DS 5220

Supervised Machine Learning and Learning Theory

Alina Oprea
Associate Professor, CCIS
Northeastern University

September 11 2019

Outline

- Finish probability review
- Linear regression
  – MSE objective
- MLE for linear regression
  – Statistical interpretation
- Simple linear regression
  – Optimal closed-from solution
- Multiple linear regression
  – Optimal closed-form solution

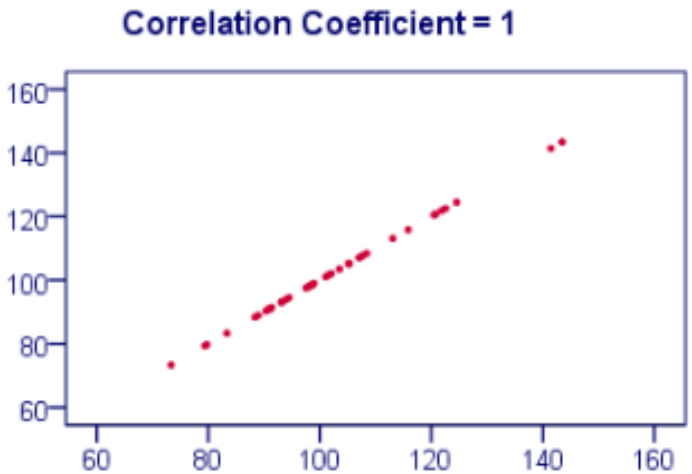
Covariance

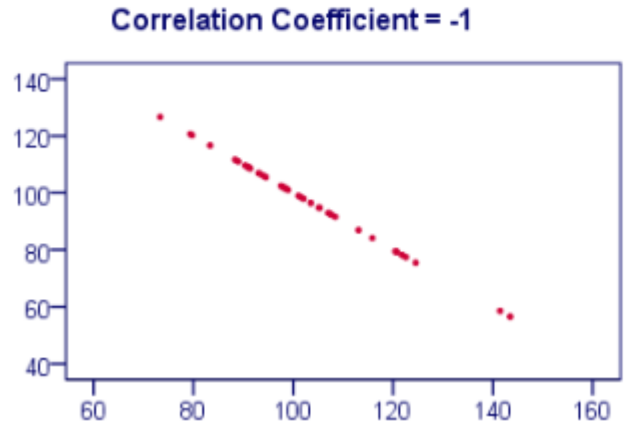
- X and Y are random variables
- $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$
- Properties

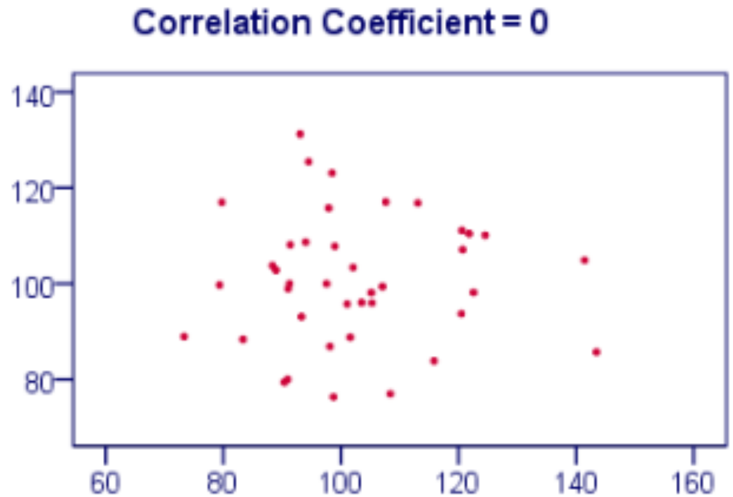
$$
\begin{aligned}
&\text{(i)} \quad \mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) \\
&\text{(ii)} \quad \mathrm{Cov}(X, X) = \mathrm{Var(X)} \\
&\text{(iii)} \quad \mathrm{Cov}(aX, Y) = a\,\mathrm{Cov}(X, Y) \\
&\text{(iv)} \quad \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{Cov}(X_i, Y_j)
\end{aligned}
$$

Pearson Correlation

$$\rho = \mathrm{Corr}(X,Y) = \frac{\mathrm{Cov(X,Y)}}{\sigma_X \sigma_Y} \in [-1,1]$$







Bivariate distributions

$$F_{X,Y}(x,y) = \mathrm{P}(X \leq x, Y \leq y) \quad \textbf{(Eq.1)}$$

Joint CDF

$$F_X(x,y) = P(X \leq \ x, Y \leq y\ ) = P[X \leq \ x|Y = y|]P[y \leq y\ ]$$

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \quad \textbf{(Eq.5)}$$

Joint PDF

$f_X(x) = \int_y f_{X,Y}(x,y)dy$ Marginal distributions

$f_Y(x) = \int_x f_{X,Y}(x,y)dx$

Bivariate normal

- $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$ are Normal
- $\mu = (E[X], E[Y]) = \binom{\mu_X}{\mu_Y}$

mean vector

- $\Sigma = \begin{pmatrix} Var(X) & Cov(X,Y) \\ Cov(X,Y) & Var(Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\, \sigma_X \sigma_Y \\ \rho\, \sigma_X \sigma_Y & \sigma_Y^2 \end{pmatrix}$

covariance matrix

Bivariate normal

$$\boxed{F_{X,Y}(x,y) = \mathrm{P}(X \leq x, Y \leq y) \quad \textbf{(\textit{Eq.1})}}$$

Joint CDF

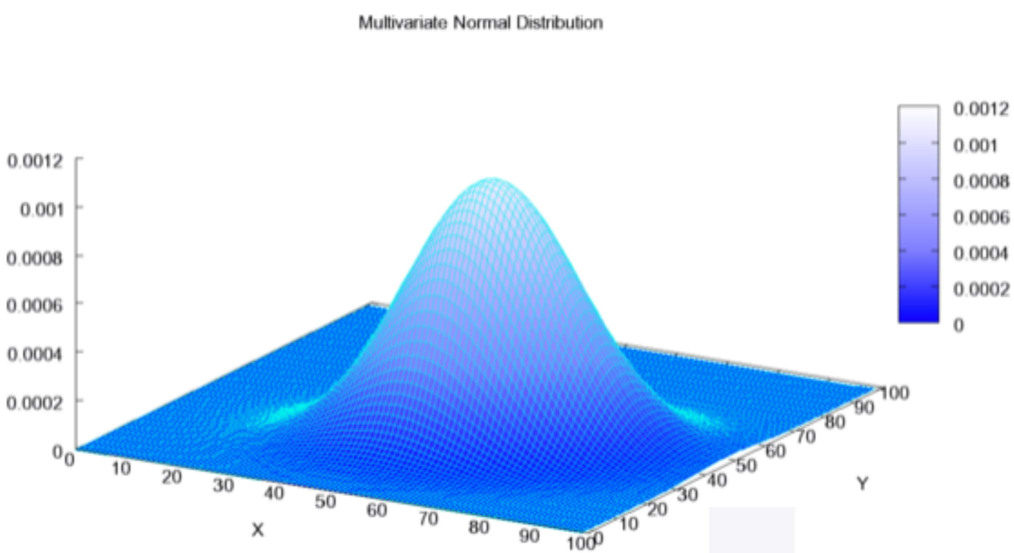
$$\boxed{f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \quad \textbf{(\textit{Eq.5})}}$$

