

Machine Learning

Lecture 3: Linear Regression

杨思蓓

SIST

Email: yangsb@shanghaitech.edu.cn

Content

- Linear regression
- Normal equation
- Maximum Likelihood Estimation
- MSE v.s. MAE

1. linear regression

Housing prices

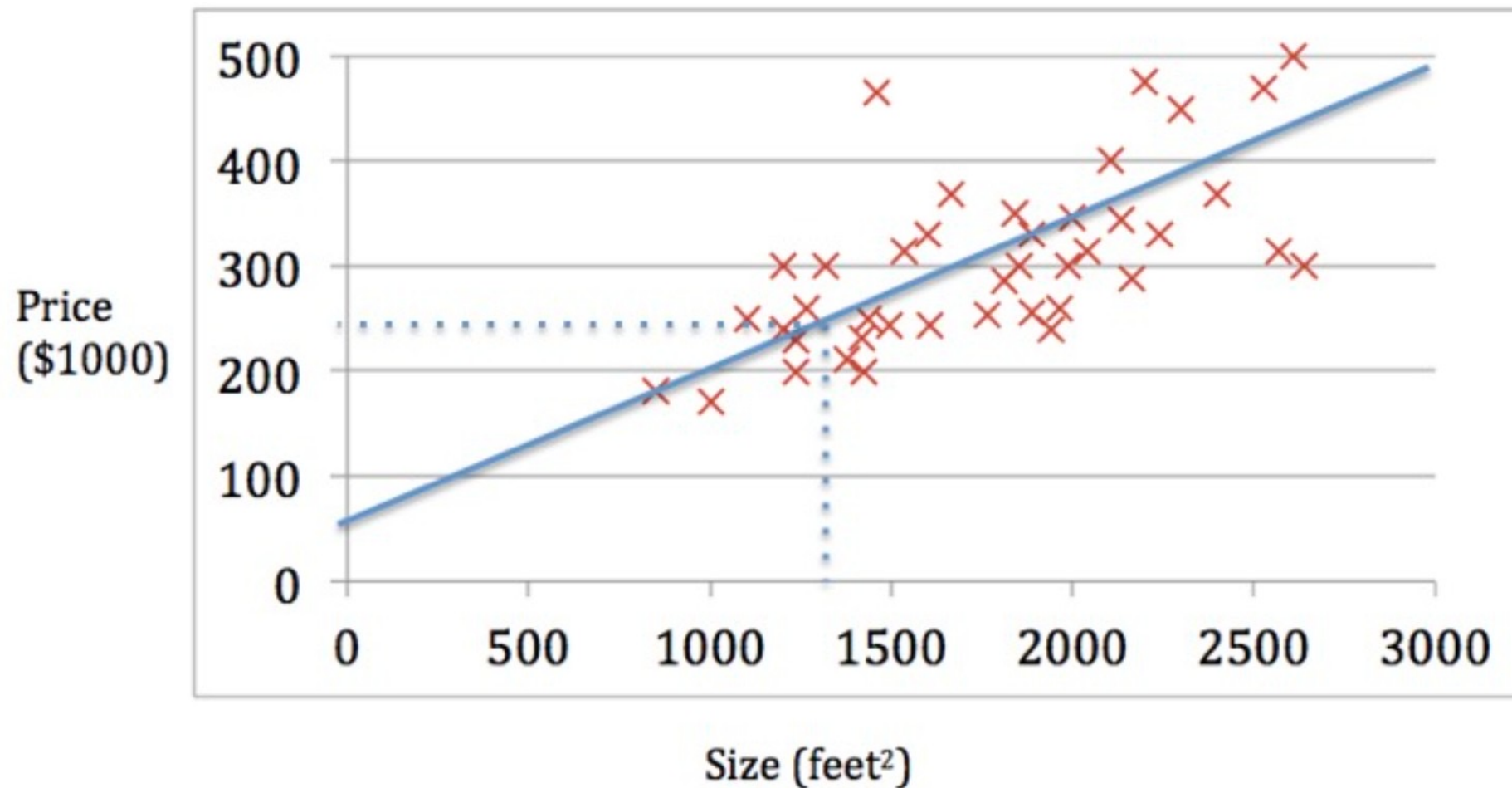
Training set of housing prices (Portland, OR)

Size in feet ₂ (x)	Price in \$1000's (y)
2104	460
1416	232
1534	315
852	178
...	...

