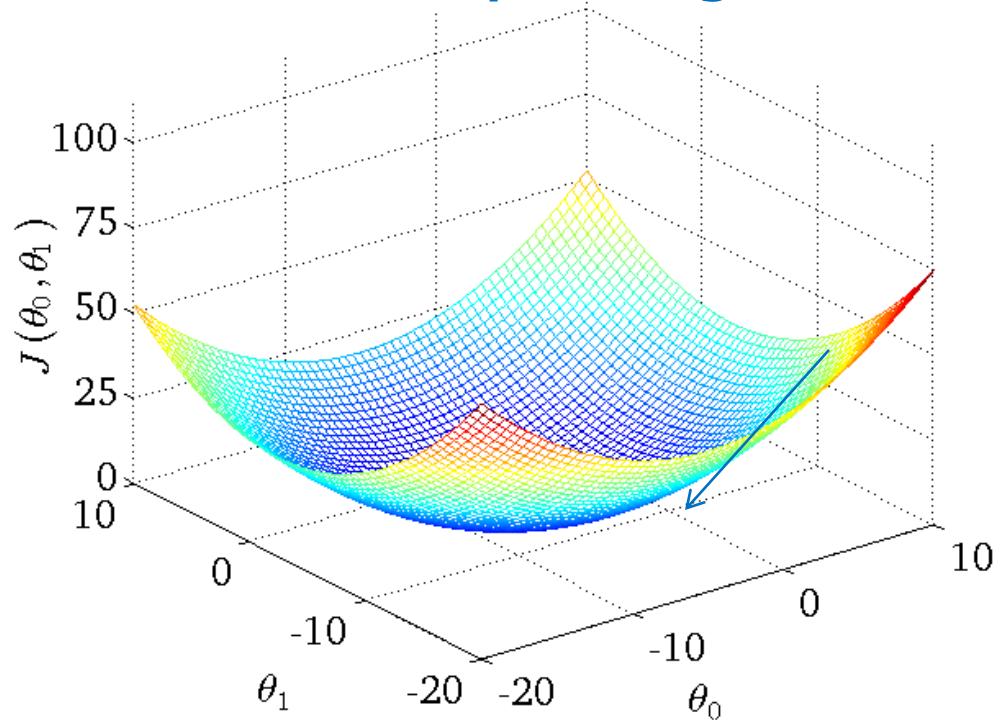
- $\theta_1=1 \Rightarrow J(\theta_1) = 0$; $\theta_1=0.5 \Rightarrow J(\theta_1) = 0.58$;
 $\theta_1=0 \Rightarrow J(\theta_1) = 2.3$

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Một ví dụ với $h_{\theta}(x) = \theta_0 + \theta_1 x$



Contour plots/figures

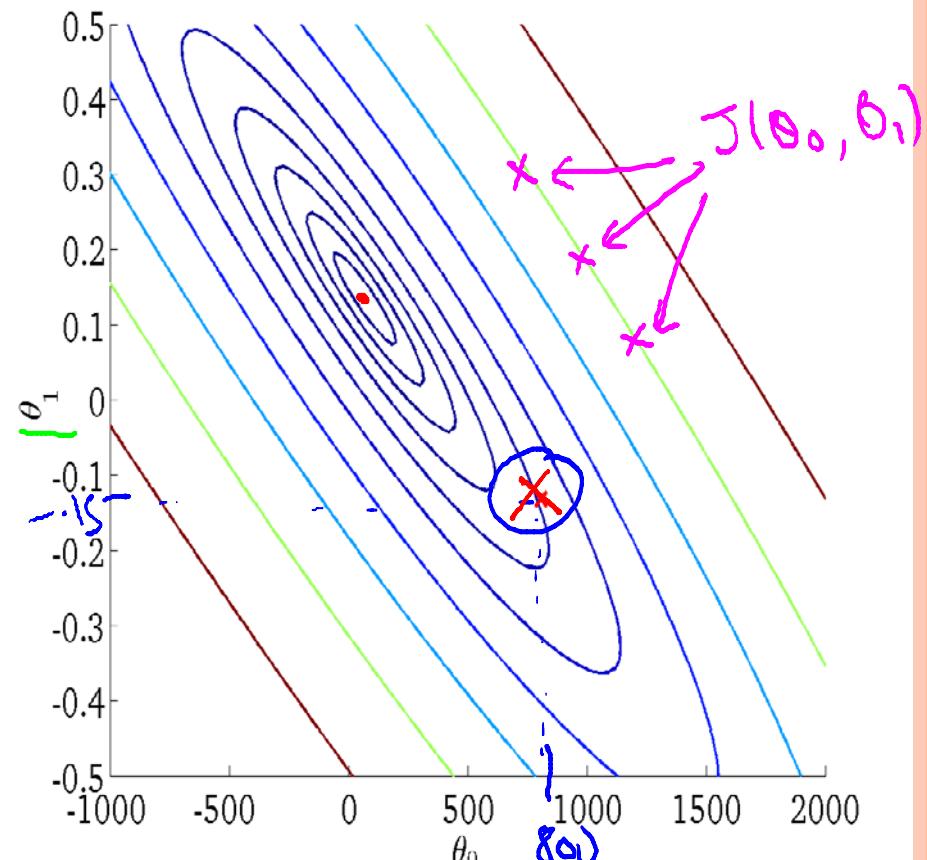
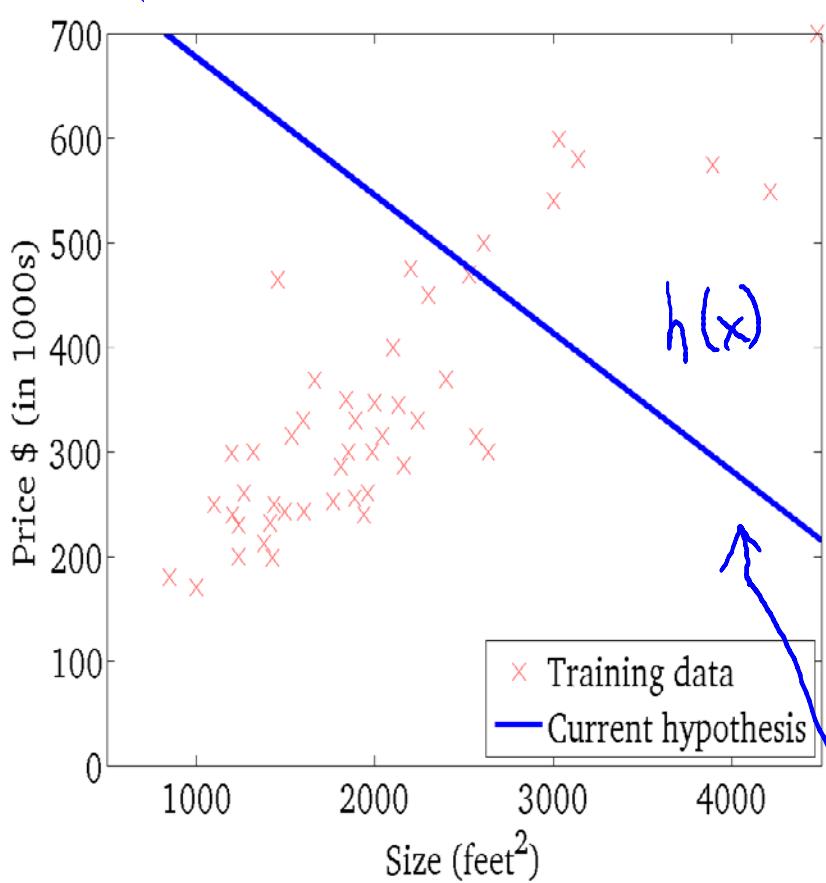


$$h_{\theta}(x)$$

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

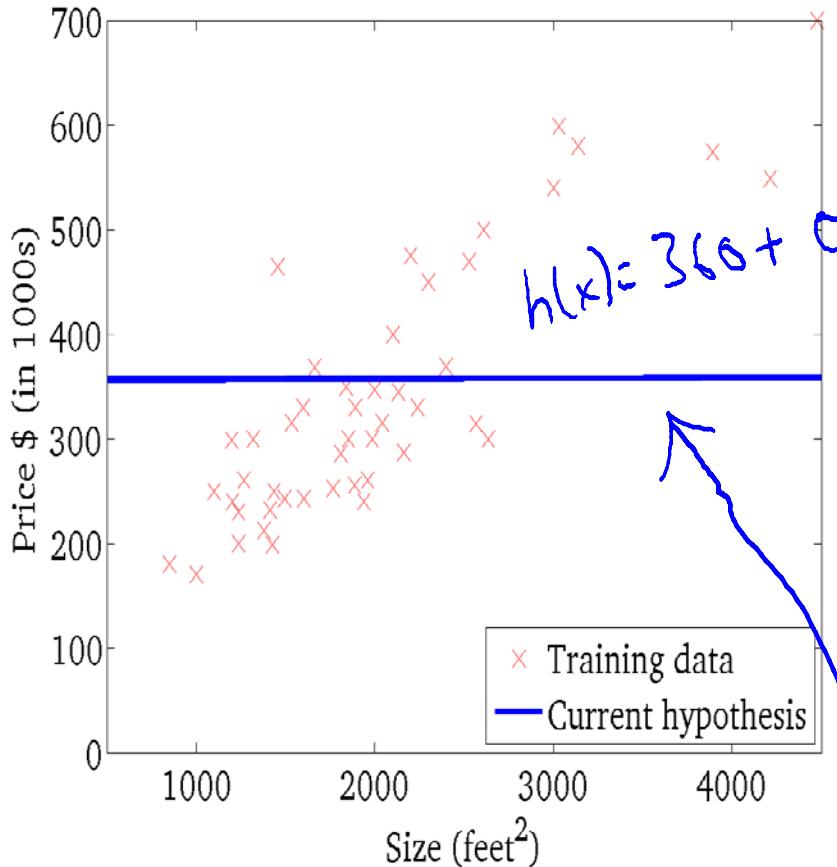
(for fixed θ_0, θ_1 , this is a function of x)

(function of the parameters θ_0, θ_1)



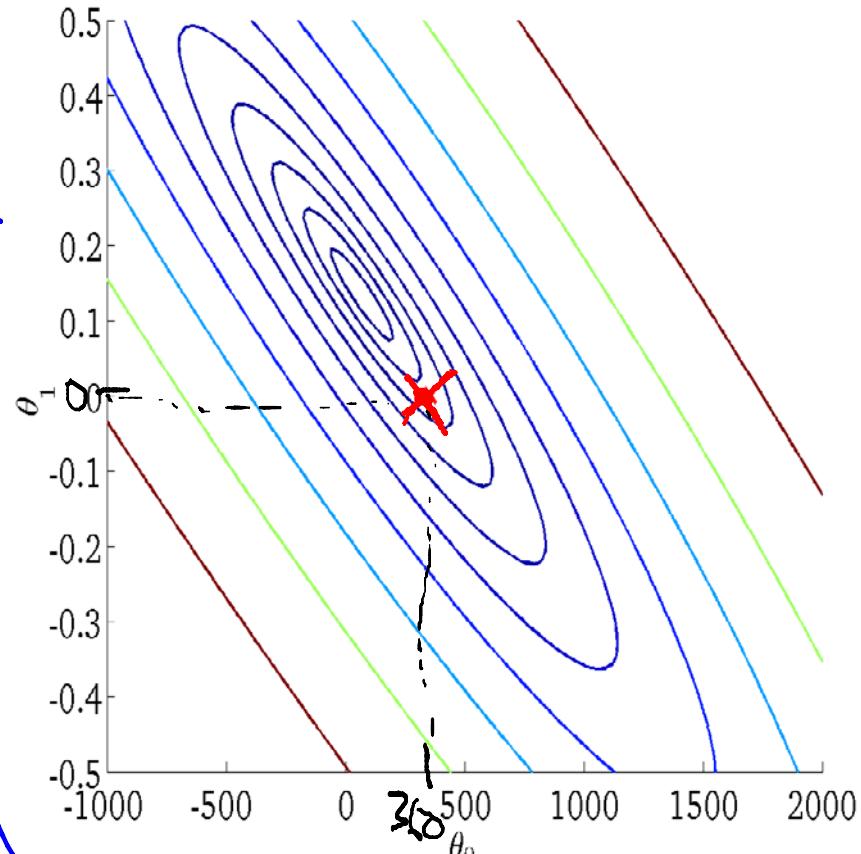
$$h_{\theta}(x)$$

(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)