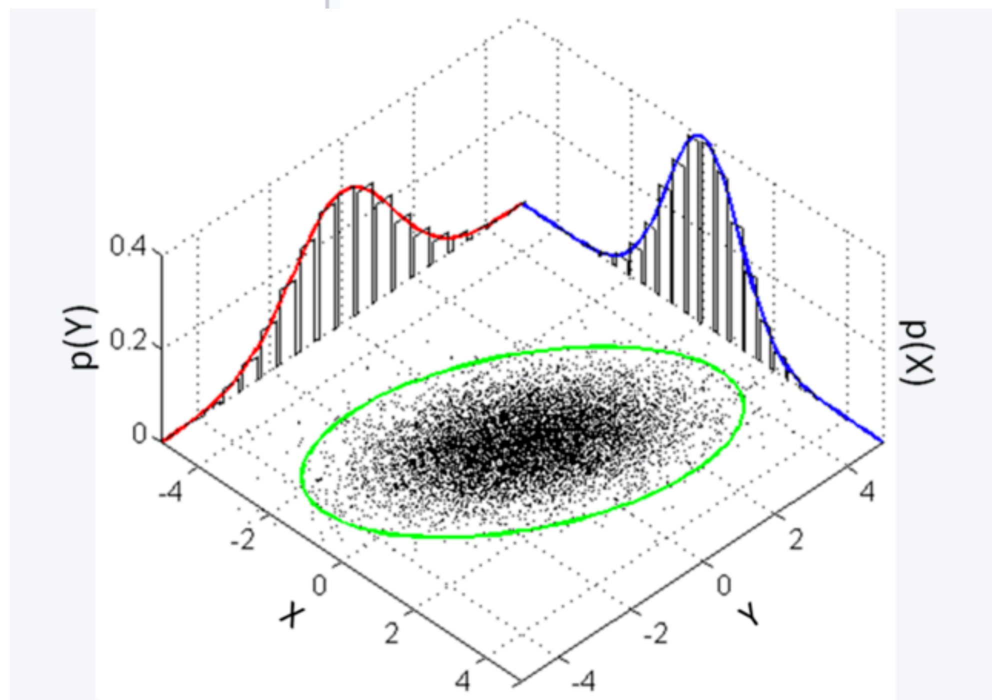
Joint PDF

$$f(z) = \frac{\exp(-\frac{1}{2}(z-\mu)^T \Sigma^{-1}(z-\mu))}{2\pi\sqrt{|\Sigma|}},$$

$$z = (x, y)$$

Bivariate normal





- Marginals of bivariate normal are normal
- Linear combinations of normal are normal

Bivariate normal

If X and Y have mean μ_X and μ_Y, general case is:

$$f_{X,Y}(x,y) =$$

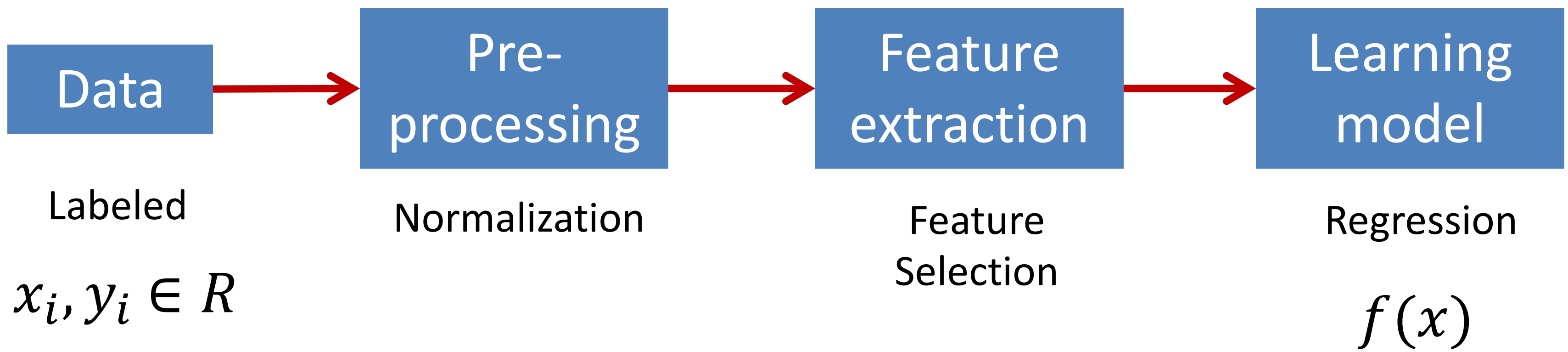
$$= \frac{1}{2\pi\sigma_X\sigma_Y(1-\rho^2)^{1/2}} \exp\left[\frac{-1}{2(1-\rho^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - 2\rho\frac{(x-\mu_X)}{\sigma_X}\frac{(y-\mu_Y)}{\sigma_Y}\right)\right]$$

If X and Y are uncorrelated ($\rho = 0$), and centered with mean 0:

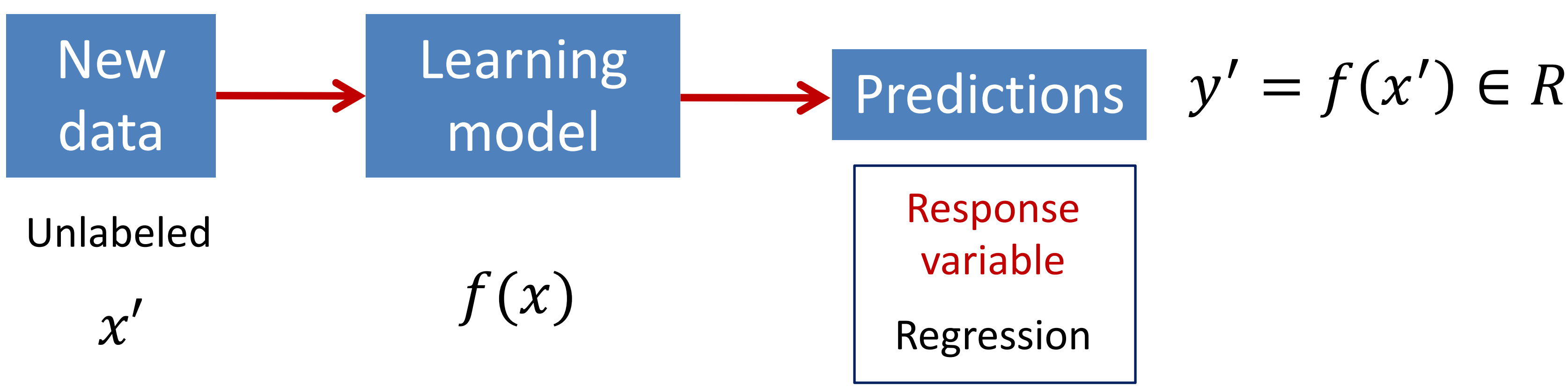
$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y} e^{-\frac{x^2}{2\sigma_X^2} - \frac{y^2}{2\sigma_Y^2}},$$