Notation:

- m : number of training examples
- x : “Input” variables/features
- y : “Output” variables
- (x_i, y_i) : i -th training example (i -th datum)
- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$: the training set

Linear regression problems with one variable



supervised learning

Given the “right output” (labels)
for each example

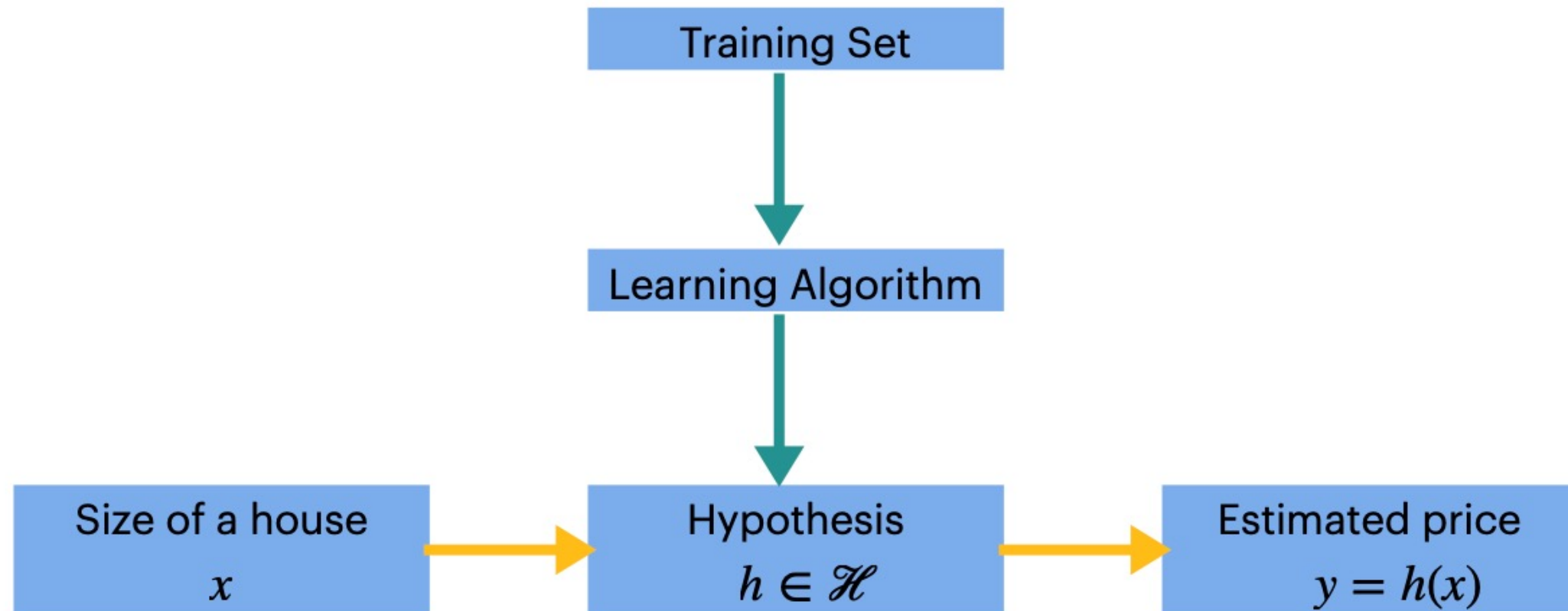
Regression problems

Predict real-valued (continuous) output

Classification problems

Predict categorical (discrete) output

Learning process



- What is our hypothesis class \mathcal{H} ?

Linear regression problems with one variable

How do we represent h ?