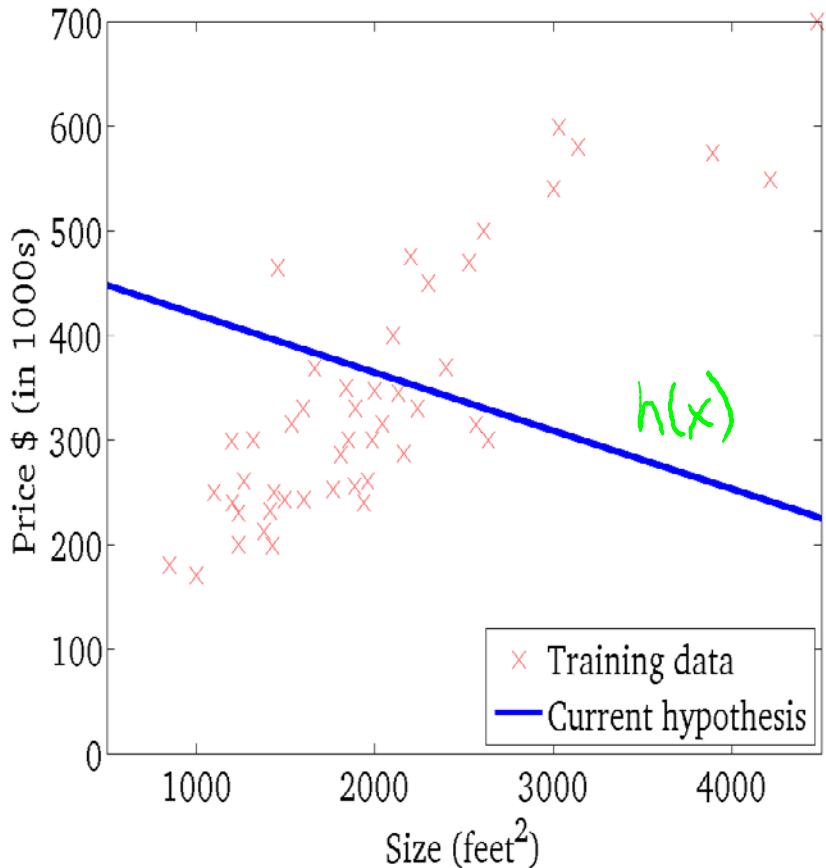


$$\begin{cases} \theta_0 = 360 \\ \theta_1 = 0 \end{cases}$$

Source: Andrew Ng

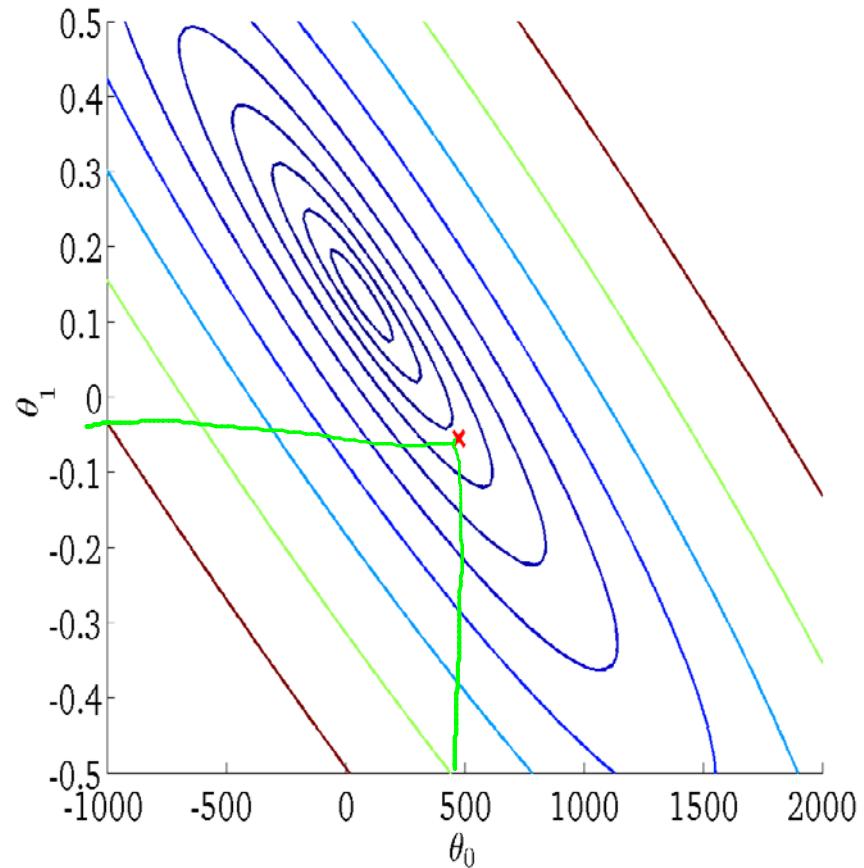
$$h_{\theta}(x)$$

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



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

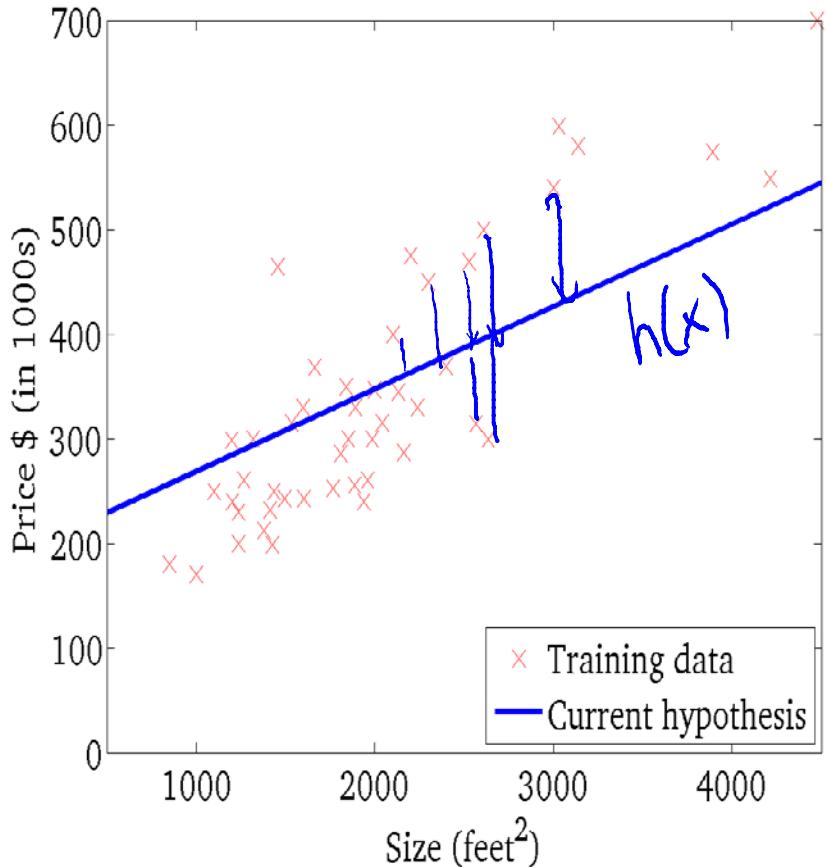
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

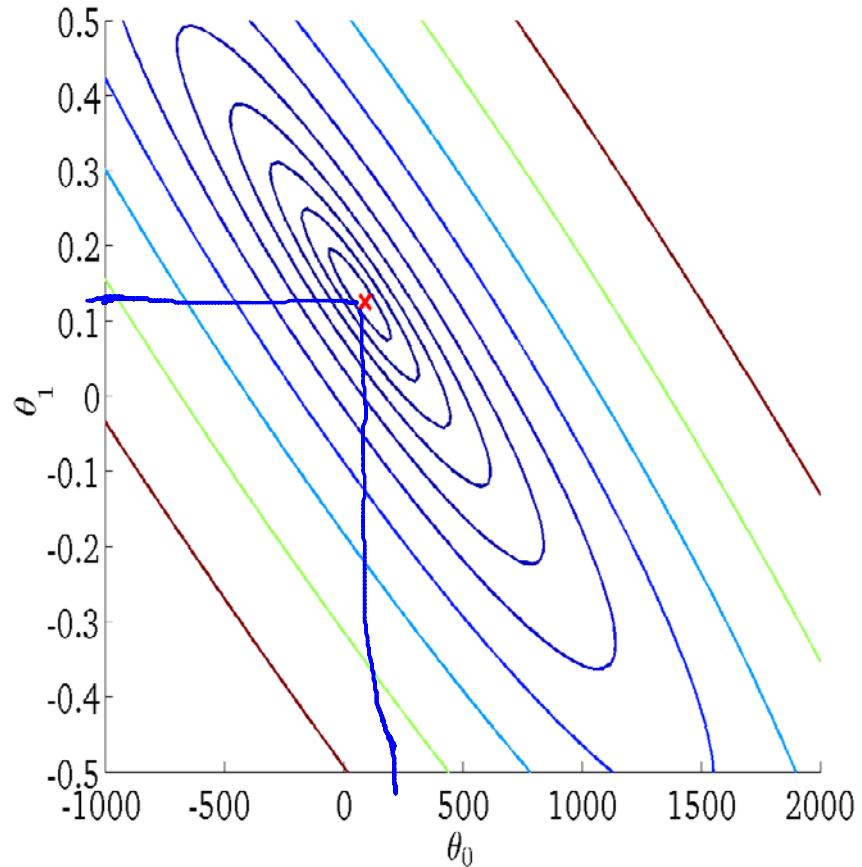
$$h_{\theta}(x)$$

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



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

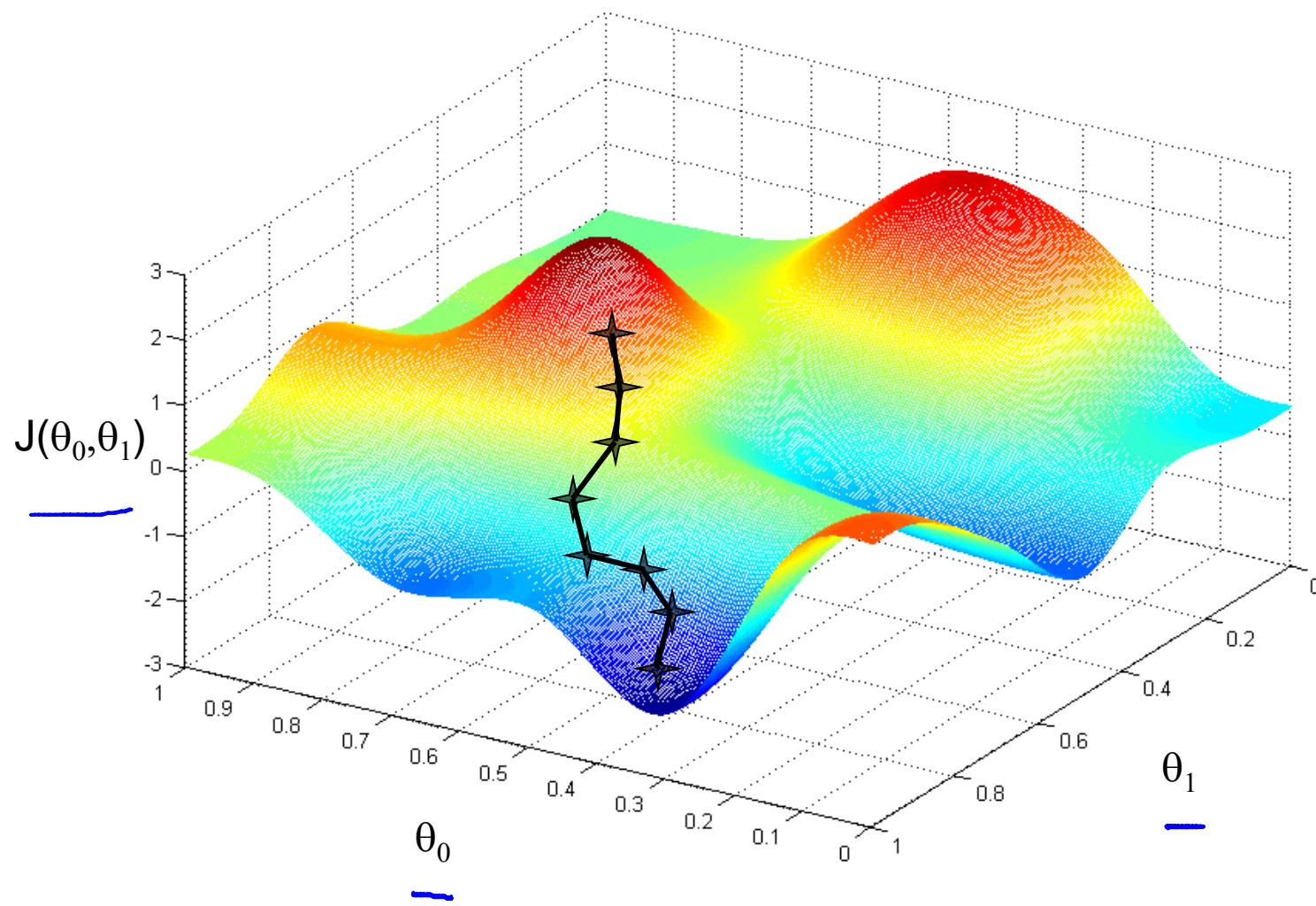
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