Supervised Learning: Regression

Training



Testing



Steps to Learning Process

- Define problem space
- Collect data
- Extract feature
- Pick a model (hypothesis)
- Develop a learning algorithm
  - Train and learn model parameters
- Make predictions on new data
  - Testing phase
- In practice, usually re-train when new data is available and use feedback from deployment

Linear regression

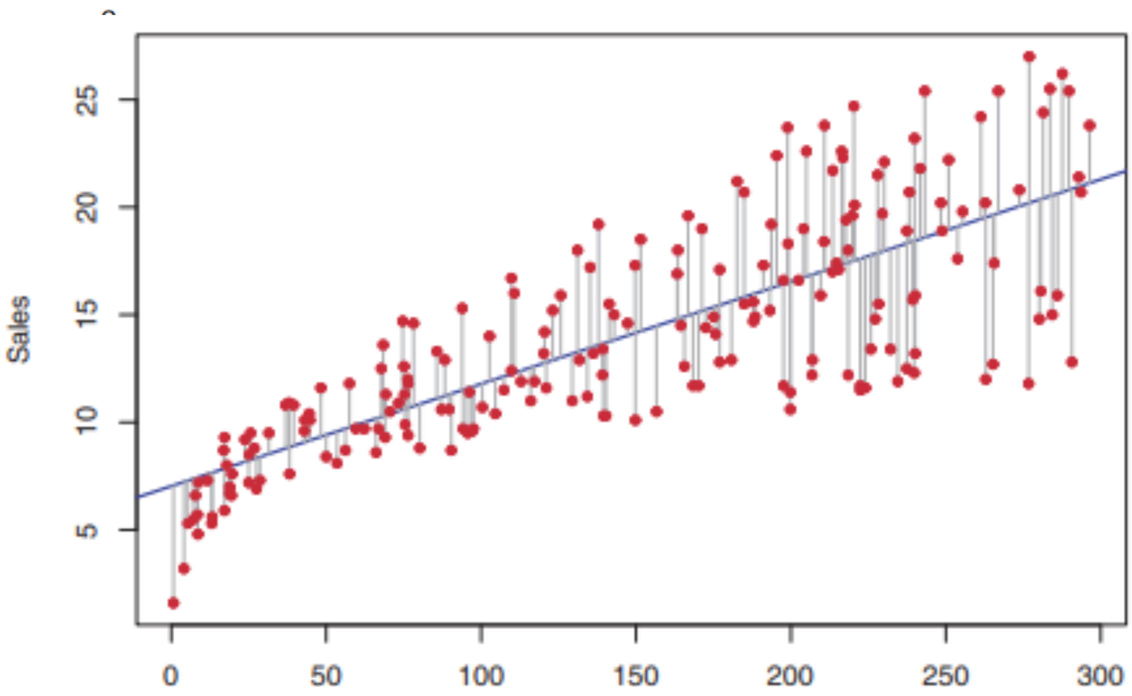
- One of the most widely used techniques
- Fundamental to many complex models
  - Generalized Linear Models
  - Logistic regression
  - Neural networks
  - Deep learning
- Easy to understand and interpret
- Efficient to solve in closed form
- Efficient practical algorithm (gradient descent)

Linear regression

Given:

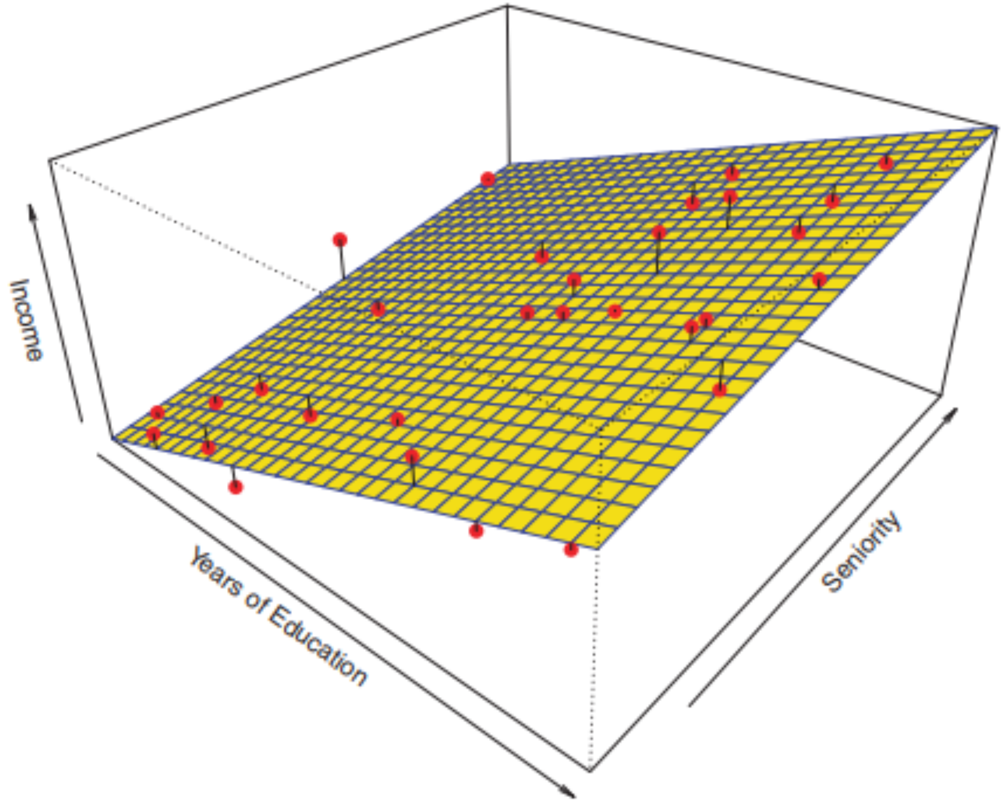
– Data $X = \{x_1, \ldots x_N\}$, where $x_i \in R^d$ Features

– Corresponding labels $Y = \{y_1, \ldots y_N\}$, where $y_i \in R$ Response variables