- Linear (affine) function

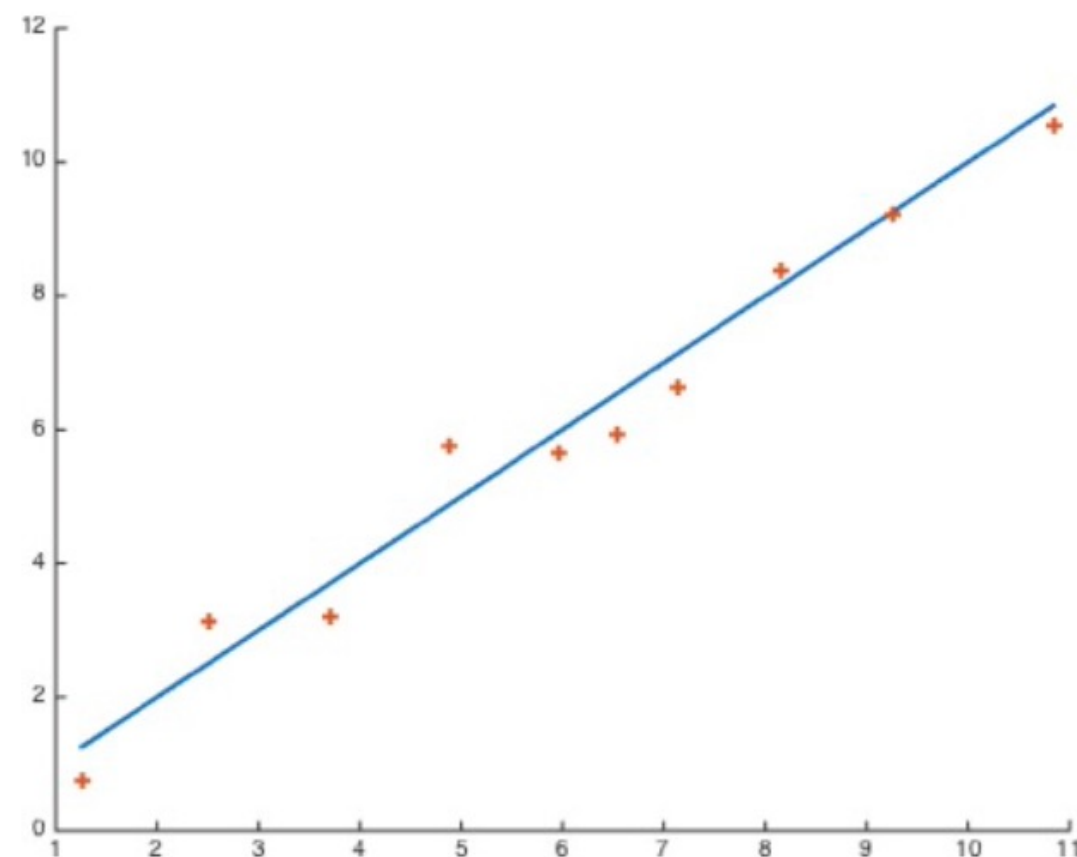
$$h(\cdot; \theta) : \mathcal{X} \rightarrow \mathcal{Y}$$

$$h(x; \theta) = \theta_0 + \theta_1 x$$

$$\theta = [\theta_0, \theta_1]^T$$

- Hypothesis class

$$\mathcal{H} = \{\theta_0 + \theta_1 x \mid \theta_0, \theta_1 \in \mathbb{R}\}$$



How do we choose the optimal h ?

Construct loss function

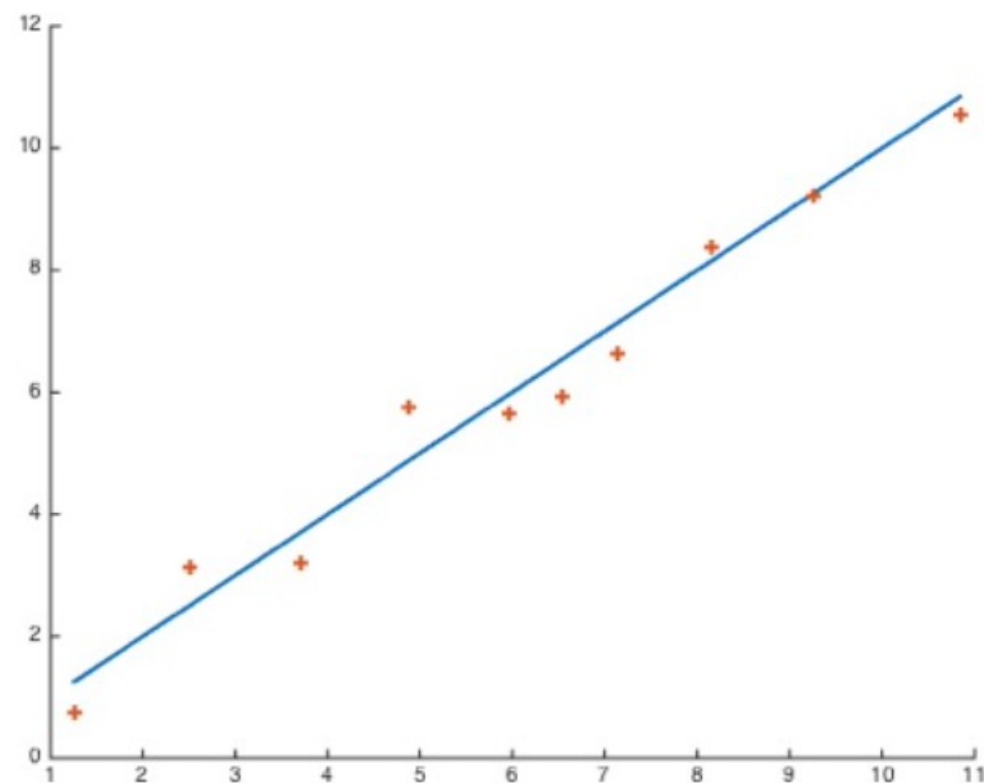
How much loss/cost is generated by this h on training set of housing prices (Portland, OR)?