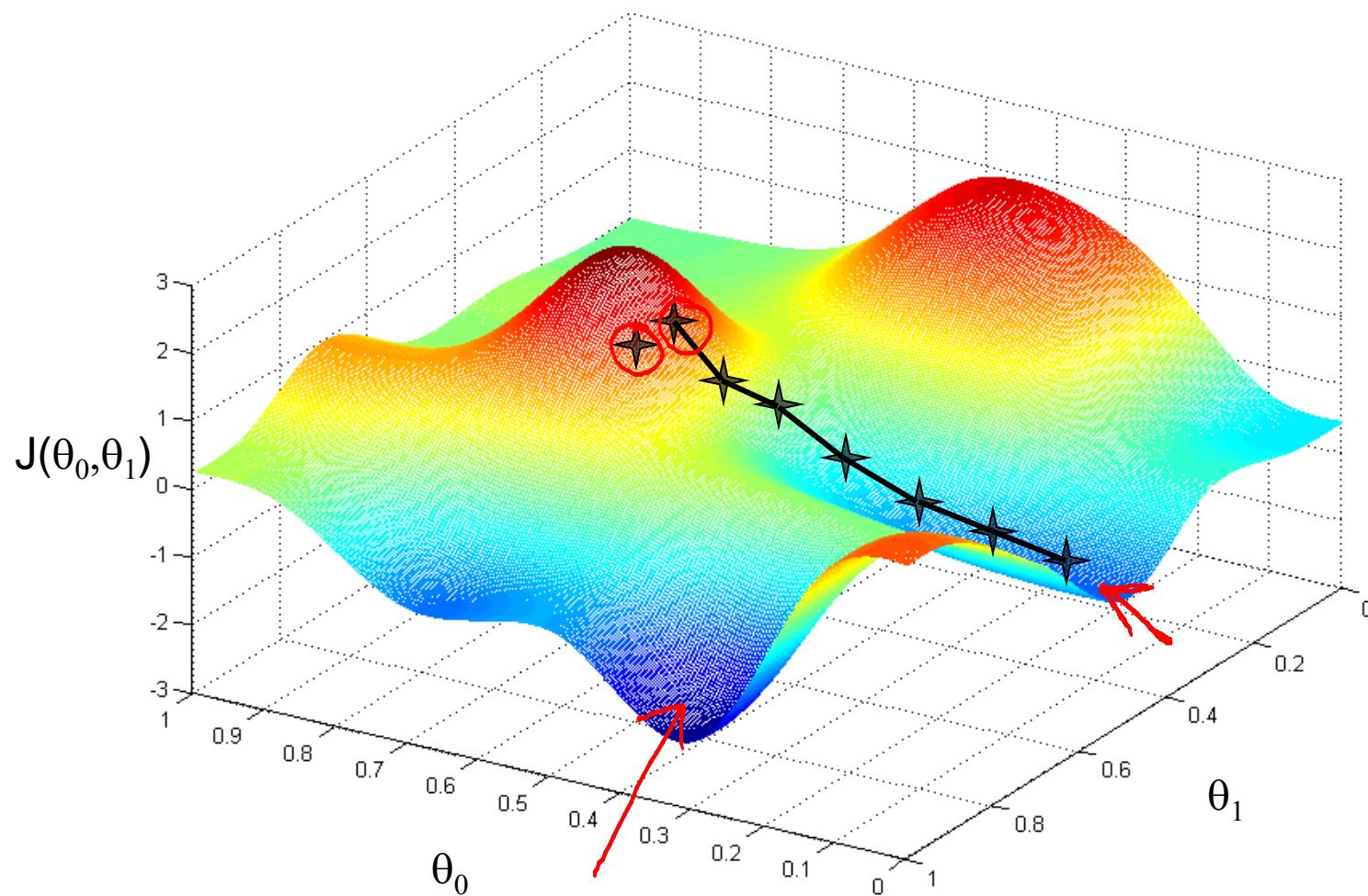
2.1. HÒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Kỹ thuật “xuống đồi” (Gradient descent) => tìm điểm tối ưu (minimize $J(\theta_0, \theta_1)$)
- Tổng quát:
 - ✓ Bắt đầu với giá trị ngẫu nhiên (θ_0, θ_1) , ex. $(\theta_0=0, \theta_1=0)$
 - ✓ Thay đổi bộ (θ_0, θ_1) nhằm giảm $J(\theta_0, \theta_1)$
 - ✓ Lặp lại bước trên cho đến khi ta “tin rằng” $J(\theta_0, \theta_1)$ là nhỏ nhất (minimum)



Source: Andrew Ng





Source: Andrew Ng

2.1. HÒI QUI TUYẾN TÍNH ĐƠN BIỀN

o Giải thuật Gradient descent)

Repeat until convergence{

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

}

↑
Learning rate

Đúng: Cập nhật đồng thời

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

Sai:

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

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Giải thuật Gradient descent): minimize $J(\theta_0, \theta_1)$

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

Repeat until convergence {

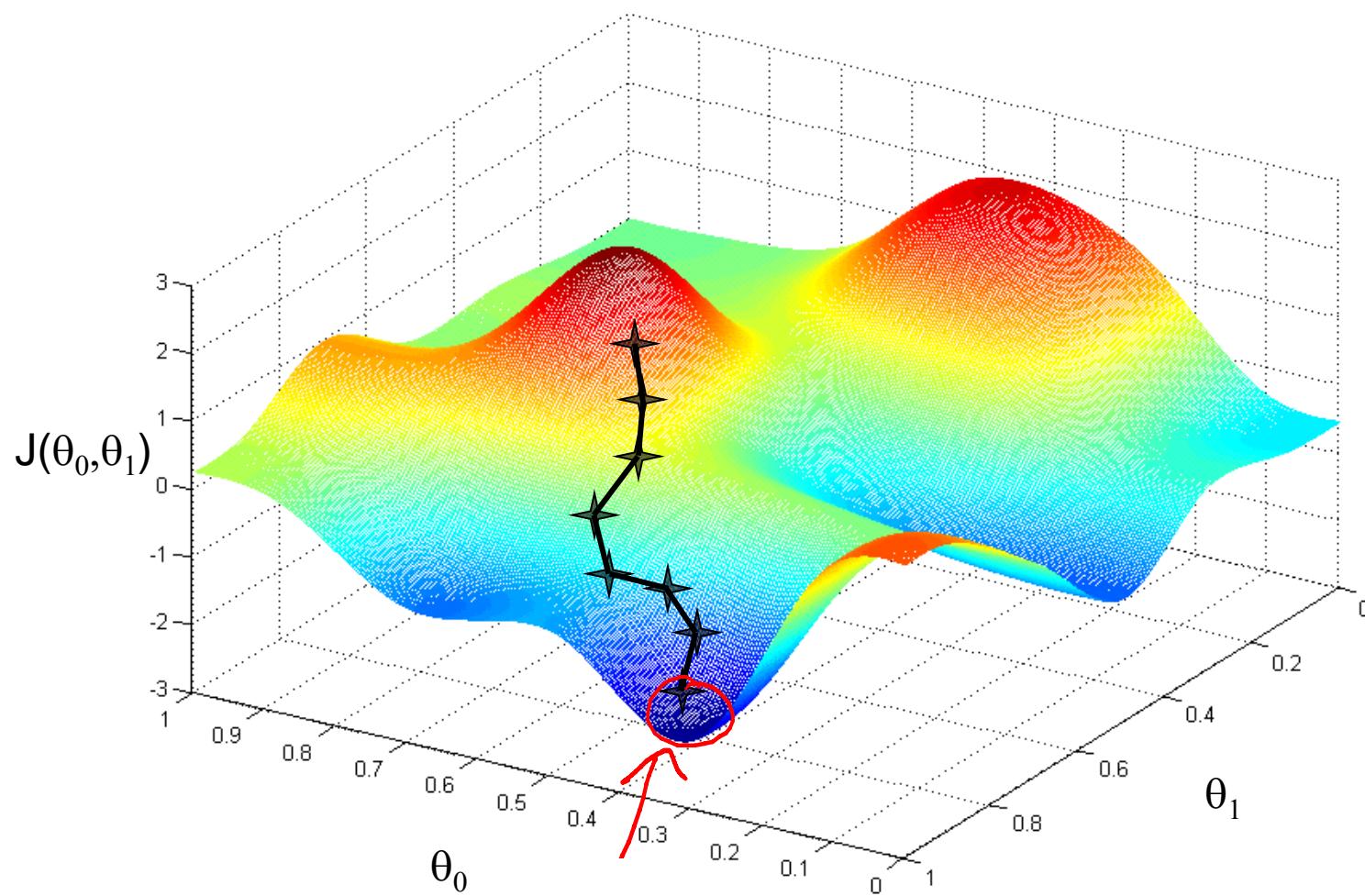
$$\theta_0 = \theta_0 - \alpha \frac{1}{N} \sum_{i=1}^N (h_\theta(x^{(i)}) - y^{(i)})$$

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

}

Update θ_0 and θ_1 simultaneously

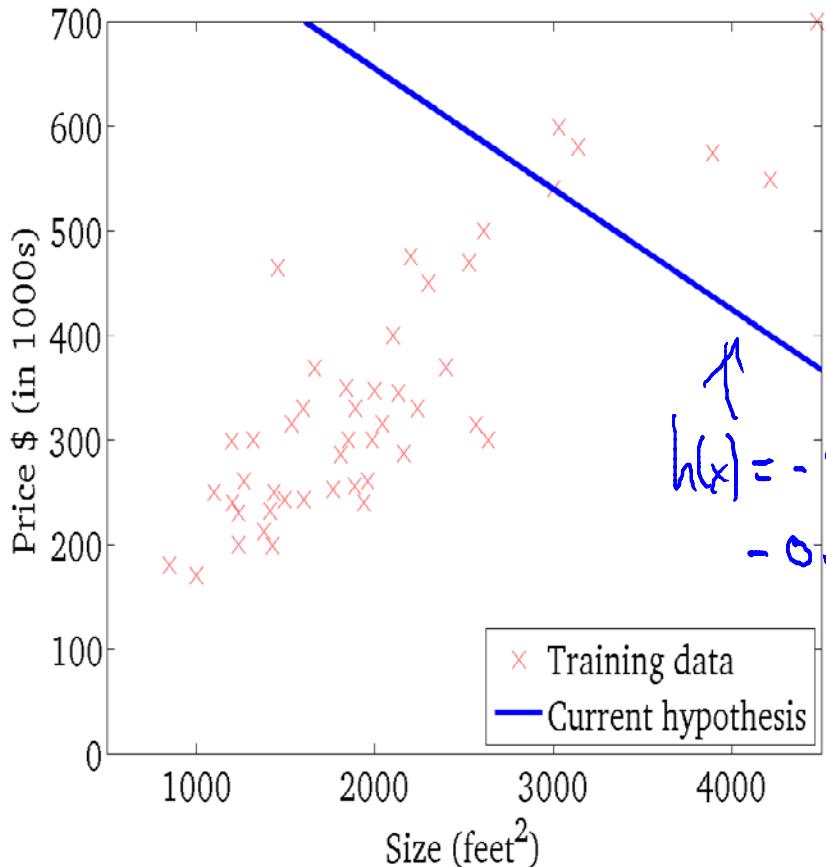
$$\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$



Source: Andrew Ng

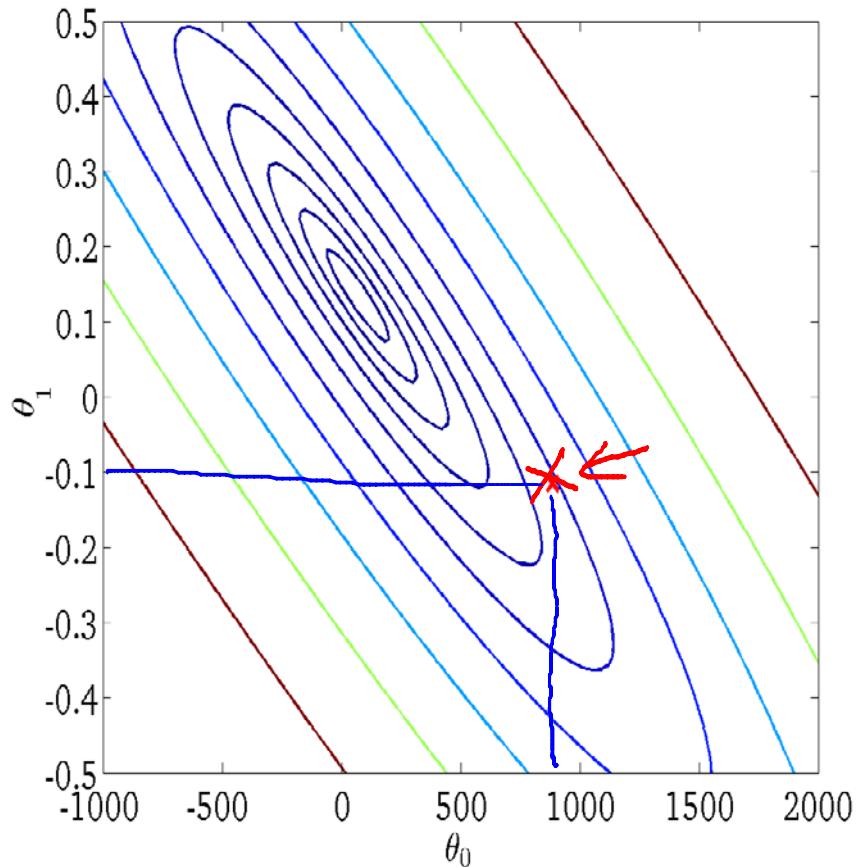
$$h_{\theta}(x)$$

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



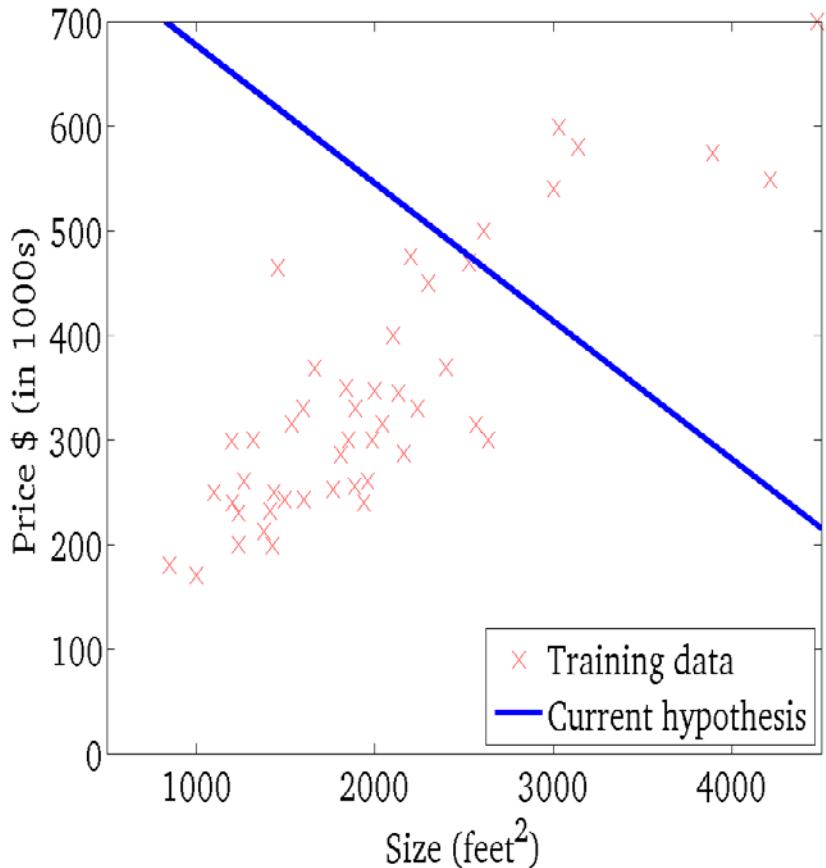
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



