Data Mining

Classification V - Simple Linear Regression

Dr. Jason T.L. Wang, Professor

Department of Computer Science

New Jersey Institute of Technology

Training data set:

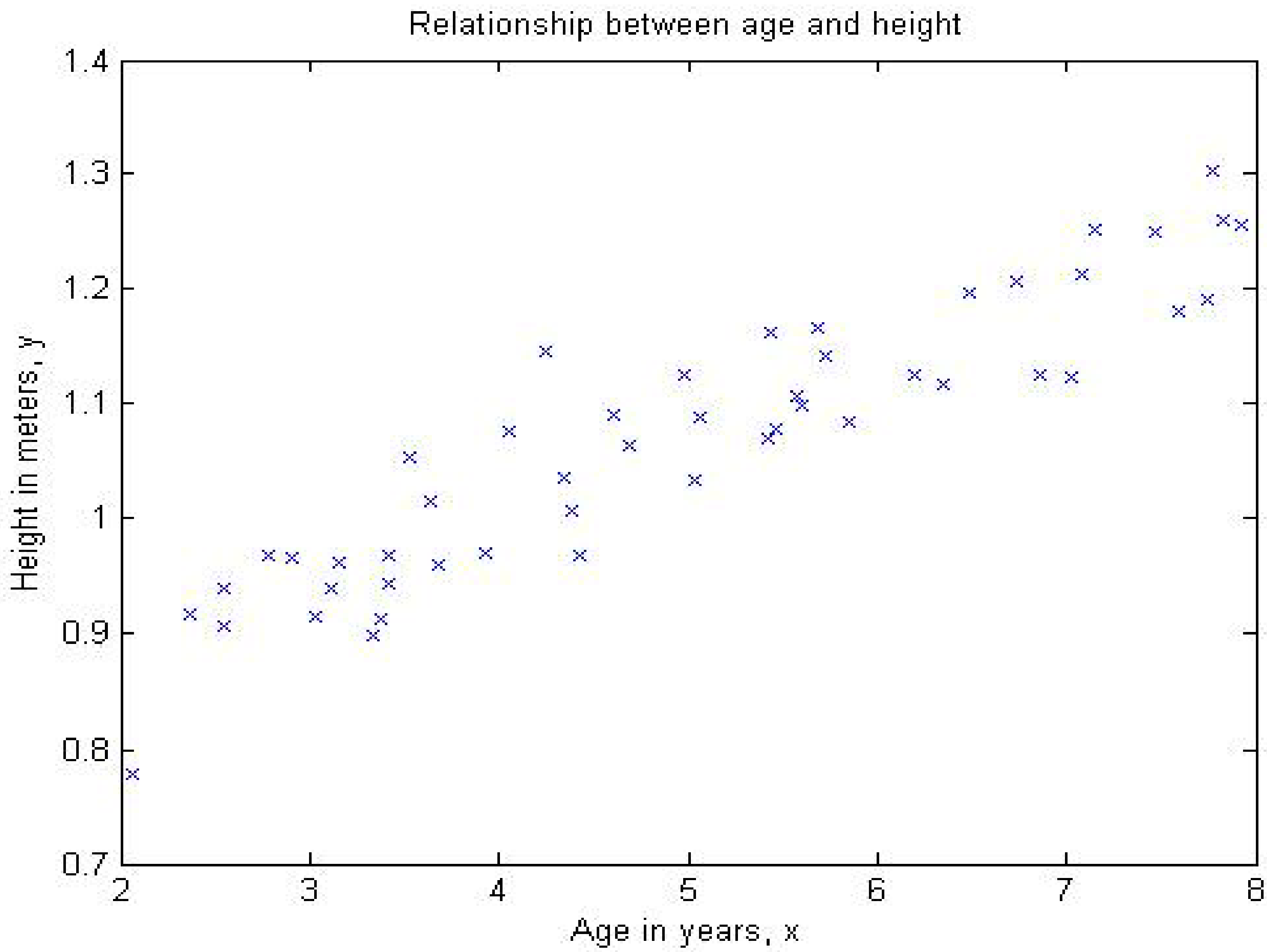
Age in years (x)	Height in meters (y)
2.06	0.78
4.25	1.15
7.47	1.25
…	…

Notation:

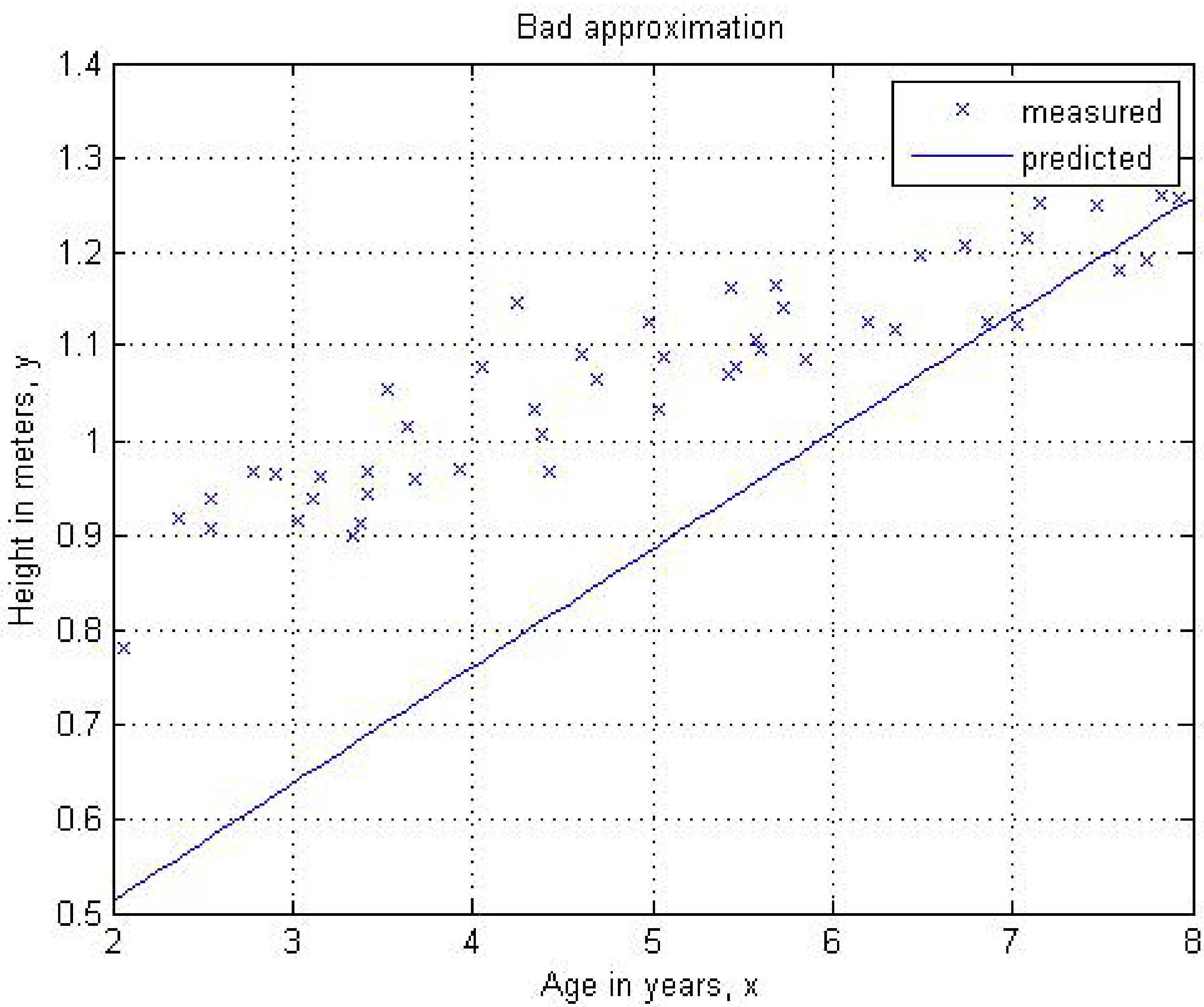
m = Number of training examples
x's = "input" variable / feature values
y's = "output" variable / "target" variable values

Regression problem:
Predict real-value output given some input.