How do we represent h ?

- The prediction of $h(x; \theta)$ for the i -th datum

$$h(x_i; \theta) = \theta_0 + \theta_1 x_i$$

- The “real” output y_i
- The “error” between the predicted and the “real” output
 $(h(x_i; \theta) - y_i)^2 = (\theta_0 + \theta_1 x_i - y_i)^2$



📖 Do we have to choose squared error?

Construct loss function

- The (averaged) total error on the training set

$$J(h(\cdot; \theta_0, \theta_1), \mathcal{D}) = \frac{1}{2m} \sum_{i=1}^m (h(x_i; \theta_0, \theta_1) - y_i)^2$$

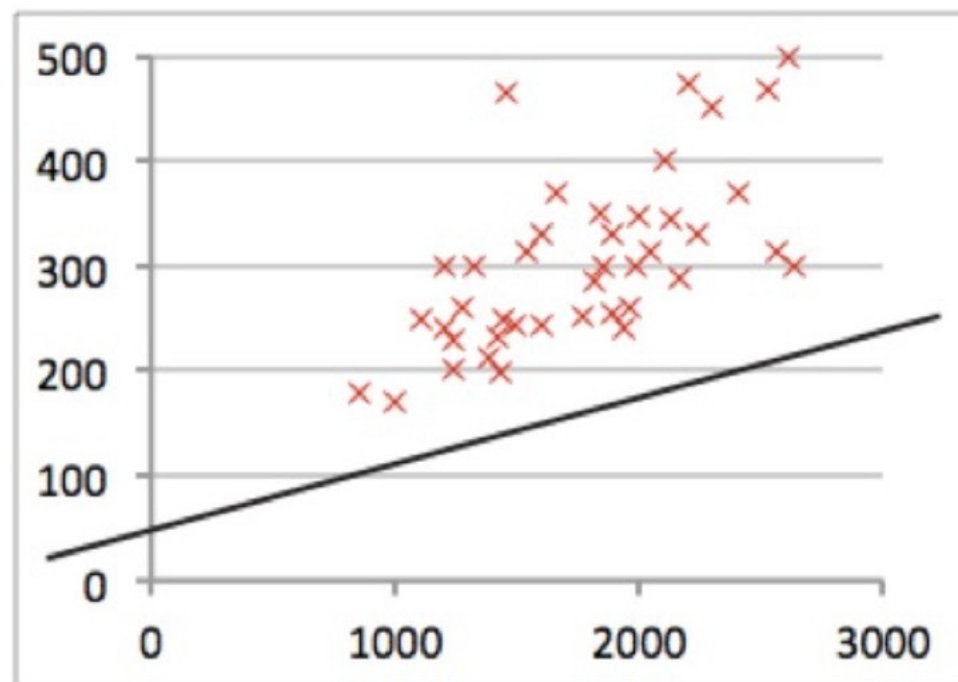
- Intuitively, h should minimize the total error $J(h(\cdot; \theta_1, \theta_2), \mathcal{D})$ over the training set \mathcal{D}
- Choose (θ_0, θ_1) as the solution of

$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h(x_i; \theta_1, \theta_2) - y_i)^2$$

