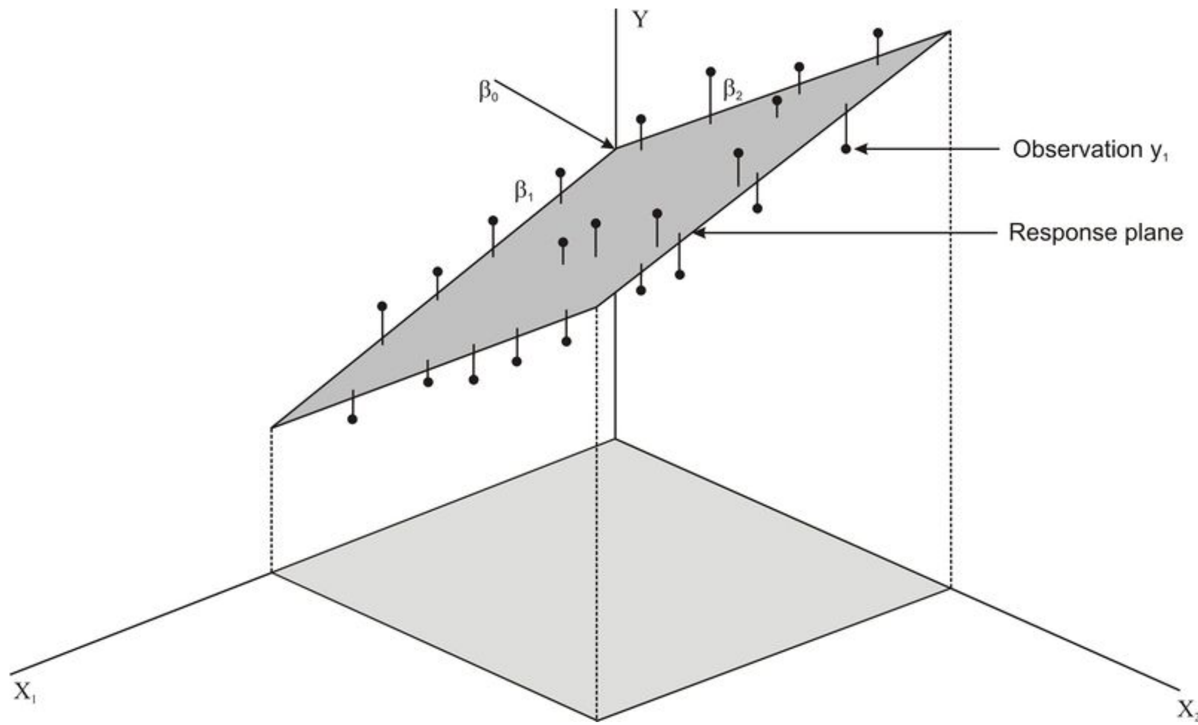
Linear Regression with Multiple Predictor Variables



ck12.org

Slides from:
Andrew Ng

CSC411/2515: Machine Learning and Data Mining, Winter 2018
Michael Guerzhoy and Lisa Zhang

Multiple variables (features) predict y

Size (feet²) x_1	Number of bedrooms x_2	Number of floors x_3	Age of home (years) x_4	Price ($1000) y
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…	…	…	…	…

$x_0 = 1$

Notation:

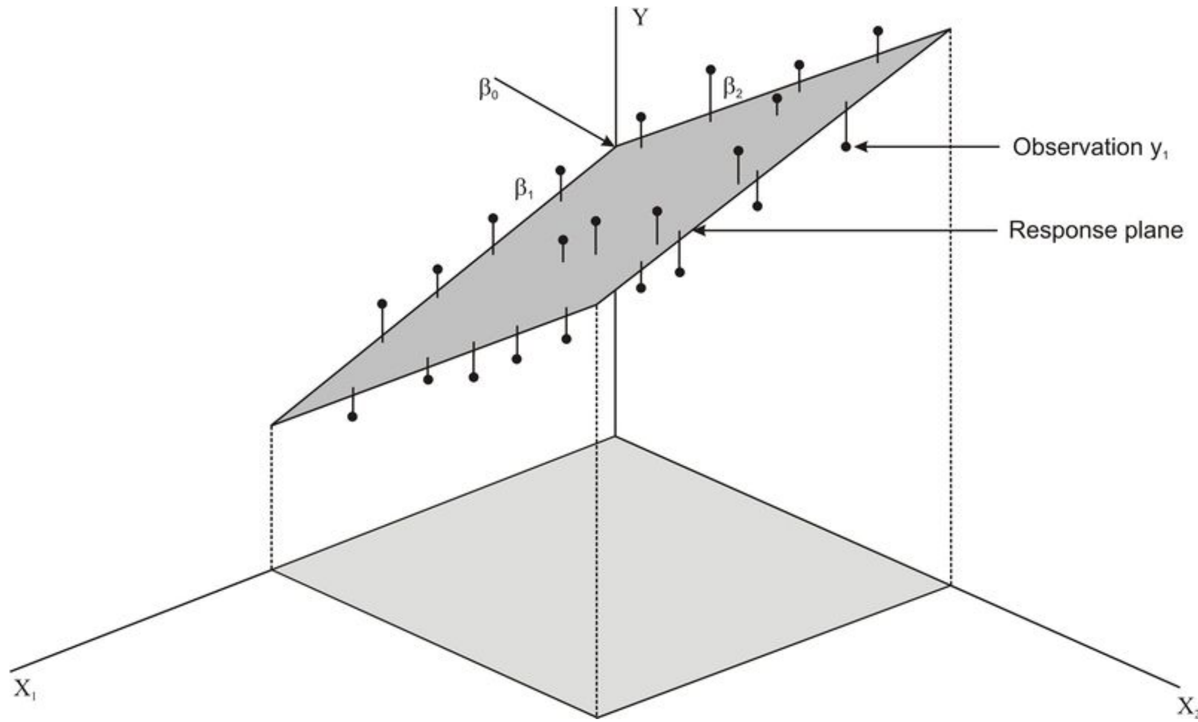
n = number of variables features

$x^{(i)}$ = input (features) of i^{th} training example.

$x_j^{(i)}$ = value of feature j in i^{th} training example.

$x_0\theta_0 = \theta_0$

$$h_\theta(\boldsymbol{x}) = h_{\boldsymbol{\theta}}(x_0, x_1, x_2, \dots, x_n) = x_0\theta_0 + x_1\theta_1 + \cdots + x_n\theta_n = \boldsymbol{\theta}^T\boldsymbol{x}$$



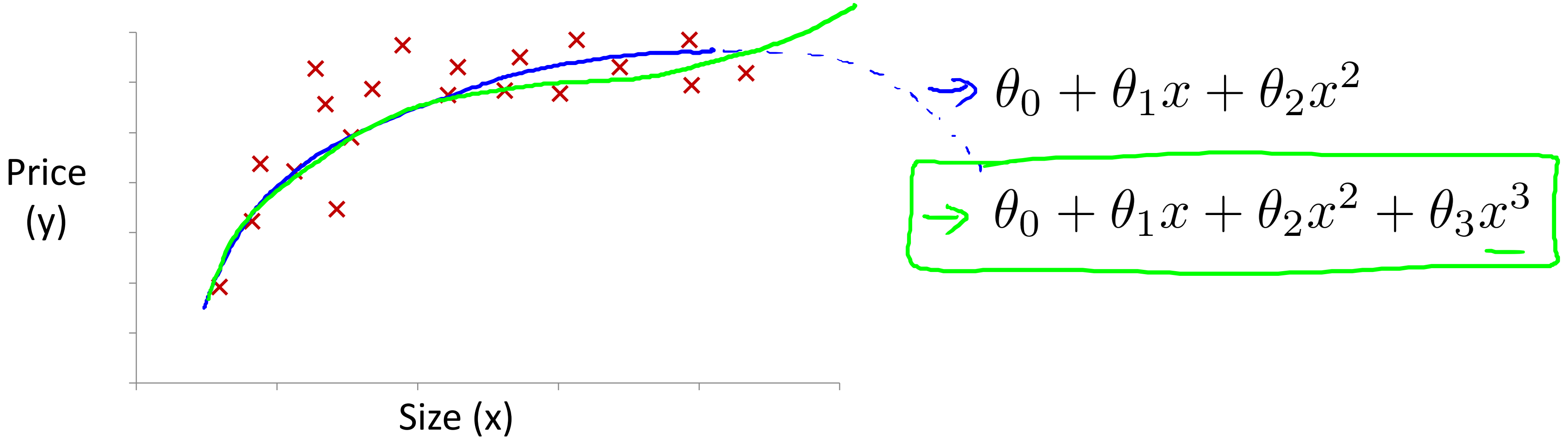
Minimize: $$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Minimizing this cost function corresponds to minimizing the distance between The observations and the hyperplane defined by $\boldsymbol{\theta}^{\boldsymbol{T}}\boldsymbol{X} - \boldsymbol{Y}$=0

Reminder about the intuition for this in 2D on the board

Computed Features

It is sometimes useful to compute more features



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$$
$$= \theta_0 + \theta_1 (size) + \theta_2 (size)^2 + \theta_3 (size)^3$$

$$x_1 = (size)$$
$$x_2 = (size)^2$$
$$x_3 = (size)^3$$

Much better fit than we would have gotten with linear regression

Computed Features Examples – cont'd

Basic Idea:

$$b_{\theta}(depth, frontage) = \theta_0 + \theta_1 depth + \theta_1 frontage$$

Better (if the price depends on the area):

$$h_{\theta}(depth, frontage) = \theta_0 + \theta_1 depth + \theta_2 frontage + \theta_3 (frontage \times depth)$$

Note: we could not represent the idea that the price is proportional to the area using The basic b_{θ}

Set:

$$x_0: 1$$
$$x_1: depth$$
$$x_2: frontage$$
$$x_3: depth \times frontage$$
$$x_4: depth^2$$
...

(But there is a limit to how much we can do this without *overfitting:* more on that later), and find the best θ such that

$$h_\theta(\boldsymbol{x}) = \theta^T x$$