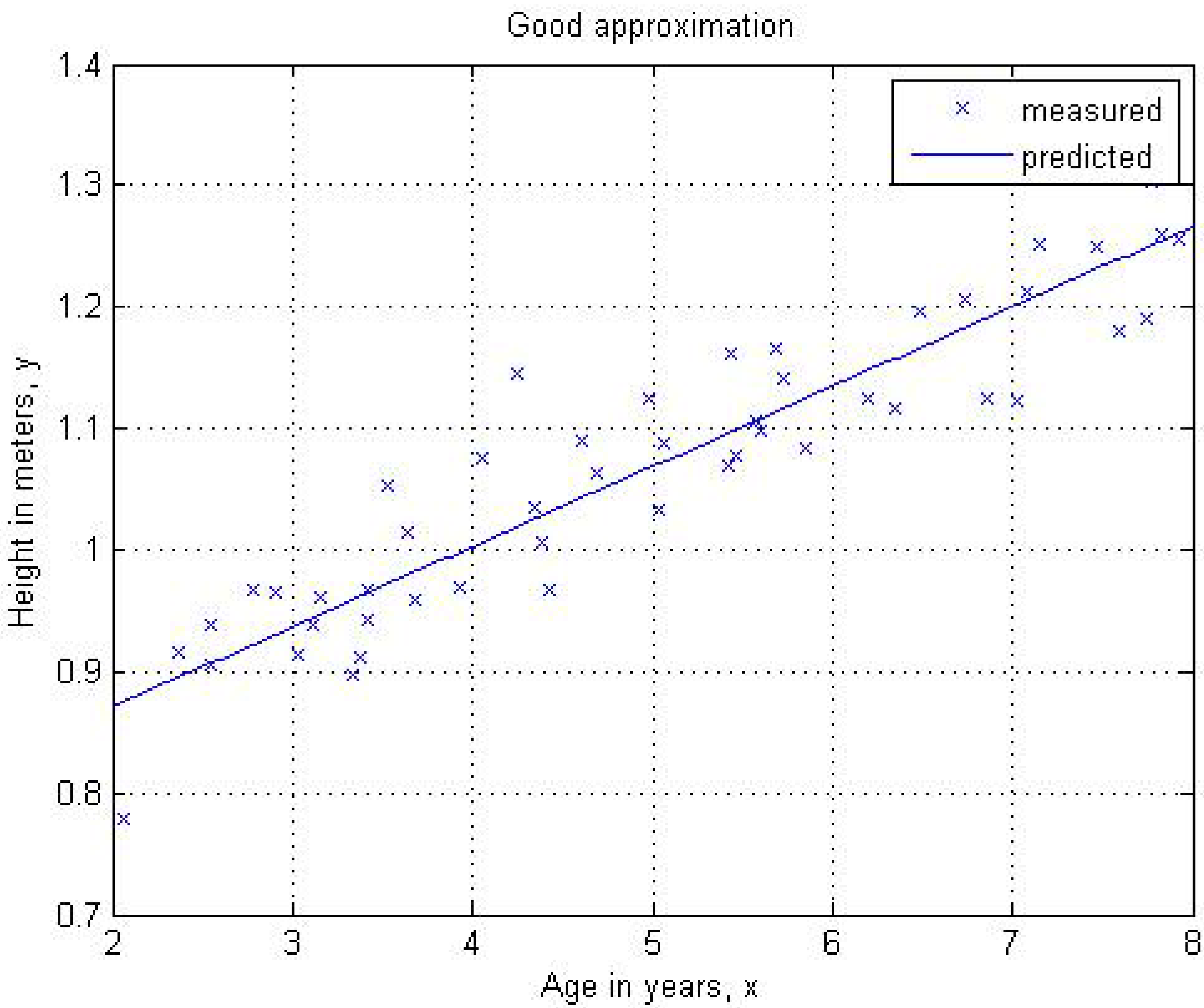
Simple linear regression



Bad approximation



Good approximation



Regression vs. Classification

- Similarities: Both algorithms learn from a training data set.
- Differences: In classification, we deal with training examples that have categorical attributes (e.g. gender) with unordered values (e.g. male, female). In regression, we deal with training examples that have continuous values.

Simple linear model $H_{a,b}(x) = ax + b$

Training data

$\downarrow$

Learning algorithm

$\downarrow$

$\mathit{Age\ in\ years} \rightarrow H_{a,b} \rightarrow \mathit{Estimated\ height}$

Cost function $Q(a,b)$

- $Model\ H(x) = ax + b$
- $Q(a,b) = \frac{1}{2m}\sum_{i=1}^{m}(H(x_i) - y_i)^2$

$$= \frac{1}{2m}\sum_{i=1}^{m}(ax_i + b - y_i)^2$$

- Goal: minimize the cost function i.e. Find $\min\limits_{a,b} Q(a,b)$

Analytical method (Ordinary Least Squares)

$$Q(a,b) = \frac{1}{2m}\sum_{i=1}^{m}(ax_i + b - y_i)^2$$

$$Let\ \frac{\partial Q(a,b)}{\partial a} = 0\ and\ \frac{\partial Q(a,b)}{\partial b} = 0$$

We have:

$$\frac{\partial Q(a,b)}{\partial a} = \frac{1}{m}\sum_{i=1}^{m} x_i(ax_i + b - y_i) = 0 \quad \text{Eq. (1)}$$

$$\frac{\partial Q(a,b)}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} (ax_i + b - y_i) = 0 \quad \text{Eq. (2)}$$

Simplifying equations (1) and (2) leads to the following linear system of a and b:

$$mb + a\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} y_i$$

$$b\sum_{i=1}^{m} x_i + a\sum_{i=1}^{m} {x_i}^2 = \sum_{i=1}^{m} y_i x_i$$

Solving the linear system, we get analytical solutions for a and b.