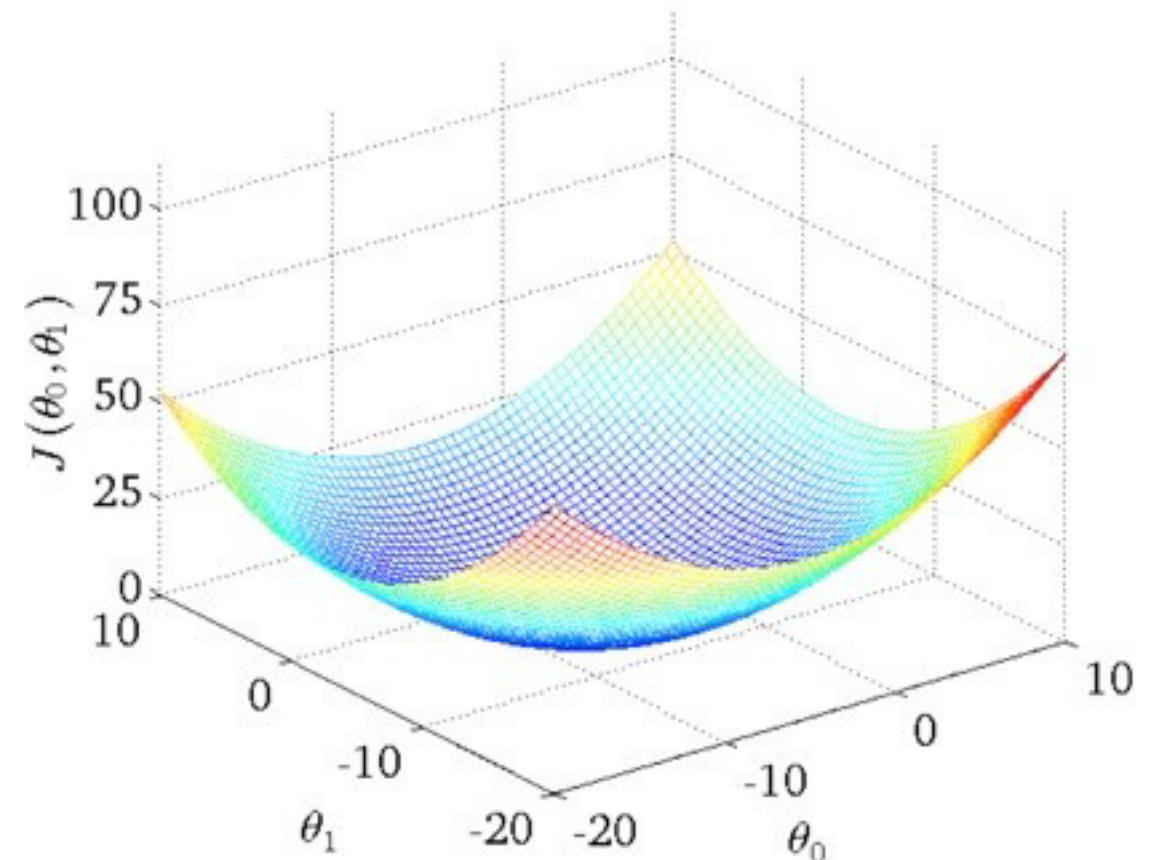
Finding optimal parameter

$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h(x_i; \theta_0, \theta_1) - y_i)^2$$

- For fixed parameters, $h(x, \theta_0, \theta_1)$ is a linear function of x
- For given training set \mathcal{D} , $J(\theta_0, \theta_1)$ is quadratic of (θ_0, θ_1)



$$h(x; 50, 0.06) = 50 + 0.06x$$



Linear regression with multiple variable



Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)	Zipcode
2104	5	1	45	460	18015
1416	3	2	40	232	18014
1534	3	2	30	315	18323
852	2	1	36	178	18001
...		

- Multiple variables

$$h(\mathbf{x}) = h(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

- For convenience of notion, define

$$x_0 = 1, \quad \mathbf{x} = [x_0, x_1, \dots, x_n]^T$$

- Example:

$$h(\mathbf{x}) = 80x_0 + 0.1x_1 + 0.01x_2 + \dots + 3x_3 - 2x_4$$

Dummy coding

- ▶ Continuous variables (连续变量) : Real-valued numbers
Examples: 23; 1000.999
- ▶ Ordinal variables (有序类别变量) : A variable that can take on one of a number of possible values, thus assigning each individual to a particular group or “category”, and is ordered or ranked
Examples: integer 1, 2, 3, 4...; “strongly disagree, disagree, neutral, agree, strongly agree”
- ▶ Categorical variables ((无序) 类别变量)
A variable that can take on one of a limited, and usually fixed, number of possible values, thus assigning each individual to a particular group or “category”, but is NOT ordered or ranked
Examples: “yes, no”; “female, male”; zipcode

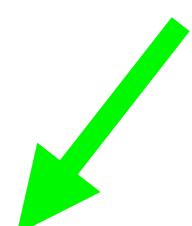
Dummy coding

District	Price 1000's (y)
徐汇	100
浦东	90
闵行	80
浦东	30
浦东	70
⋮	⋮

One way of coding:

徐汇	闵行	浦东	Price 1000's (y)
1	0	0	100
0	0	1	90
0	1	0	80
0	0	1	30
0	0	1	70
⋮	⋮		

Nonlinear Transformation: “Linear” w.r.t. parameters



Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)	Zipcode
2104	5	1	45	460	18015
1416	3	2	40	232	18014
1534	3	2	30	315	18323
852	2	1	36	178	18001
...		