Simple Linear Regression: 1 predictor

Income Prediction



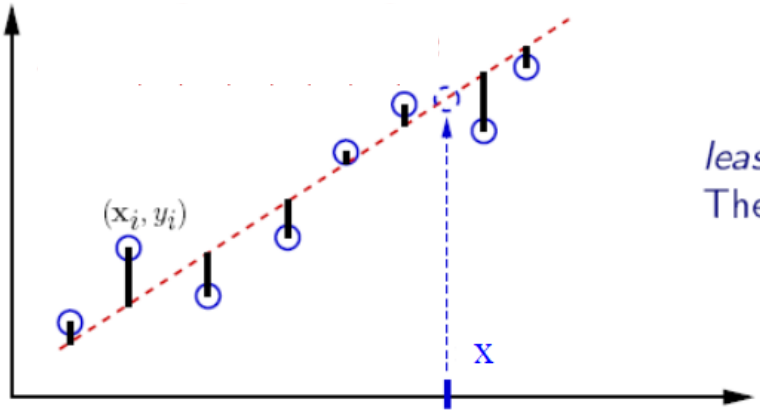
Linear Regression with 2 predictors
Multiple Linear Regression

Hypothesis: linear model

- Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Simple linear regression
Regression model is a line with 2 parameters: θ_0, θ_1

- Fit model by minimizing sum of squared errors



least squares (LSQ)
The fitted line is used as a predictor

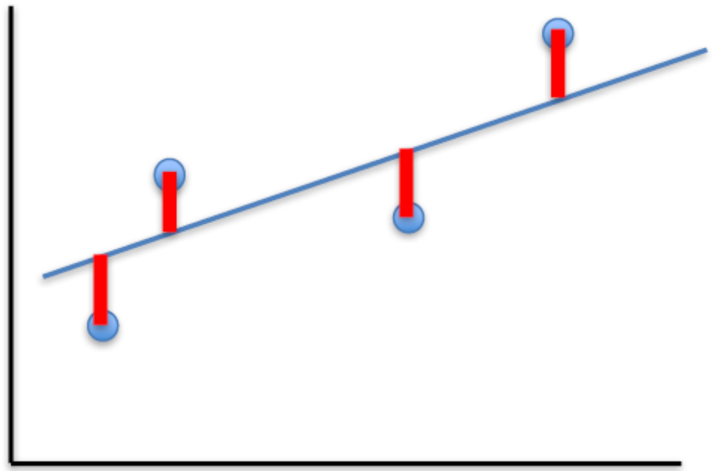
Least-Squares Linear Regression

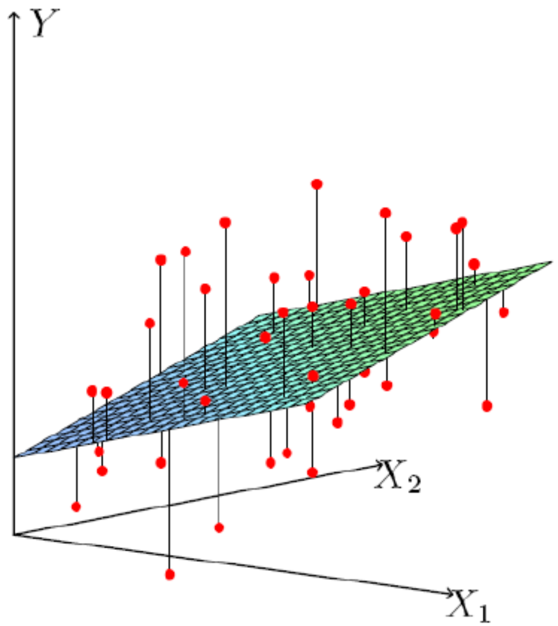
- Cost Function

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

Mean Square Error (MSE)

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

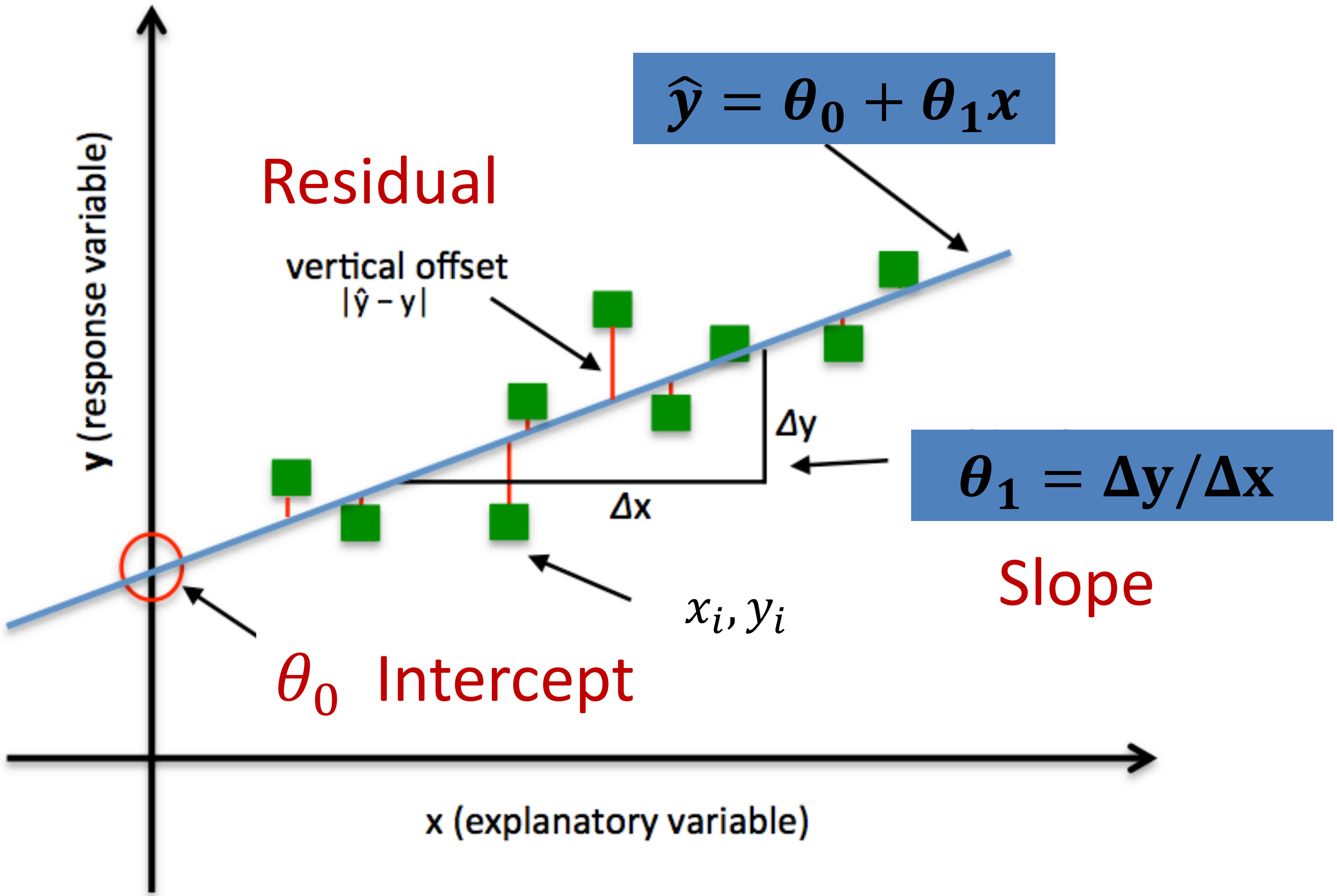




Terminology and Metrics

- Residuals
  - Difference between predicted values and actual values
  - Predicted value for example i is: $\widehat{y_i} = h_\theta(x_i)$
  - $R_i = \left|y_i - \widehat{y_i}\right| = |y_i - (\theta_0 + \theta_1 x_i)|$
- Residual Sum of Squares (RSS)
  - $RSS = \sum {R_i}^2 = \sum \left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$
- Mean Square Error (MSE)
  - $MSE = \frac{1}{N} \sum R_i^2 = \frac{1}{\mathrm{N}} \sum \left[y_i - (\theta_0 + \theta_1 x_i)\right]^2$