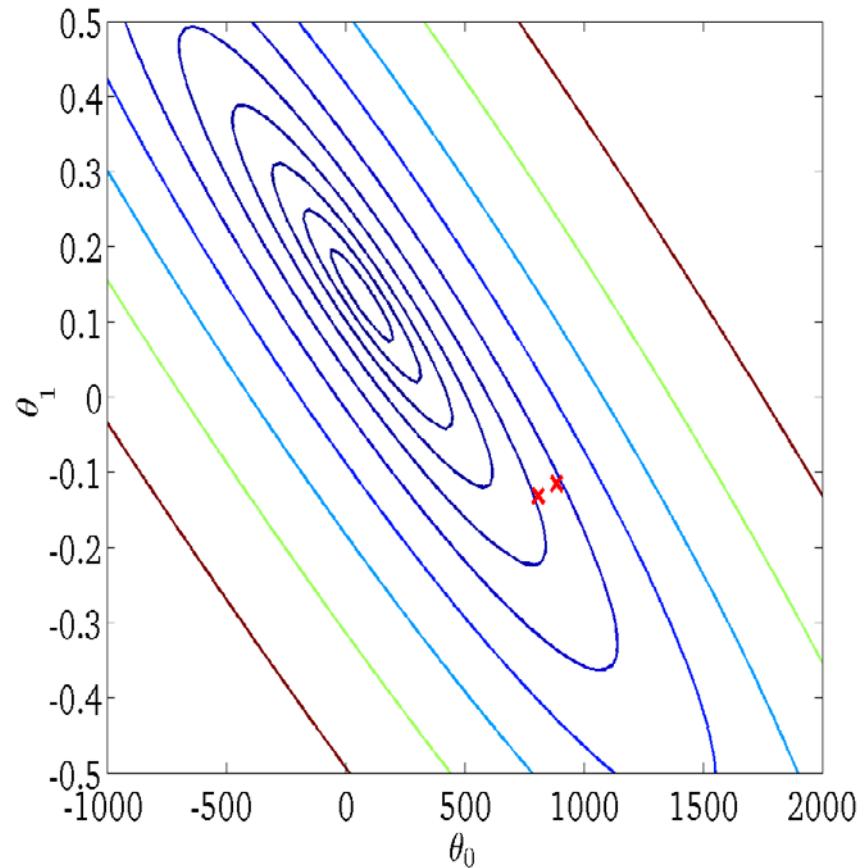
$$h_{\theta}(x)$$

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



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

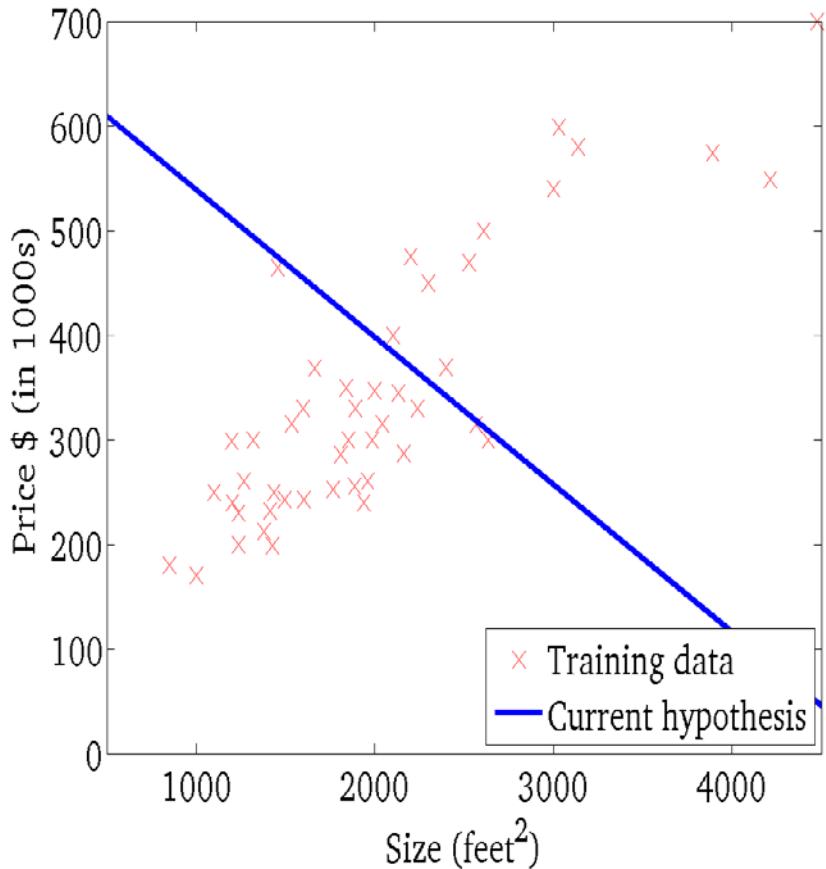
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

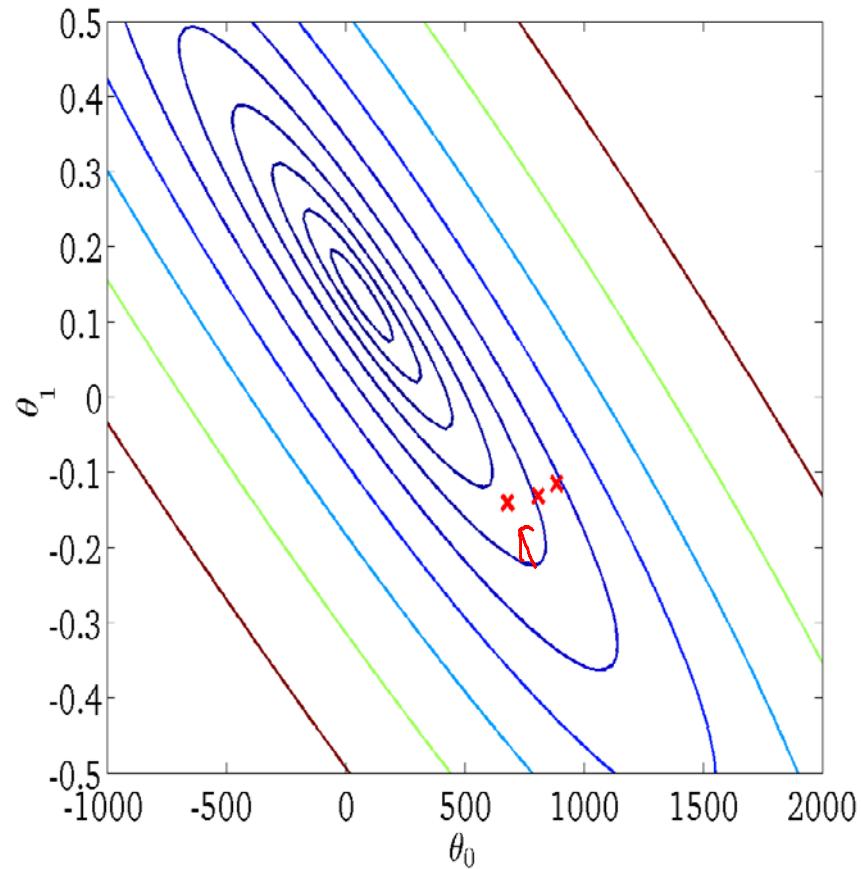
$$h_{\theta}(x)$$

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



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

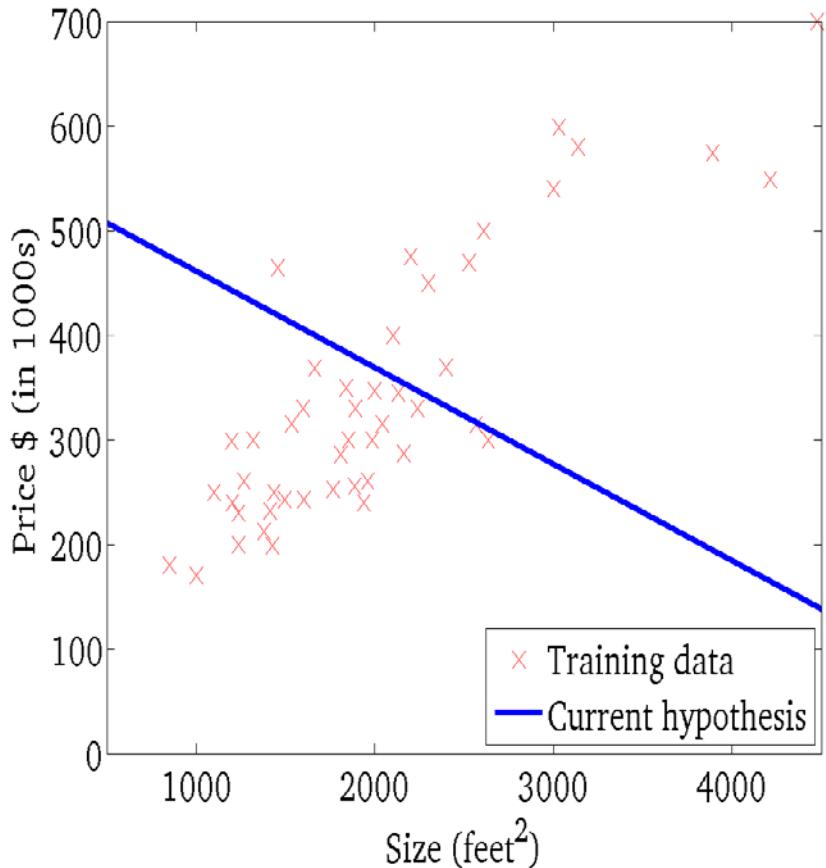
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

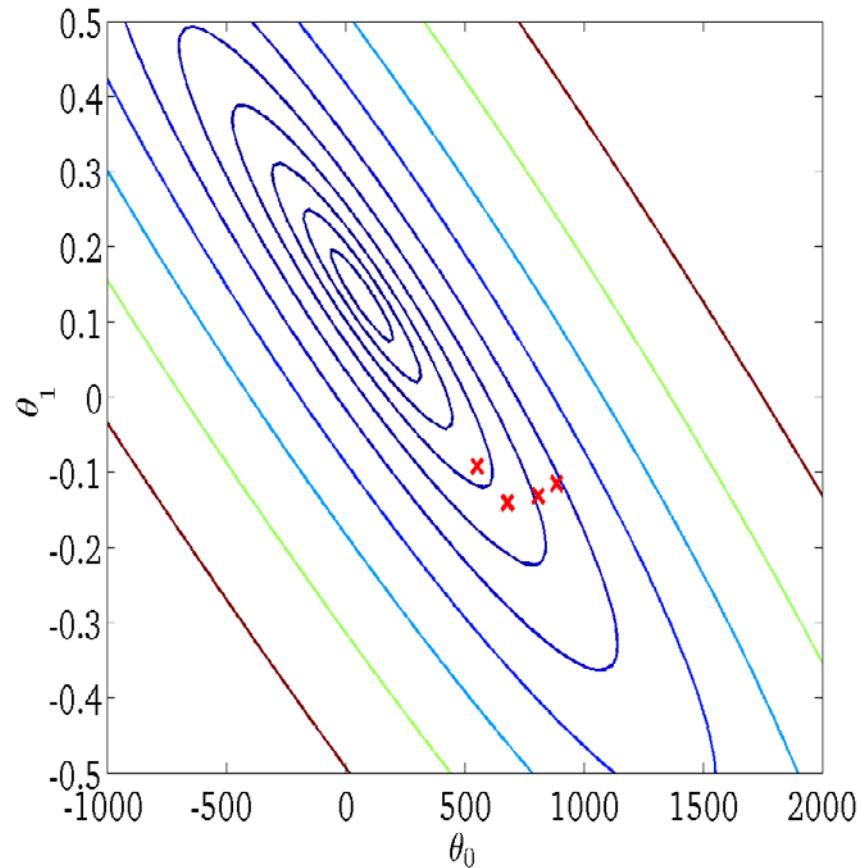
$$h_{\theta}(x)$$

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



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

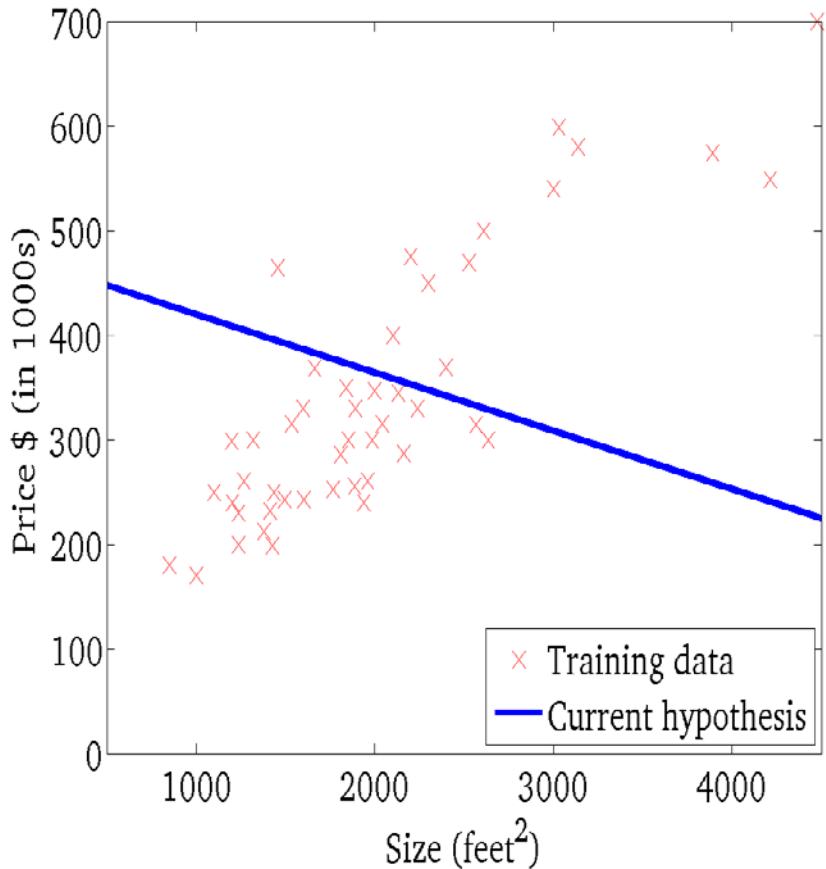
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