Linear models:

$$h(\mathbf{x}) = h(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_n x_n$$

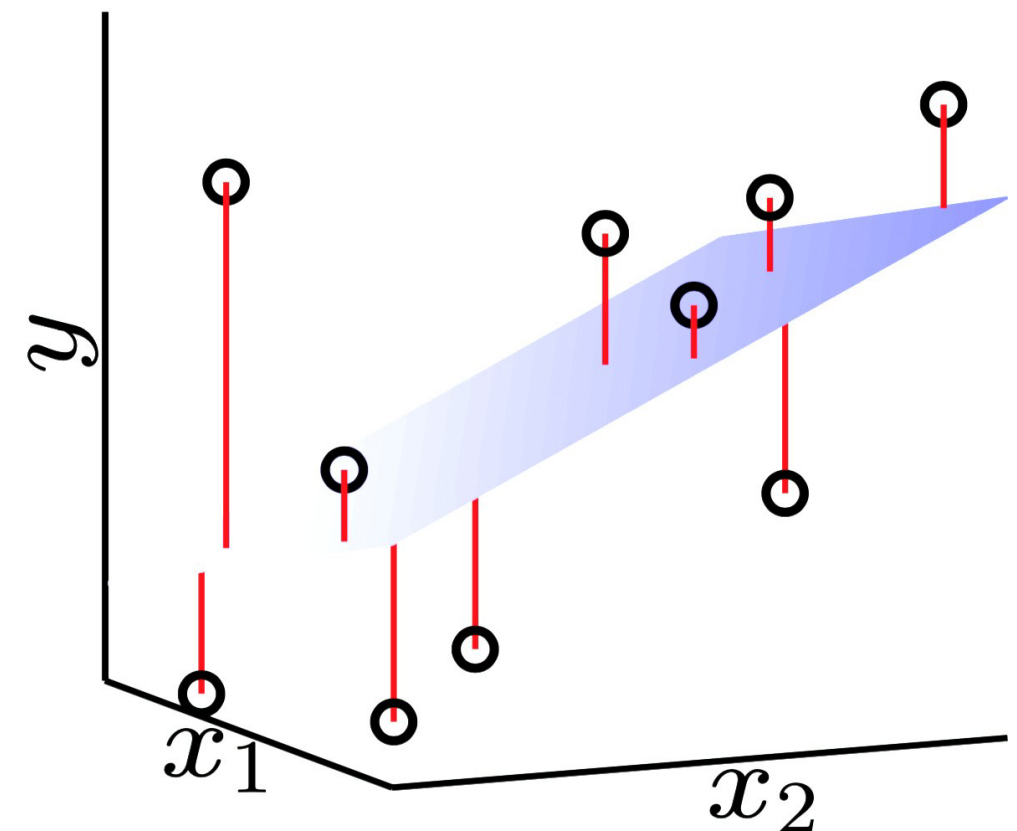
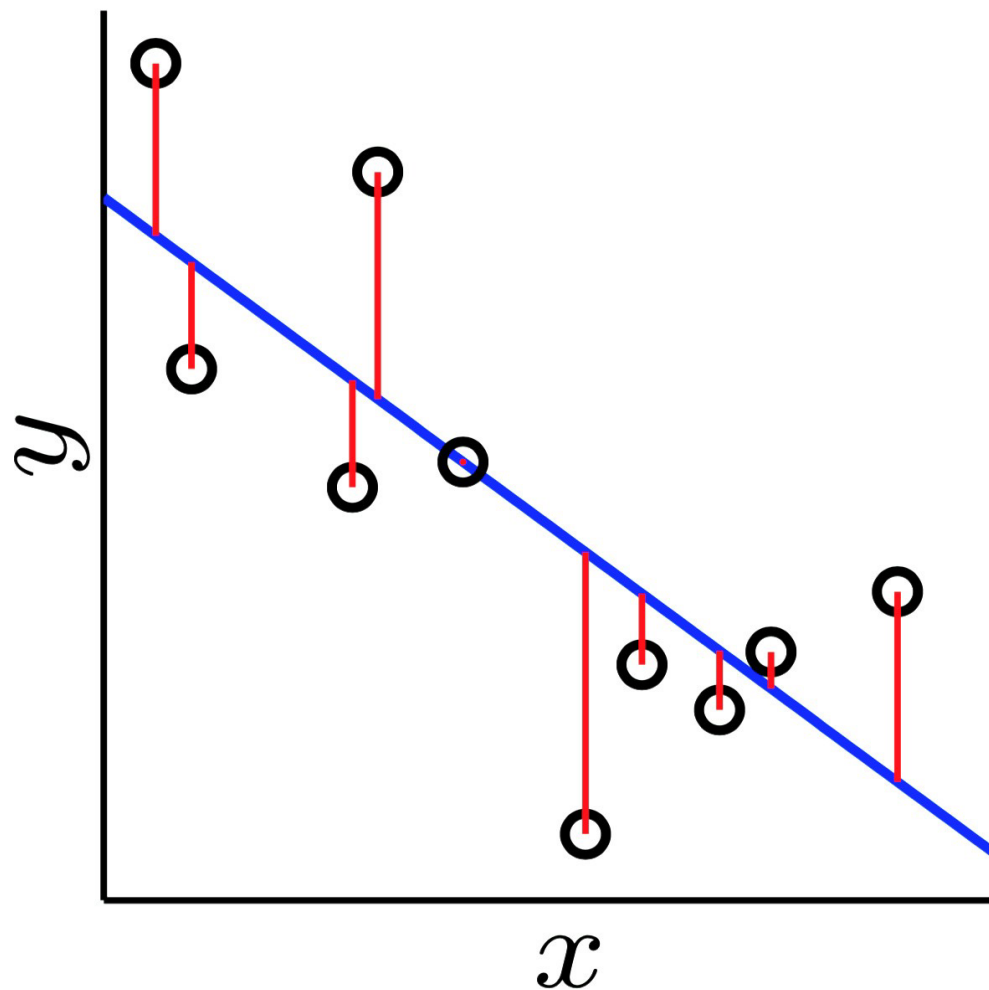
$$h(\mathbf{x}) = h(\mathbf{x}; \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 x_2 + \theta_2 x_1^2 (x_2 - x_3)$$