Interpretation



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

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} [h_\theta(x_i) - y_i]^2$$

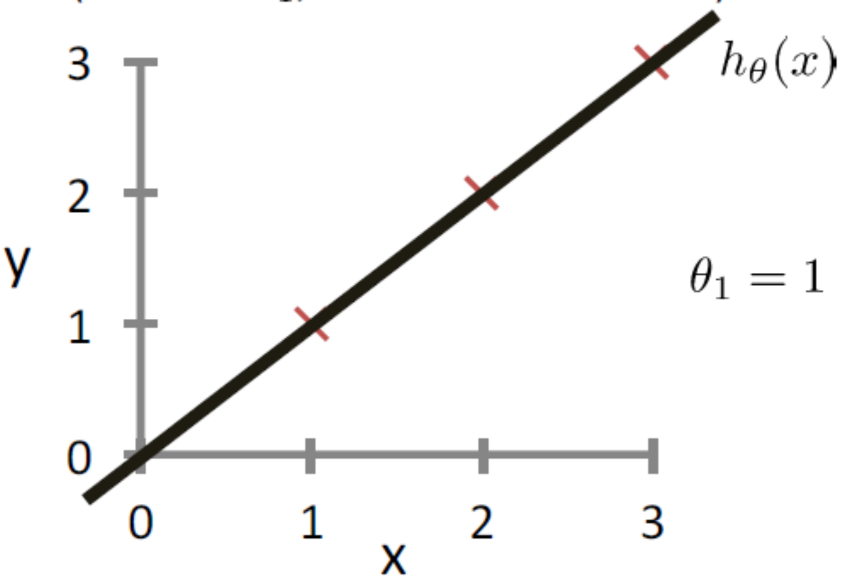
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

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

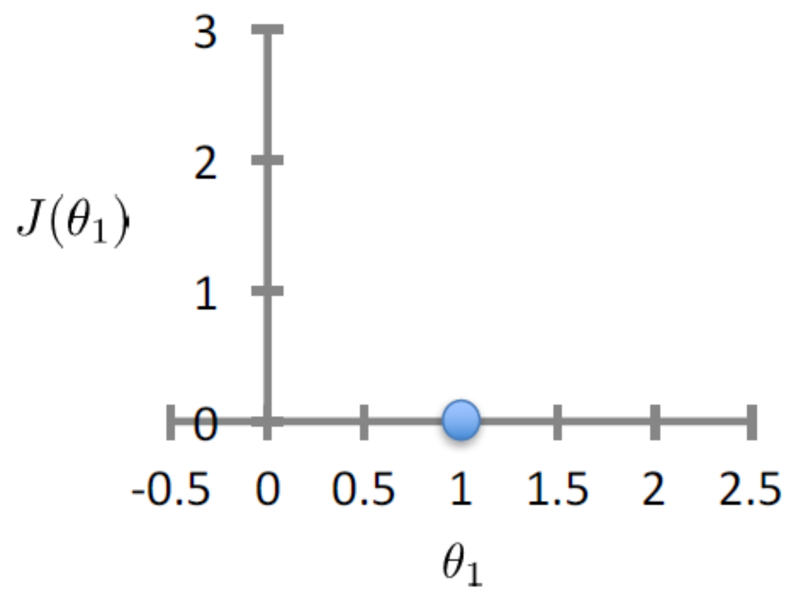
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Fix θ_0 =0

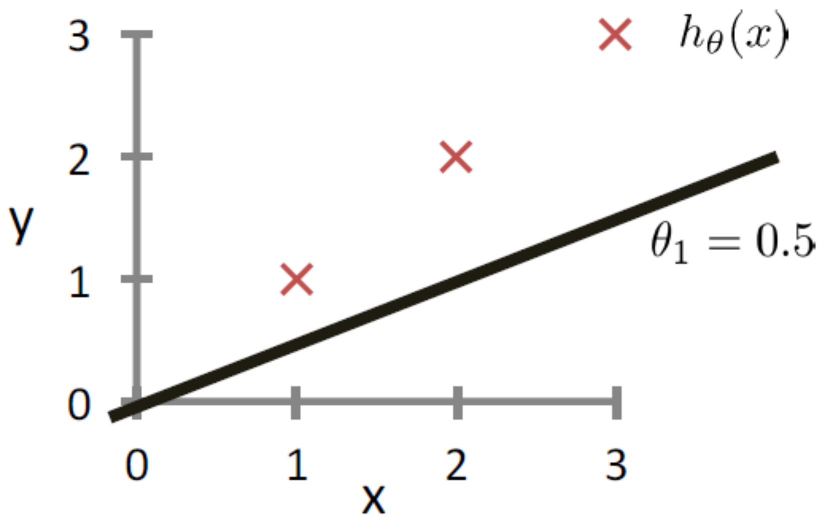
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

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

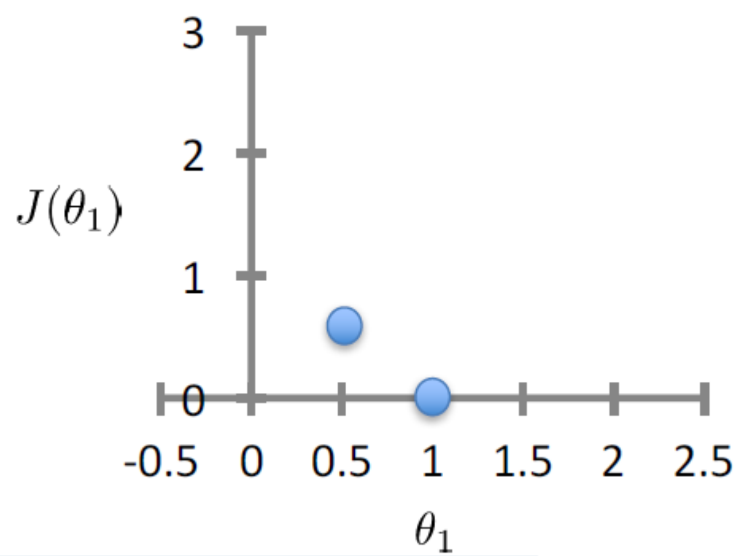
$h_\theta(x)$

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



$J(\theta_1)$

(function of the parameter θ_1)