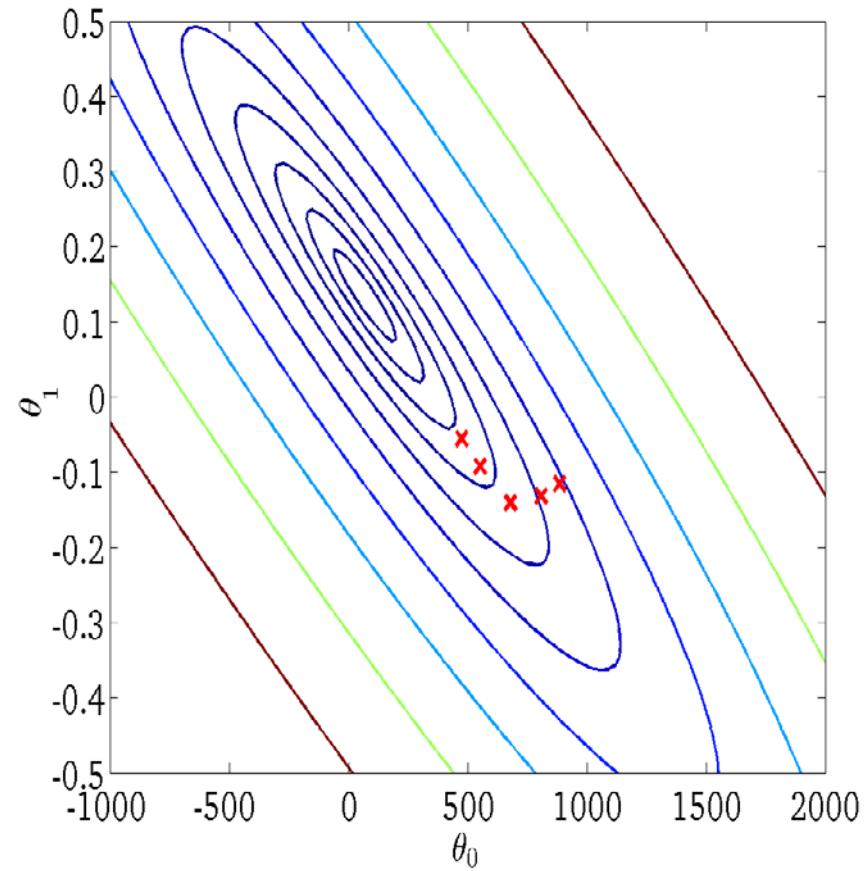
$$h_{\theta}(x)$$

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



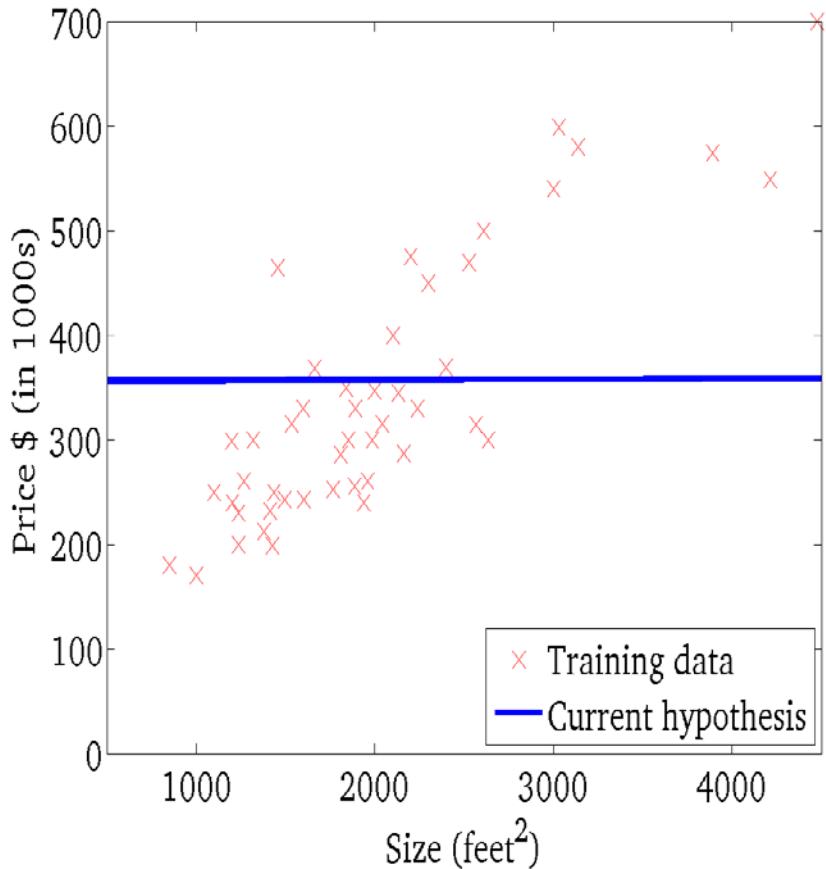
$$J(\theta_0, \theta_1)$$

(function of the parameters θ_0, θ_1)



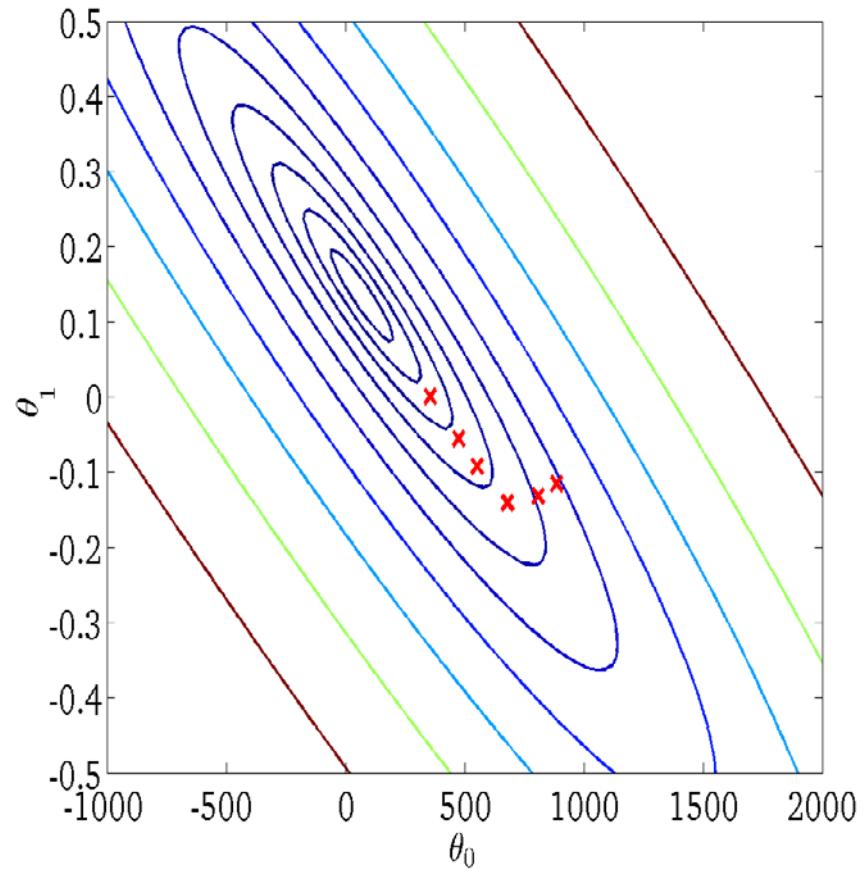
$$h_{\theta}(x)$$

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



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

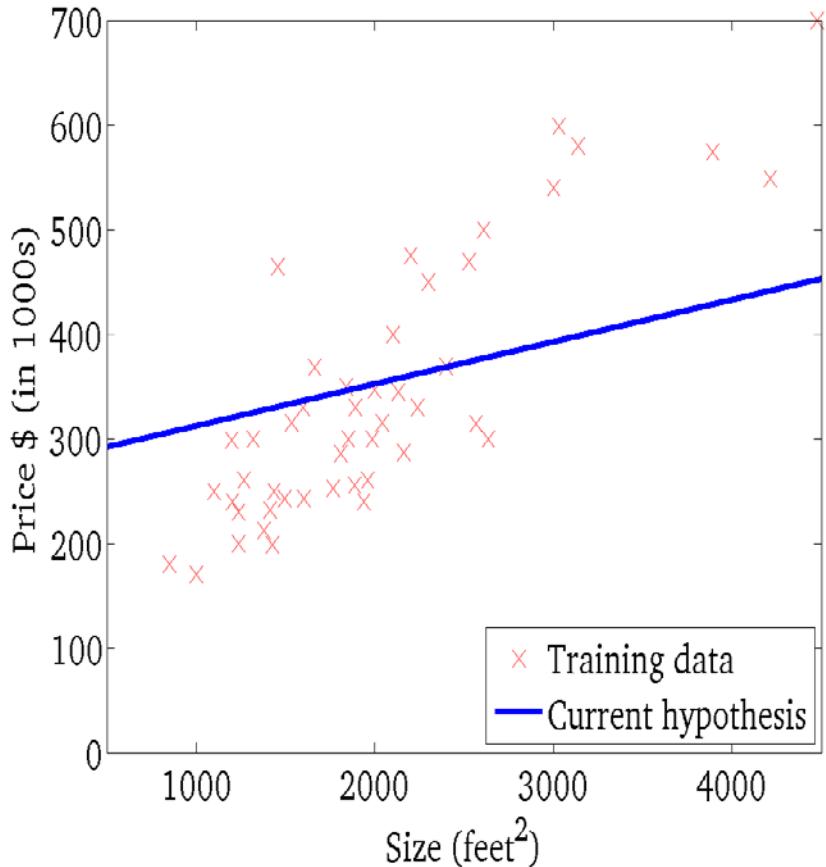
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

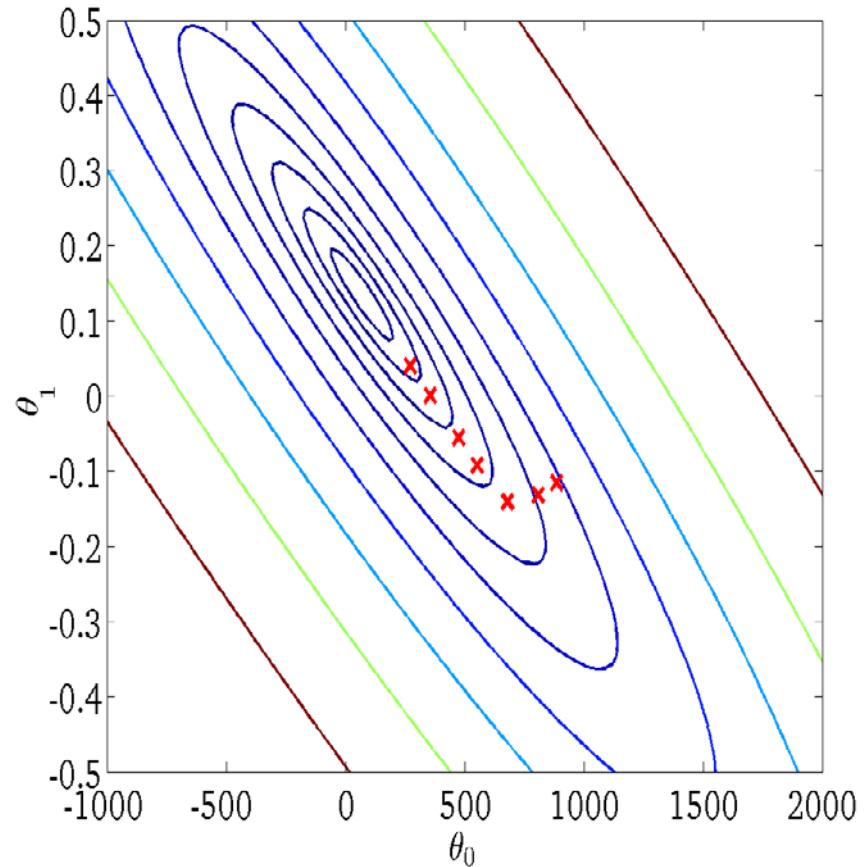
$$h_{\theta}(x)$$

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



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

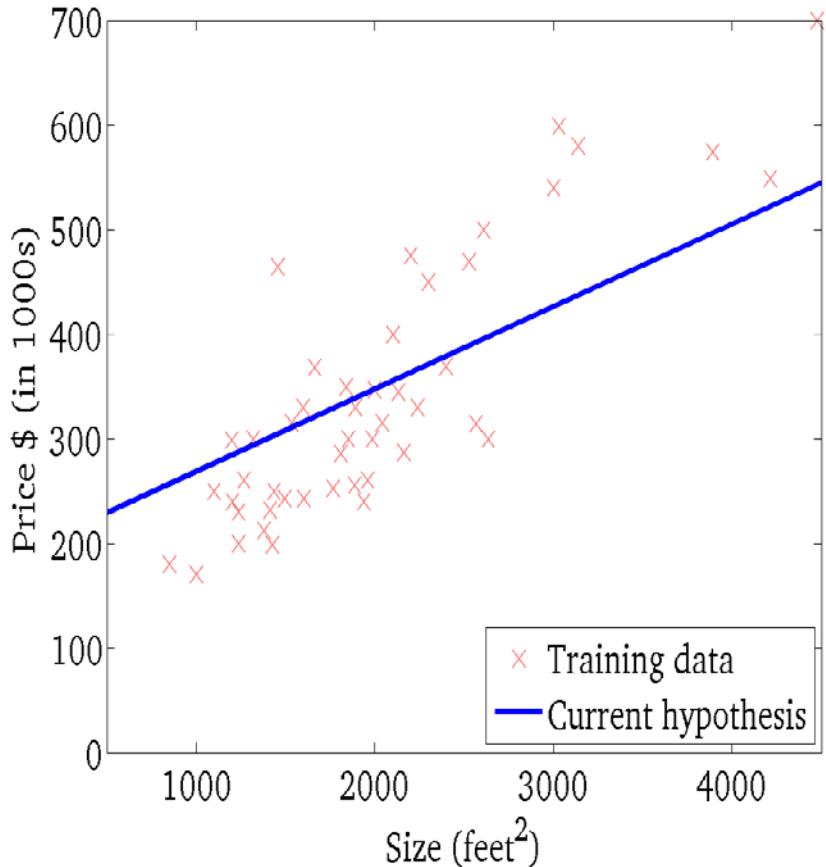
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