Linear regression with multiple variable

- Multivariate linear regression

$$h(\mathbf{x}) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2$$

Illustration of linear regression



Linear regression with multiple variable

► Notation:

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_m^T \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y^m \end{bmatrix}$$

$$\mathbf{x}_i = [x_{i,0}, x_{i,1}, x_{i,2}, \dots, x_{i,n}]^T$$

$$x_{i,j}, i = 1, \dots, m; j = 1, \dots, n$$

- Hypothesis: $h(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x}$
- Hypothesis class: $\mathcal{H} = \{h(\mathbf{x}) \mid \boldsymbol{\theta} \in \mathbb{R}^{n+1}\}$
- Parameters/Cost function variables: $\boldsymbol{\theta}$
- Cost function: $J(\boldsymbol{\theta}) = \frac{1}{2} \|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$ with gradient

$$\nabla J(\boldsymbol{\theta}) = \mathbf{X}^T (\mathbf{X}\boldsymbol{\theta} - \mathbf{y})$$

Finding the optimal parameters: Gradient Descent Method

- ▶ Batch Gradient: $\nabla J(\boldsymbol{\theta}) = \sum_{i=1}^m (\mathbf{x}_i^T \boldsymbol{\theta} - y_i) \mathbf{x}_i$
- ▶ Batch Gradient Descent: until convergence, repeat

$$\boldsymbol{\theta}^{k+1} \leftarrow \boldsymbol{\theta}^k - \alpha^k \nabla J(\boldsymbol{\theta}^k)$$

2. Normal Equation

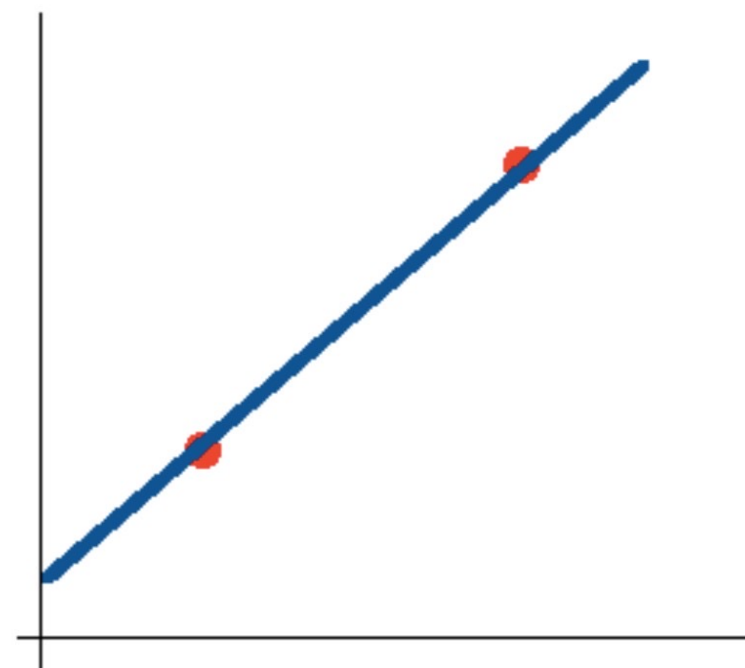
MSE (均方差) and SSE (和方差)