Based on example by Andrew Ng

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

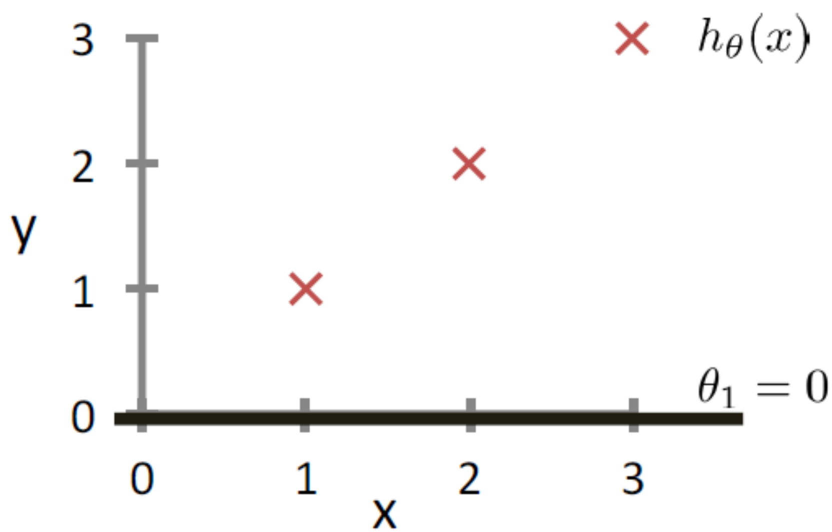
Intuition on MSE

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}[h_\theta(x_i) - y_i]^2$$

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

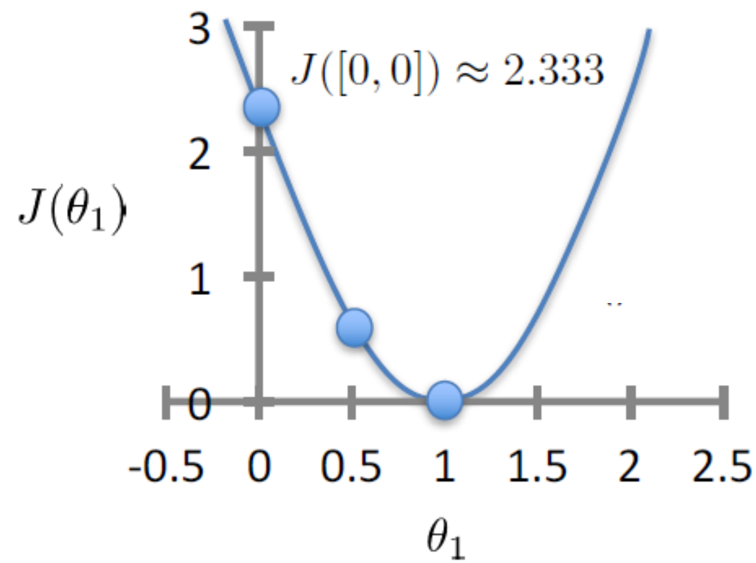
$h_\theta(x)$

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



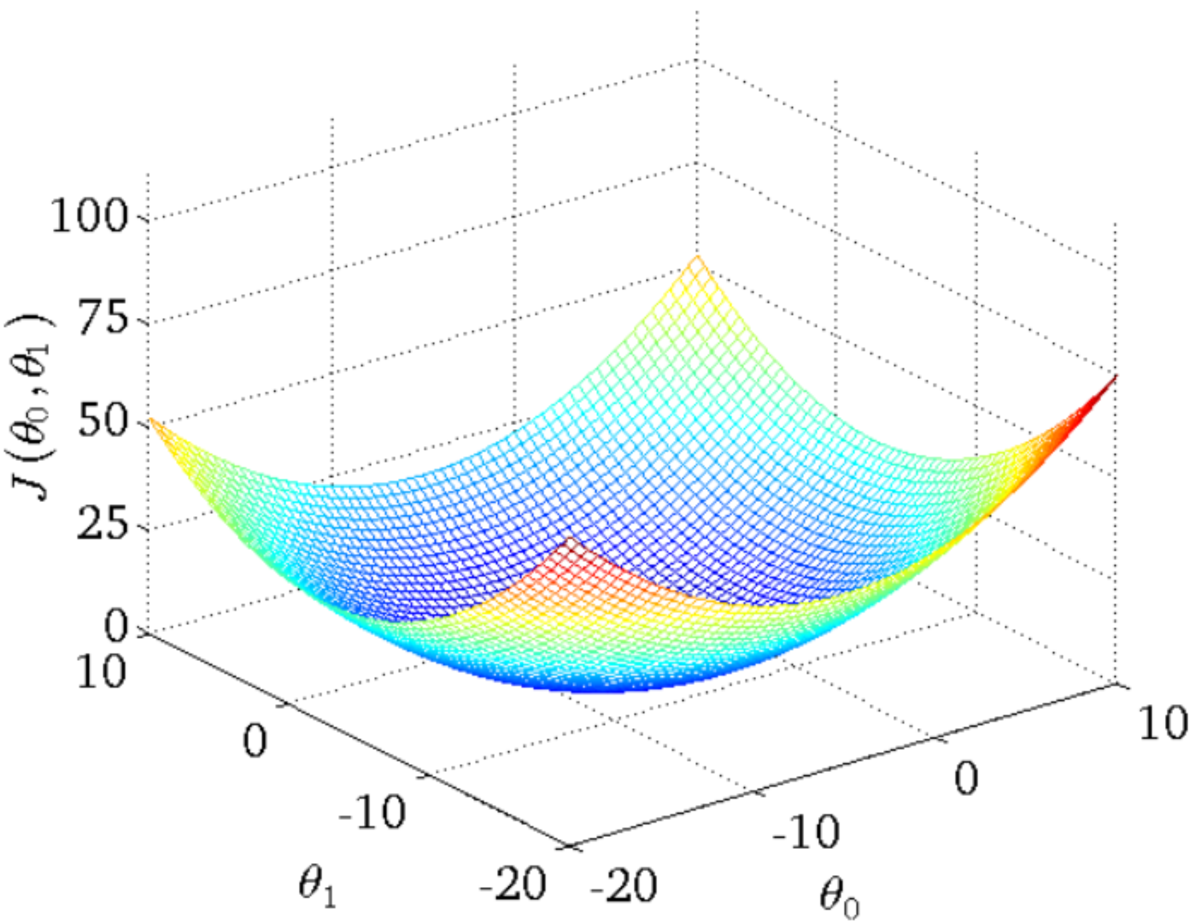
$J(\theta_1)$

(function of the parameter θ_1)