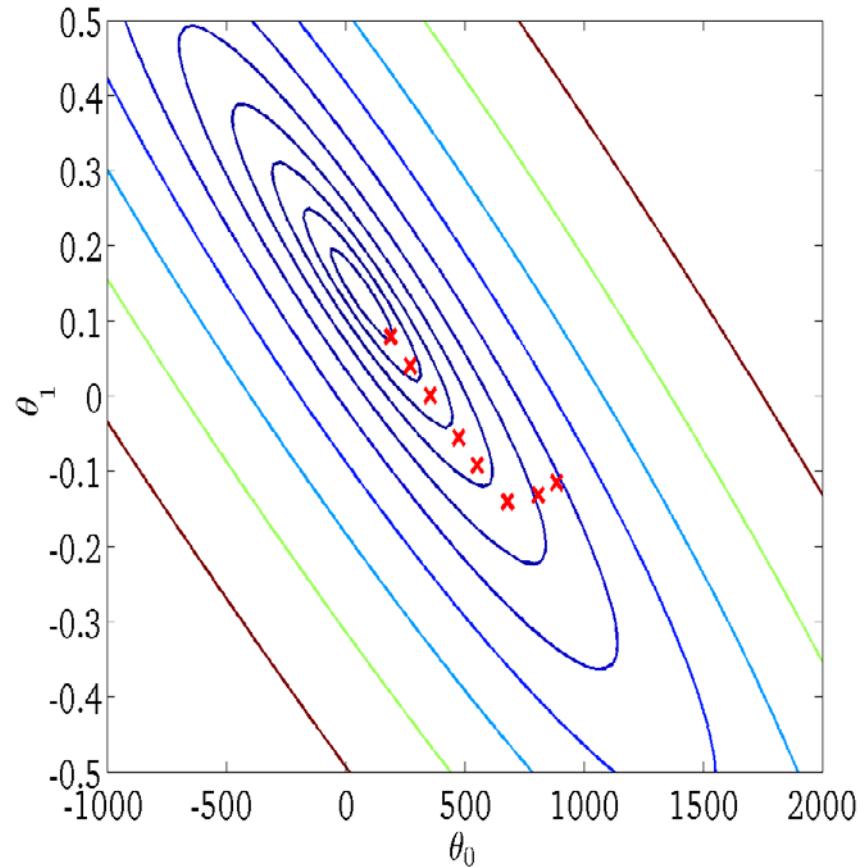
$$h_{\theta}(x)$$

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



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

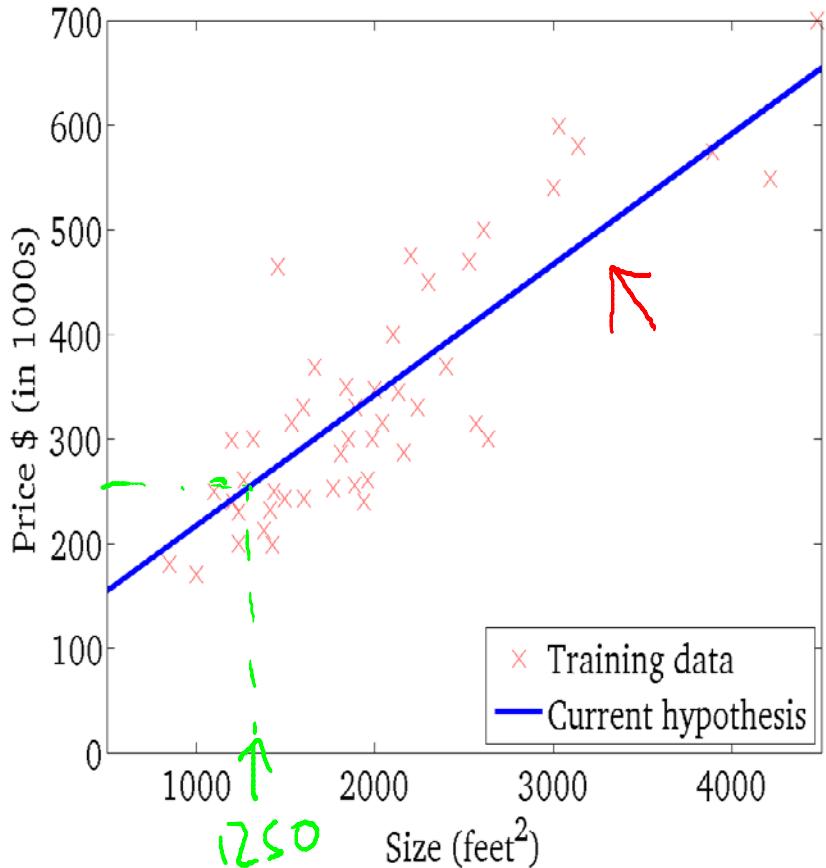
(function of the parameters θ_0, θ_1)



Source: Andrew Ng

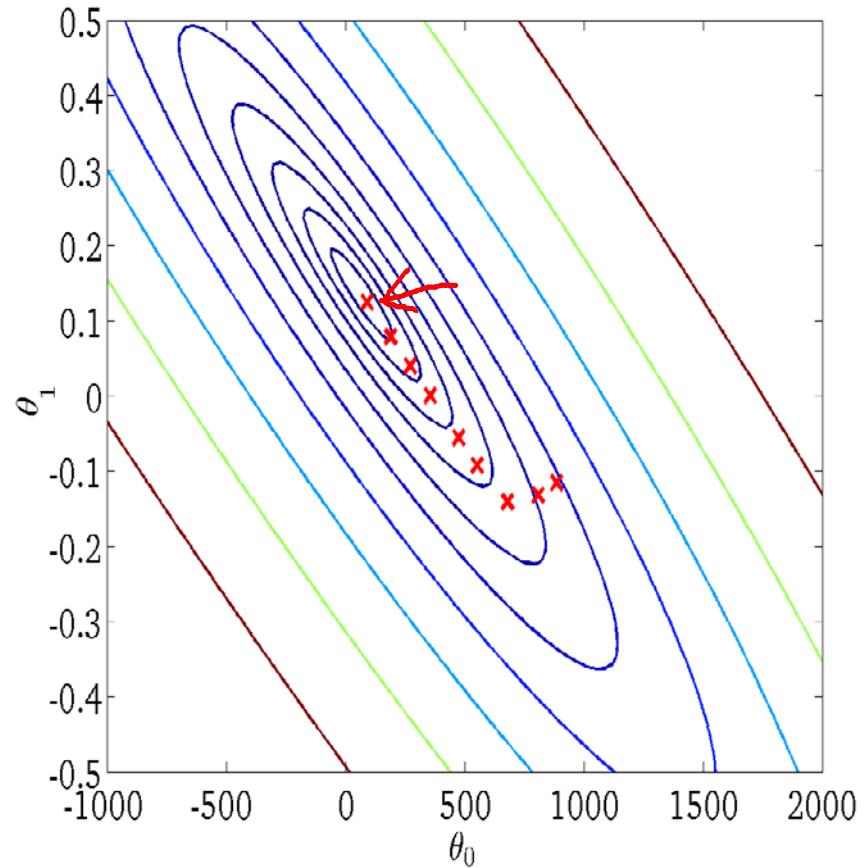
$$h_{\theta}(x)$$

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



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

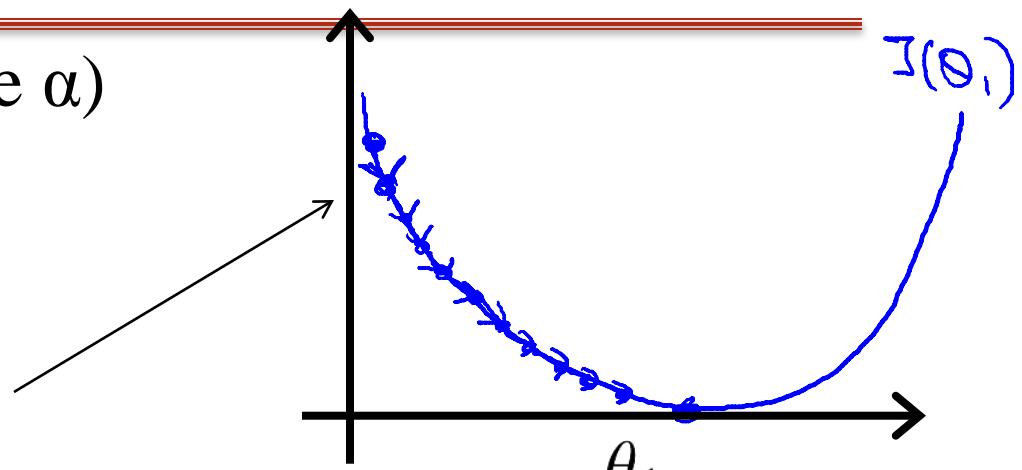
(function of the parameters θ_0, θ_1)



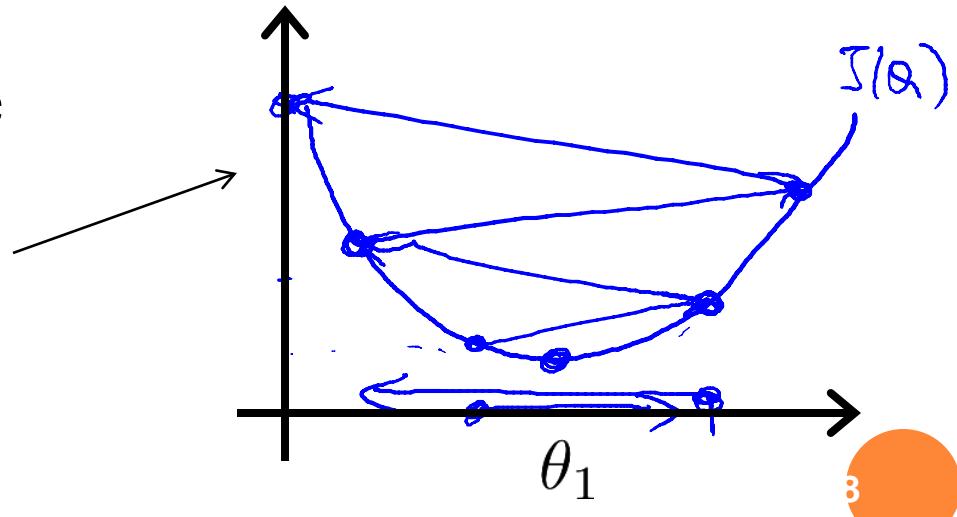
2.1. HÒI QUI TUYẾN TÍNH ĐƠN BIỀN

- Hệ số học (learning rate α)

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



- ✓ Quá nhỏ: Học chậm
- ✓ Quá lớn: Khó hội tụ về điểm tối ưu mà ở đó $J(\theta_0, \theta_1)$ là nhỏ nhất



2.1. HÒI QUI TUYẾN TÍNH ĐƠN BIẾN

- Tóm tắt: Cho N đối tượng đã được quan sát, m.hình HQTT đơn biến được cho dưới dạng sau (e_i giữ phần biến thiên của đáp ứng Y không được giải thích từ X)

✓ Dạng đường thẳng

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

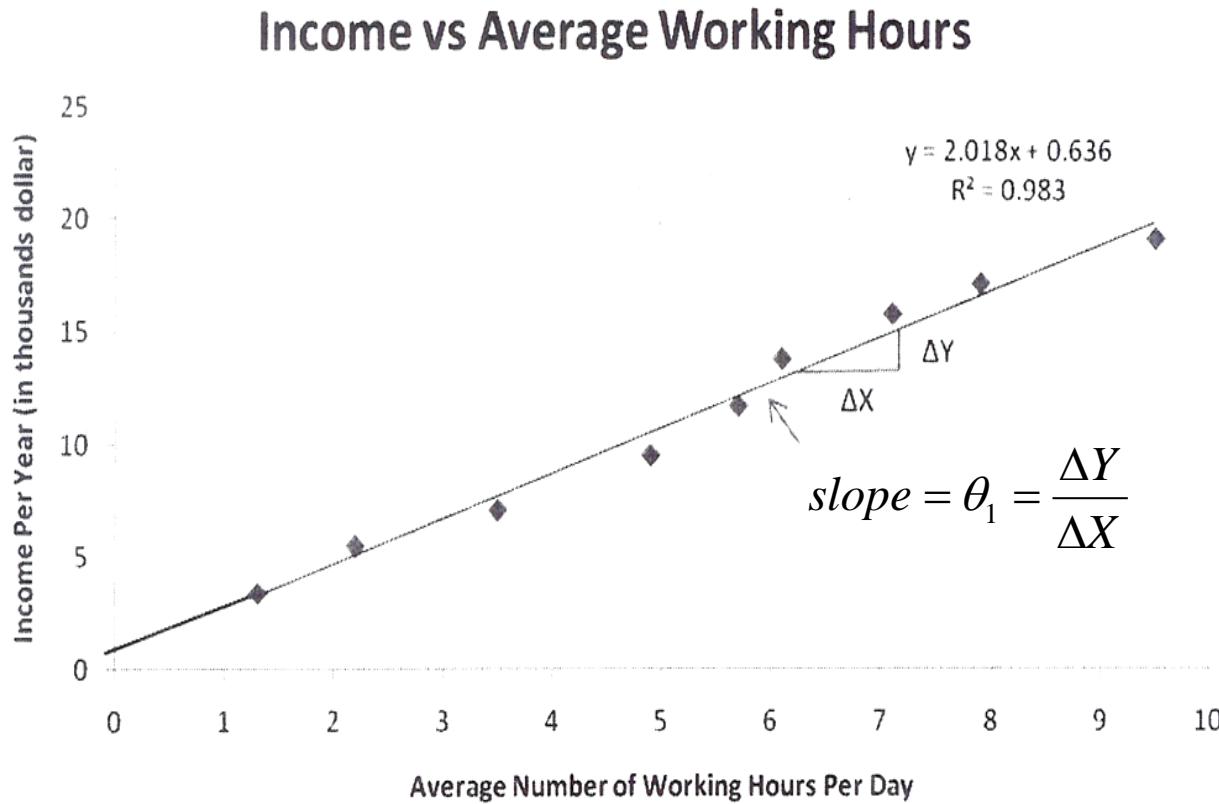
✓ Dạng parabol

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

- Ước lượng bộ thông số (θ_0, θ_1) bằng pp xuống đối hoặc có thể ước lượng nhanh bằng

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

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN



- $Y = \theta_0 + \theta_1 * X_1 \rightarrow Y = 0.636 + 2.018 * X$
- Dấu của θ_1 cho biết sự ảnh hưởng của X đối với Y.