- ▶ Consider a simple problem

- One feature, two data points
- Two unknowns: θ_0, θ_1
- Two equations

$$y_1 = \theta_0 + \theta_1 x_1 \quad (4.1)$$

$$y_2 = \theta_0 + \theta_1 x_2 \quad (4.2)$$



- ▶ Can solve this system directly

$$\mathbf{X}\boldsymbol{\theta} = \mathbf{y} \quad \rightarrow \quad \hat{\boldsymbol{\theta}} = \mathbf{X}^{-1}\mathbf{y}$$

- ▶ However, most of the time, $m > n$

- ▶ There may be no linear function that hits all the data exactly
- ▶ Instead, solve directly for minimum of MSE function

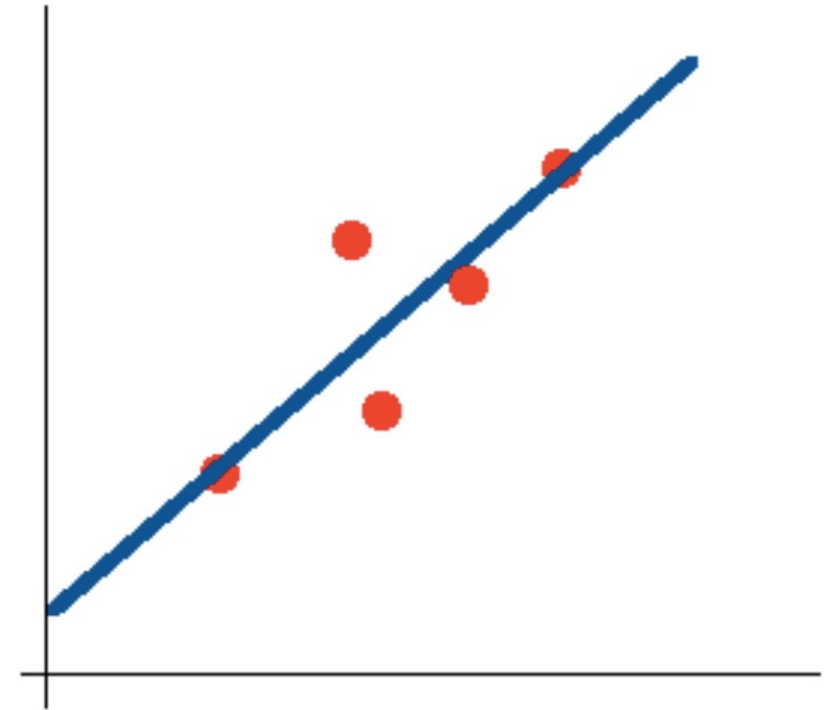
Normal equation (法方程)

- ▶ Reordering, we have

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} - \mathbf{X}^T \mathbf{y} = 0$$

$$\mathbf{X}^T \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^T \mathbf{y}$$

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^+ \mathbf{X}^T \mathbf{y}$$



- ▶ $(\mathbf{X}^T \mathbf{X})^+$ is called the “pseudo-inverse” of $\mathbf{X}^T \mathbf{X}$ (伪逆矩阵)
- ▶ If $\mathbf{X}^T \mathbf{X}$ has full rank, this is the inverse $(\mathbf{X}^T \mathbf{X})^+ = (\mathbf{X}^T \mathbf{X})^{-1}$
- ▶ If $m > n$, overdetermined; gives minimum MSE fit

A^+ is called the pseudo-inverse of A if it satisfies:

- ▶ $AA^+A = A$
- ▶ $(AA^+)^T = AA^+$
- ▶ $A^+AA^+ = A^+$
- ▶ $(A^+A)^T = A^+A$

Examples: $m = 4$.

	Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)
x_0	x_1	x_2	x_3	x_4	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

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \end{bmatrix}$$

$$y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

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

Examples: $m = 5$.

	Size (feet ²)	Number of bedrooms	Number of floors	Age of home (years)	Price (\$1000)
x_0	x_1	x_2	x_3	x_4	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
1	3000	4	1	38	540

$$X = \begin{bmatrix} 1 & 2104 & 5 & 1 & 45 \\ 1 & 1416 & 3 & 2 & 40 \\ 1 & 1534 & 3 & 2 & 30 \\ 1 & 852 & 2 & 1 & 36 \\ 1 & 3000 & 4 & 1 & 38 \end{bmatrix}$$

$$y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \\ 540 \end{bmatrix}$$

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

Interpretation

- ▶ $\mathbf{X}\boldsymbol{\theta} - \mathbf{y}$ is the vector of errors in each example
- ▶ \mathbf{X} are the features we have to work with for each example
- ▶ Inner product = 0: orthogonal (垂直)