Based on example
by Andrew Ng

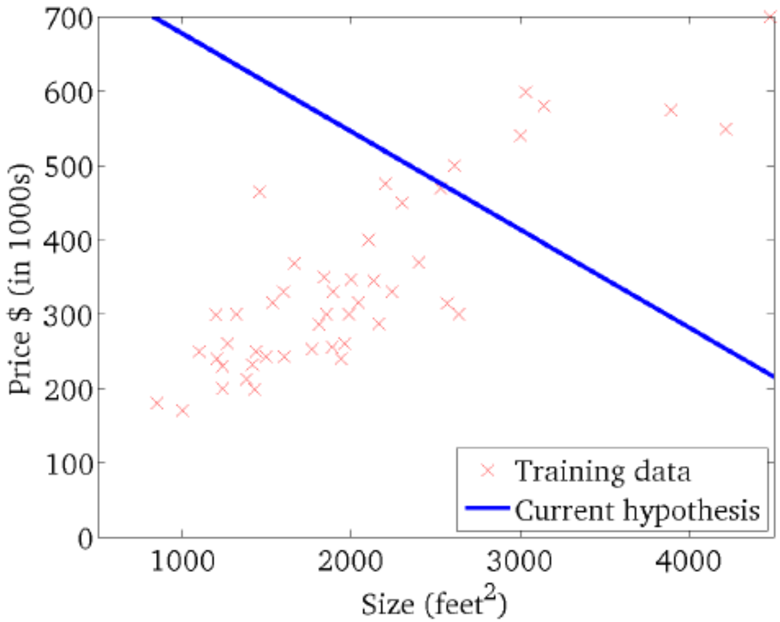
MSE function



Relation between h and J

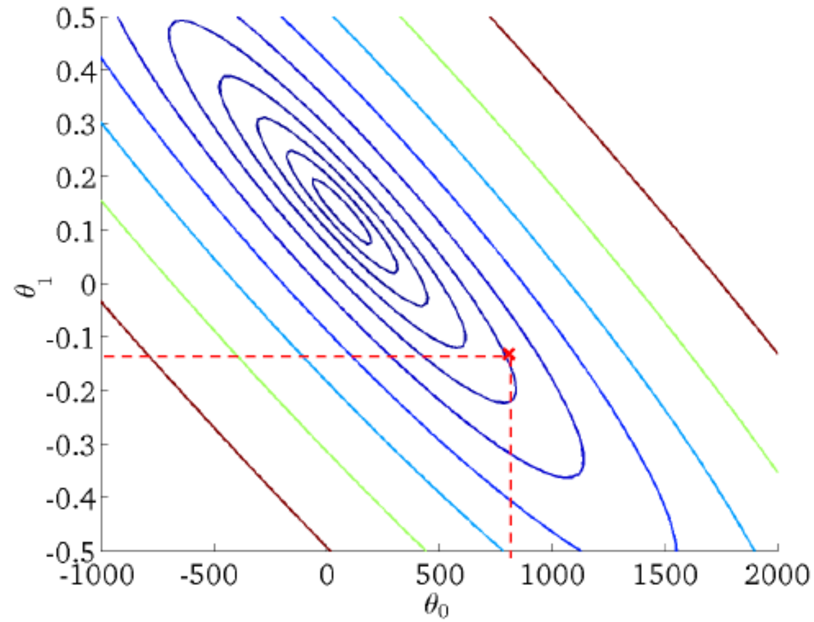
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

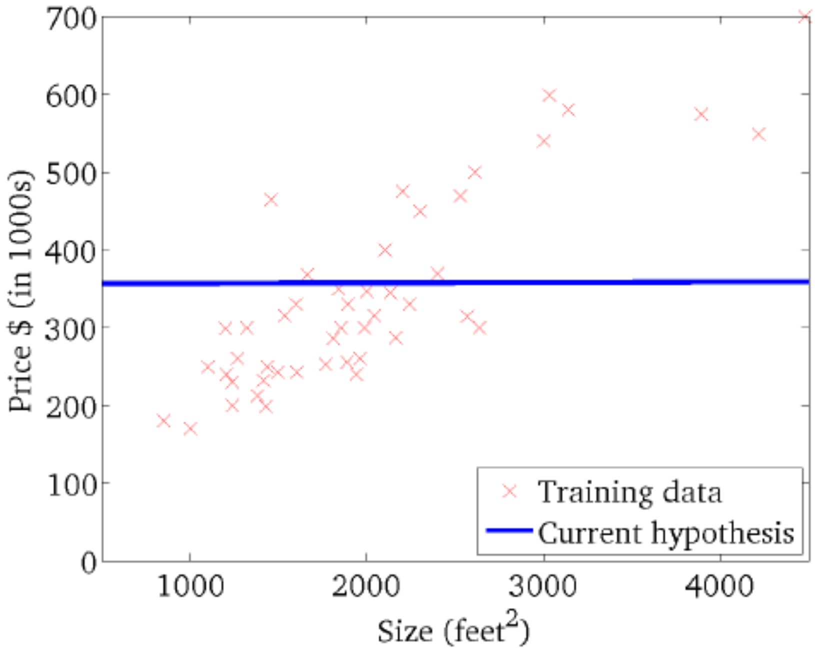
(function of the parameters θ_0, θ_1)



Relation between h and J

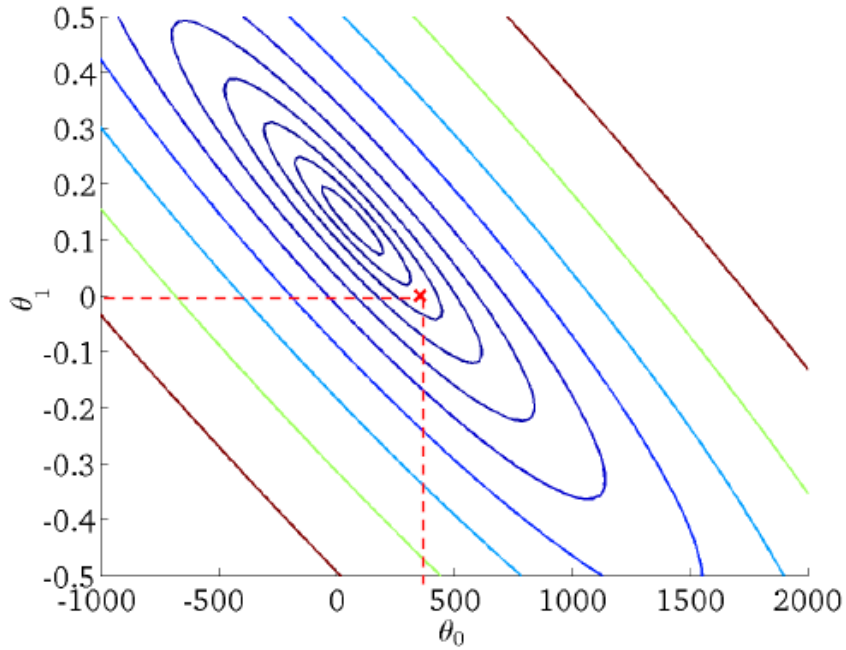
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