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

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



$$y = \text{quantitySold} = 9323 - 823 * \text{price}$$

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

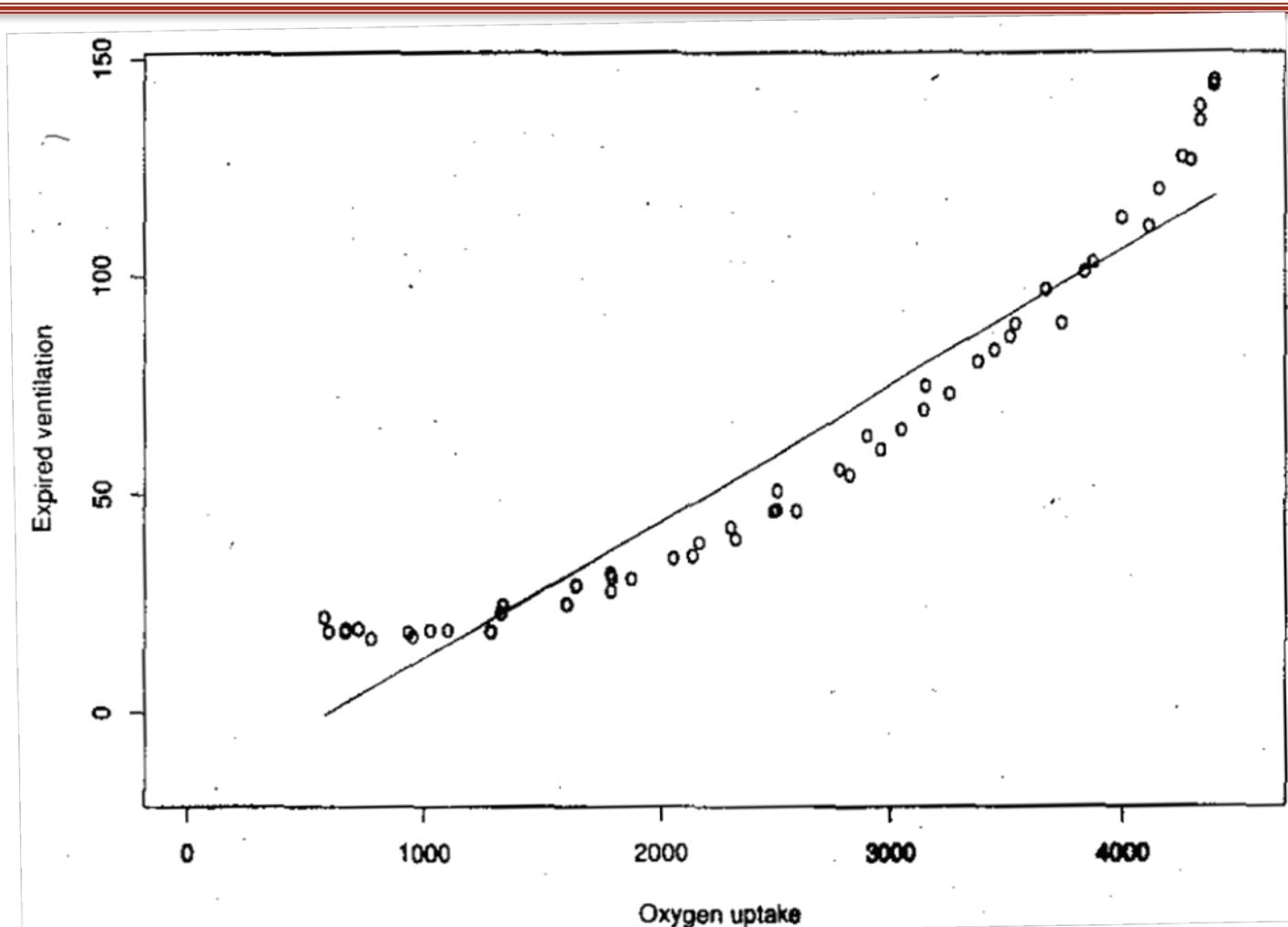


Figure 11.1, [2], pp. 372. Expired ventilation plotted against oxygen uptake in a series of trials, with fitted straight line: $y = \theta_0 + \theta_1 x$.

2.1. HỒI QUI TUYẾN TÍNH ĐƠN BIỀN

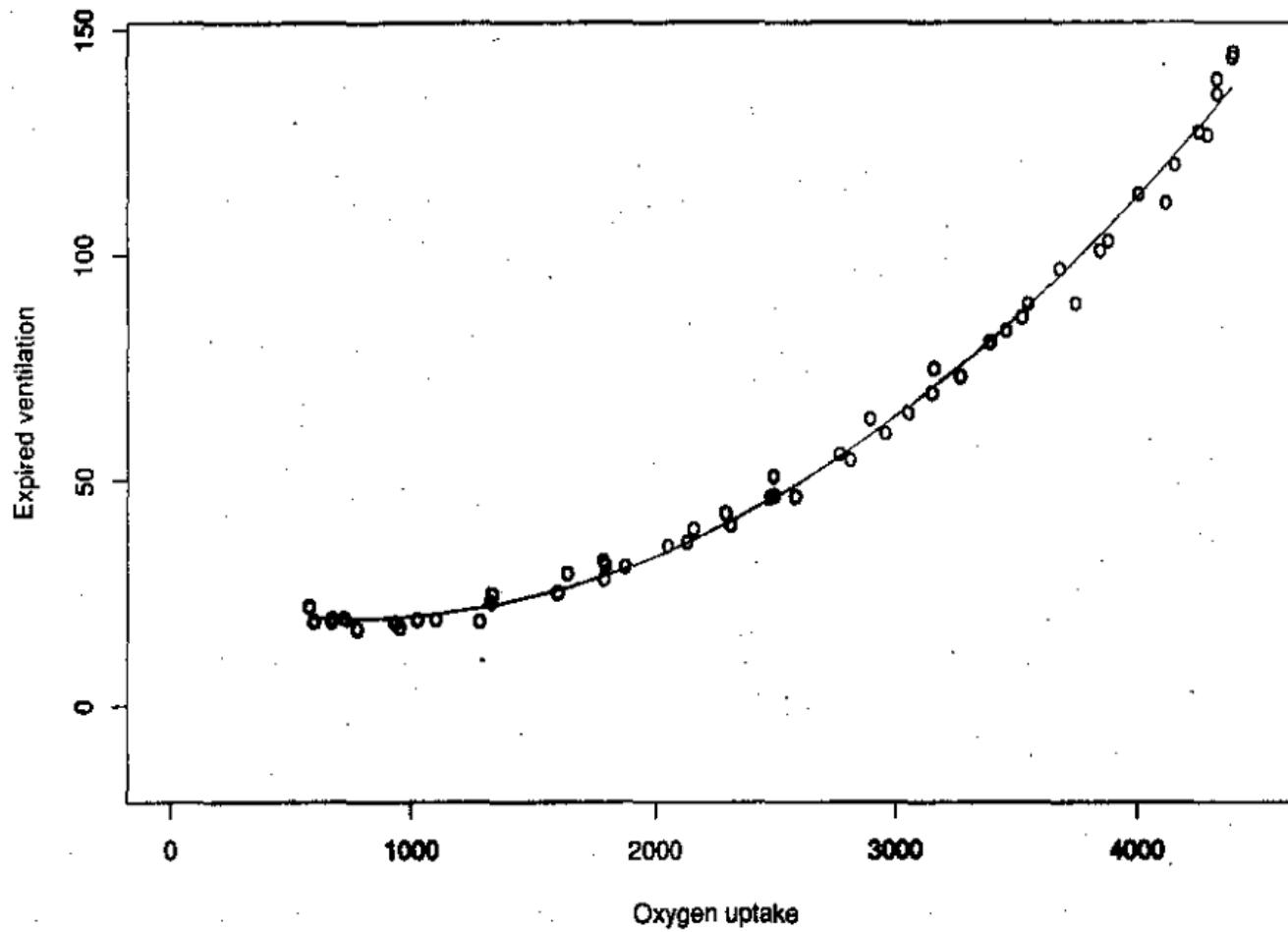


Figure 11.2, [2], pp. 373. The data from Figure 11.1 with a model that includes a term in x^2 : $y = \theta_0 + \theta_1x + \theta_2x^2$.

2.2. HỎI QUI TUYẾN TÍNH ĐA BIỂN

- Giá nhà bị tác động bởi nhiều yếu tố (biến)

Size (feet ²)	Số p. ngủ	Số tầng	Tuổi nhà	Giá(\$1K)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
...

n: số lượng thuộc tính (e.g., n=4)

$x^{(i)}$: input (features) của dữ liệu đào tạo thứ i^{th}

$x_j^{(i)}$: giá trị của thuộc tính j trong mẫu đào tạo thứ i^{th}

$y^{(i)}$: output thứ i^{th} trong mẫu đào tạo

E.x.,

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

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

- Giả thuyết về mô hình hồi qui

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

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

- Biểu diễn dưới dạng ma trận (đặt $x_0=1$)

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

45

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

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

- Giải thuật Gradient descent cho hồi qui đa biến

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

- Bộ thông số $\theta(\theta_0, \dots, \theta_n)$: Vector n+1 chiều
- Tối thiểu hóa $J(\theta) = J(\theta_0, \dots, \theta_n)$

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

Repeat until convergence {

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

}

2.2. HÒI QUI TUYẾN TÍNH ĐA BIỀN

Repeat until convergence{

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

}

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

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

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

....

2.2. HÒI QUI TUYẾN TÍNH ĐA BIỂN

- Vấn đề co giãn giá trị thuộc tính (feature scaling):
Đảm bảo các thuộc tính có cùng độ co giãn

- Ví dụ: x_1 =kích thước (0-2000 feet²)

$$x_2 = \text{số giường} (1-5)$$

\Rightarrow tốc độ hội tụ của giải thuật sẽ bị ảnh hưởng bởi độ co giãn này

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

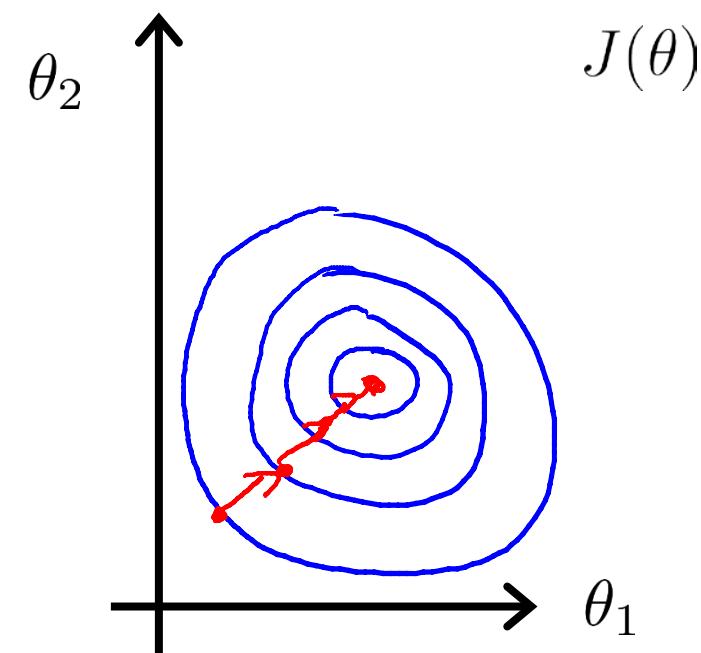
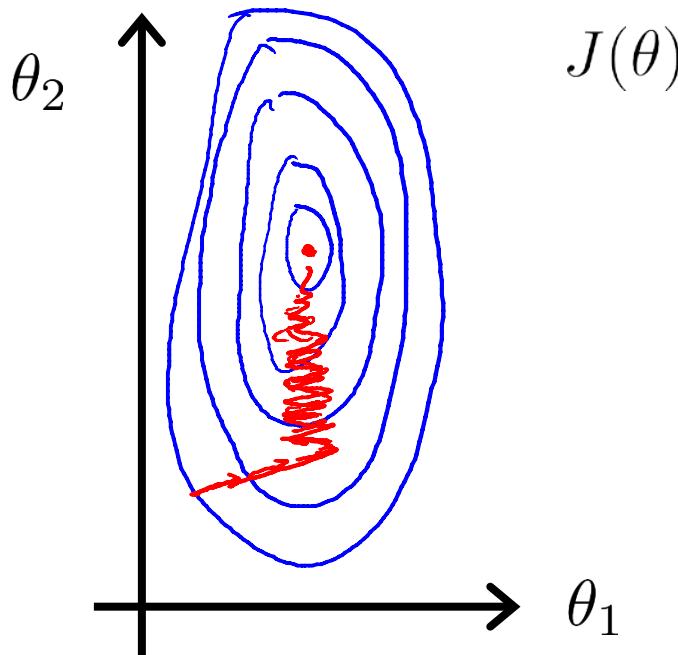
- Đảm bảo các thuộc tính có cùng độ co giãn

E.g. $x_1 = \text{size } (0-2000 \text{ feet}^2)$

$x_2 = \text{number of bedrooms } (1-5)$

Scaled: $x_1 = \text{size}/2000$

$x_2 = \text{number of bedrooms } /5$



Source: Andrew Ng

2.2. HÒI QUI TUYẾN TÍNH ĐA BIỂN

- Co giản giá trị thuộc tính (feature scaling)
 - Chuyển mọi thuộc tính về giá trị trong khoảng [-1,1]
 - Ex., $x_0=1$; $0 \leq x_1 \leq 3$; $-2 \leq x_2 \leq 0.5 \Rightarrow \text{OK}$
 $-100 \leq x_3 \leq 100$; $-0.0001 \leq x_4 \leq 0.0001 \Rightarrow \text{điều chỉnh}$

- Dùng các phương pháp chuẩn hóa giá trị dữ liệu học trong chương 2:

- Ex.

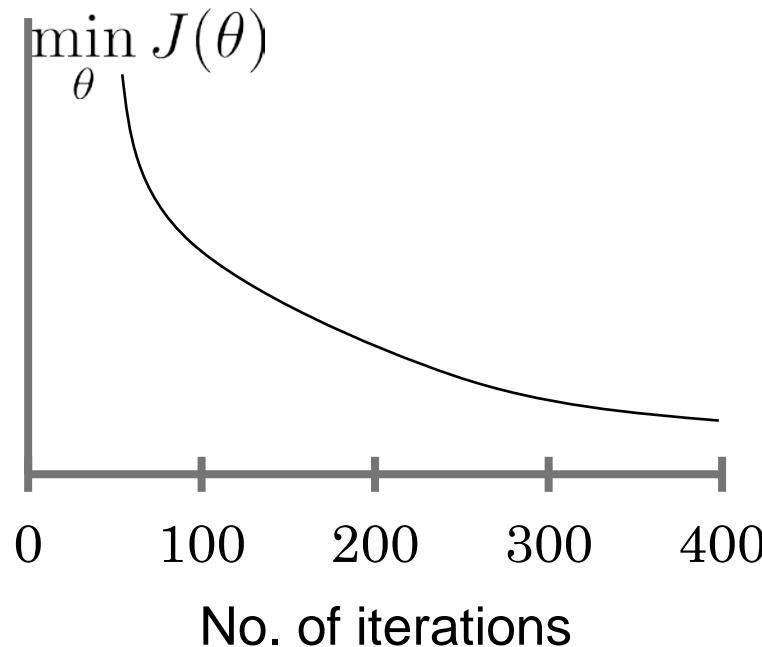
$$v' = \frac{v - \bar{v}}{\sigma v}$$

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

2.2. HÒI QUI TUYẾN TÍNH ĐA BIỀN

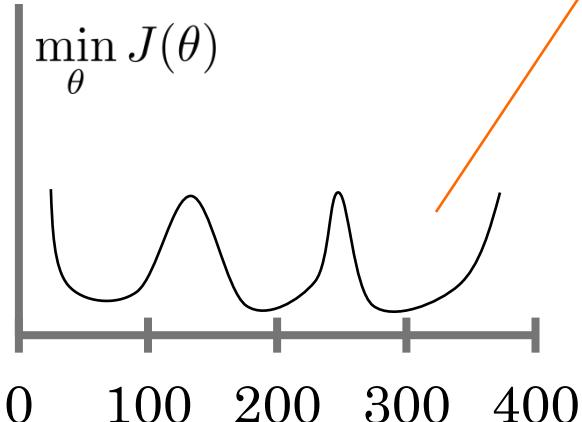
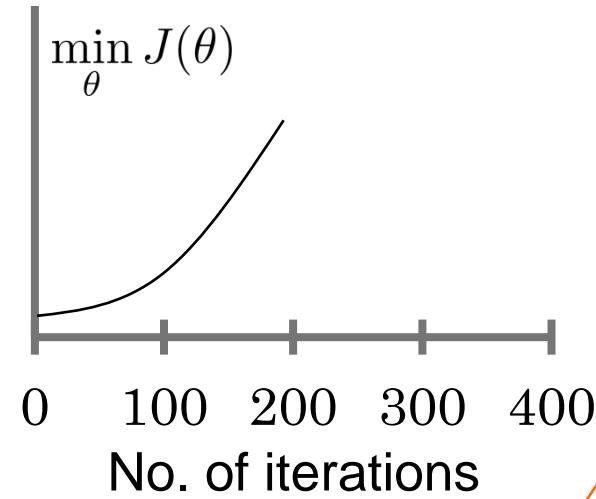
○ Kiểm tra giải thuật Gradient descent

- $J(\theta)$ phải giảm sau mỗi bước lặp
- Vẽ đồ thị biến thiên của $J(\theta)$ theo θ để kiểm tra khả năng hội tụ của giải thuật

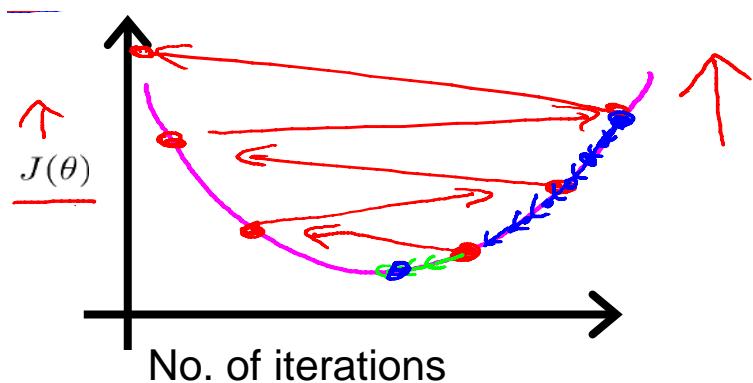


2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

Giải thuật Gradient descent không hội tụ



Dùng α nhỏ hơn



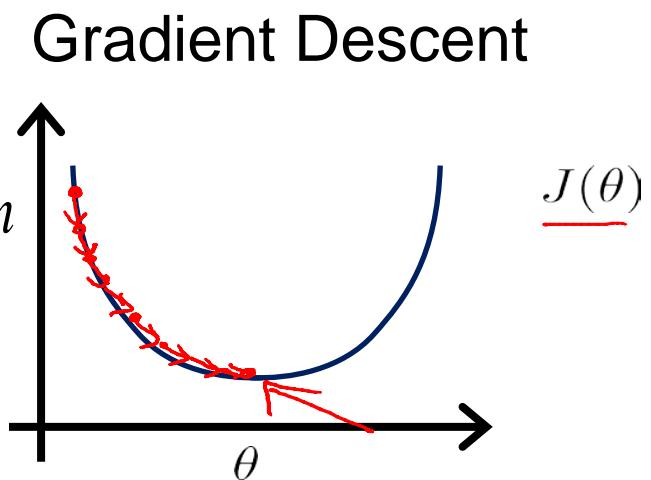
- α quá nhỏ: chậm hội tụ
- α quá lớn: $J(\theta)$ có thể không giảm ở mỗi bước lặp \Rightarrow gt có thể không hội tụ
- **thử α :** 0.001, 0.003, 0.01, 0.03, 0.1, 0.3, ...

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

- Giải các phương trình chuẩn (normal equation) để tìm bộ thông số θ

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

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



2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

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

x_0	k. Thước (feet ²) x_1	Số p. ngũ x_2	Số tầng x_3	Tuổi (years) x_4	Giá (\$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

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

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

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

- Cho tập training: N mẫu (example), n thuộc tính (features)

Gradient descent

- Phải chọn α
- Cần thử qua nhiều vòng lặp
- Làm việc tốt ngay cả khi n lớn (e.g., $n=10^6$)

Normal equation

- Không cần chọn α
- không cần vòng lặp
- Cần phải tính $(X^T X)^{-1}$
- Không hiệu quả với số chiều lớn ($n=10^4$ thì nên dùng gradient descent)

2.2. HÒI QUI TUYẾN TÍNH ĐA BIẾN

- Chú ý trường hợp ma trận không nghịch đảo được (non-invertible) trong phương pháp normal equation: $(X^T X)$ không thể nghịch đảo được
- Xử lý:
 - Kiểm tra các thuộc tính phụ thuộc (linearly dependent). Ex., Kích thước theo mét (x_1) và kích thước theo feet (x_2) => Loại biến phụ thuộc
 - Quá nhiều thuộc tính ($n > N$). Ex., $n=20$, $N=10$ => Xóa thuộc tính; tìm thuộc tính thay thế; thu thập thêm dữ liệu,...

2.2. HỒI QUI TUYẾN TÍNH ĐA BIỂN

- Một ví dụ khác:

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. HỎI QUI TUYẾN TÍNH ĐA BIỂN

SUMMARY OUTPUT

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

$$y = \text{Quantity Sold} = 8536.214 - 835.722 * \text{Price} + 0.592 * \text{Advertising}$$

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. HỒI QUI PHI TUYẾN

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

- \mathbf{Y} là hàm phi tuyến cho việc kết hợp các thông số $\boldsymbol{\theta}$.
- Ví dụ: hàm mũ, hàm logarit, hàm Gauss, ...

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

- Xác định bộ thông số $\boldsymbol{\theta}$ tối ưu: các giải thuật tối ưu hóa

- Tối ưu hóa cục bộ
- Tối ưu hóa toàn cục cho tổng thặng dư bình phương (sum of squared residuals/errors)

4. ỨNG DỤNG

- Quá trình khai phá dữ liệu
 - Tiền xử lý dữ liệu: Smoothing, loại nhiễu,...
 - KPDL: dự báo các giá trị số (numerical-values prediction), mô tả dữ liệu (predictive analysis)
- Ứng dụng trong nhiều lĩnh vực: sinh học (biology), nông nghiệp (agriculture), xã hội (social issues), kinh tế (economy), kinh doanh (business), ...

5. CÁC VẤN ĐỀ TRONG HỒI QUI

- Các giả định (assumptions)

- Phân bố dữ liệu
- Độc lập và giá trị liên tục của các biến độc lập/giải thích
- Thành phần lỗi: phân bố, trung bình, phương sai, độc lập

- Lượng dữ liệu được xử lý thường không lớn

- Đánh giá mô hình hồi qui

- Các kỹ thuật tiên tiến cho hồi qui:

- Artificial Neural Network (ANN)
- Support Vector Machine (SVM)

5. CÁC VẤN ĐỀ TRONG HỒI QUI

○ Đánh giá mô hình hồi qui:

- Thu thập dữ liệu mới để kiểm tra kết quả dự báo
- Dùng dữ liệu hiện có để đo độ chính xác của dự báo
- Phân tách dữ liệu (data splitting)
 - ✓ Training data
 - ✓ Test data → *đo độ chính xác của dự báo*
- Kiểm định chéo k phần (k-fold cross-validation)
 - ✓ Lặp k lần:
 - ✓ Training data: ($k-1$) phần
 - ✓ Test data: phần thứ k → *đo độ chính xác của dự báo*
 - ✓ Độ chính xác trung bình của các dự báo từ k lần thực thi

5. CÁC VẤN ĐỀ TRONG HỒI QUI

- Đánh giá mô hình hồi qui:

- Độ chính xác của dự báo
 - Tổng bình phương sai số (Sum of squared errors, SSE)
-> thể hiện cách đánh giá chung về sai số (overall measure of errors):
Càng nhỏ càng tốt

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

- Trung bình sai số bình phương (Mean squared error, MSE): Đánh giá sự biến thiên trong đó biến đáp ứng không giải thích được bởi hồi qui (measure of the variability in the response variable left unexplained by the regression): **Càng nhỏ càng tốt**

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

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

5. CÁC VẤN ĐỀ TRONG HỒI QUI

○ Độ chính xác của dự báo (tt)

- ✓ Sai số chuẩn (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
- ✓ Thể hiện tính chính xác (precision) của dự báo tạo ra bởi mô hình hồi qui

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

5. CÁC VẤN ĐỀ TRONG HỒI QUI

- Các yếu tố ảnh hưởng đến sự thành công trong việc xây dựng các mô hình hồi qui
 - ✓ Mô hình đúng vấn đề (proper problem formulation)
 - ✓ Chọn các biến quan trọng và dạng mô hình (selection of important variables and model form)
 - ✓ Tập hợp dữ liệu tốt (cả về số lượng lẫn chất lượng)
 - ✓ Sử dụng các thủ tục hợp lý cho việc ước lượng thông số (the use of good coefficient estimation procedures)
 - ✓ Các kỹ thuật đánh giá mô hình (model validation techniques)

6. TÓM TẮT

○ Hồi qui

- Kỹ thuật thống kê, được áp dụng cho các thuộc tính liên tục (continuous attributes/features)
- Có lịch sử phát triển lâu đời
- Đơn giản nhưng rất hữu dụng, được ứng dụng rộng rãi
- Cho thấy sự đóng góp đáng kể của lĩnh vực thống kê trong lĩnh vực khai phá dữ liệu

○ Các dạng mô hình hồi qui: tuyến tính/phi tuyến, đơn biến/đa biến, có thông số/phi thông số/thông số kết hợp, đối xứng/bất đối xứng

Q&A

quangtran@hcmut.edu.vn