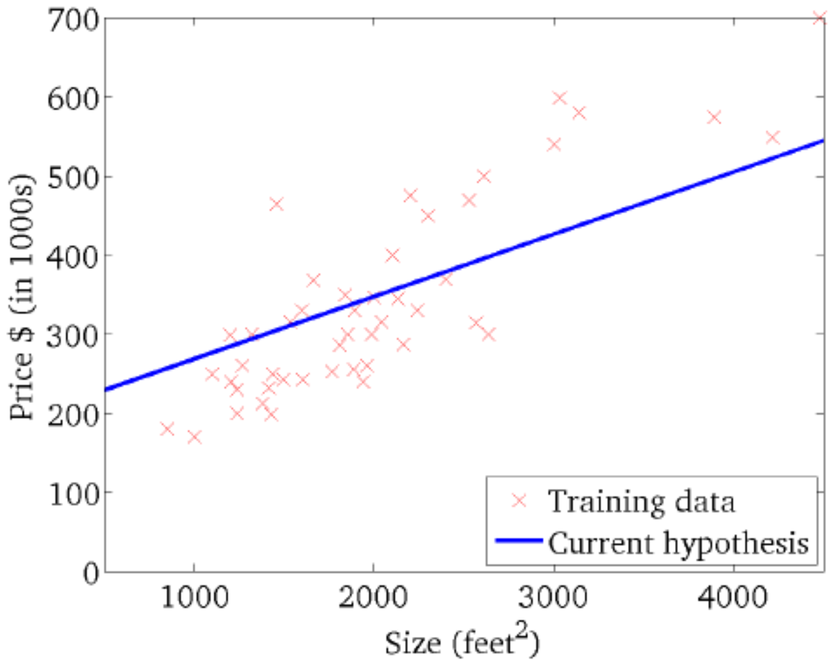
(function of the parameters θ_0, θ_1)



Relation between h and J

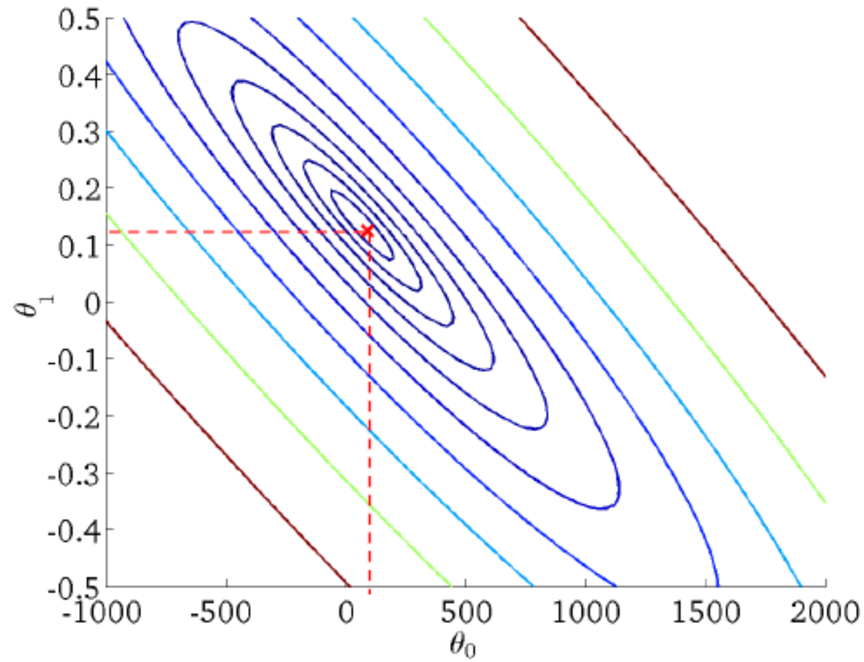
$h_\theta(x)$

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



$J(\theta_0, \theta_1)$

(function of the parameters θ_0, θ_1)



Find optimal model parameters θ to minimize MSE J

Statistical perspective

- Response has linear dependence on input with Normal noise
  - $y_i = \theta_0 + \theta_1 x_i + \epsilon_i$, $\epsilon_i \in N(0, \sigma^2)$ noise
  - $y_i | x_i \sim N(0, \sigma^2)$
  - $f(y_i | x_i; \theta, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} \mathrm{e}^{-\frac{1}{2\sigma^2}[y_i - (\theta_0 + \theta_1 x_i)]^2}$ PDF
  - One training example
- Training dataset
  - $f(y_1, \ldots, y_N | x_1, \ldots, x_N; \theta, \sigma) = \prod_{i=1}^{N} f(y_i | x_i; \theta, \sigma)$
  - Assume independence

Maximum Likelihood Estimation (MLE)

Given training data $X = \{x_1, \dots, x_N\}$ with labels $\mathrm{Y} = \{y_1, \dots, y_N\}$

What is the likelihood of training data for parameter θ?

Define likelihood function

$$Max_\theta\, L(\theta) = P[Y|X;\theta] = f(y_1, \dots, y_N|x_1, \dots, x_N;\theta)$$

Assumption: training points are independent!

$$L(\theta) = \prod_{i=1}^{N} P[y_i|x_i;\theta]$$

Log Likelihood

- Max likelihood is equivalent to maximizing log of likelihood

$$L(\theta) = \prod_{i=1}^{N} P[y_i|x_i, \theta]$$

$$\log L(\theta) = \sum_{i=1}^{n} \log P[y_i|x_i, \theta]$$

- They both have the same maximum

MLE for Linear Regression

$$L(\theta) = \prod_{i=1}^{N} P[y_i|x_i;\theta] = \prod_{i=1}^{N} f(y_i|x_i;\theta,\sigma)$$

$$\log L(\theta) = -\mathrm{c} \sum_{i=1}^{N} [y_i - (\theta_0 + \theta_1 x_i)]^2$$

Max likelihood θ is the same as Min MSE θ!

The MSE metric has statistical motivation

Solution for simple linear regression

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

$$\frac{\partial J(\theta)}{\partial \theta_0} = \frac{2}{N}\sum_{i=1N}(\theta_0 + \theta_1 x_i - y_i) = 0$$

$$\frac{\partial J(\theta)}{\partial \theta_1} = \frac{2}{N}\sum_{i=1}^{N} x_i(\theta_0 + \theta_1 x_i - y_i) = 0$$

- Solution of min loss

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

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

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

How Well Does the Model Fit?

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

$$\rho = \ Corr(X, Y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}} = \frac{\mathrm{Cov(X, Y)}}{\sigma_X \sigma_Y}$$

- Measures linear dependence between X and Y
- Positive coefficient implies positive correlation
  - The closer to 1 the coefficient is, the stronger the correlation
- Negative coefficient implies negative correlation
  - The closer to -1 the coefficient is, the stronger the correlation
- $\theta_1 = \frac{\sum \ (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
- If σ_X = σ_Y, then $\theta_1 = \mathrm{Corr(X, Y)}$

Regression vs Correlation

- Correlation
  - Find a numerical value expressing the relationship between variables
- Regression
  - Estimate values of response variable on the basis of the values of fixed variable.
- The slope of linear regression is related to correlation coefficient
- Regression scales to more than 2 variables, but correlation does not