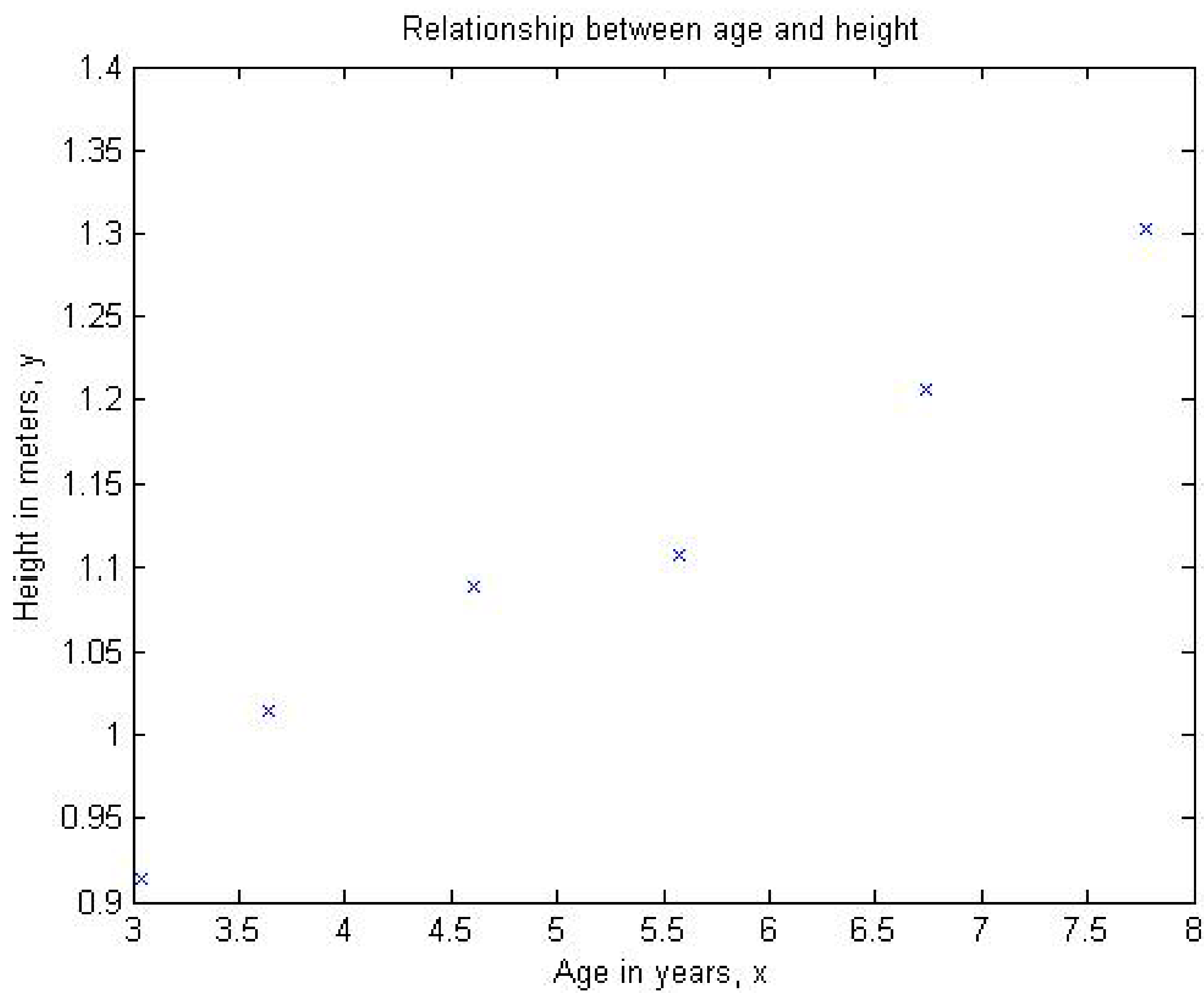
$$a = \frac{\sum_{i=1}^{m} y_i x_i \; - \; \frac{\left(\sum_{i=1}^{m} y_i\right)\left(\sum_{i=1}^{m} x_i\right)}{m}}{\sum_{i=1}^{m} {x_i}^2 \; - \; \frac{\left(\sum_{i=1}^{m} x_i\right)^2}{m}}$$

$$b = \frac{1}{m}\sum_{i=1}^{m} y_i \; - \frac{a}{m}\sum_{i=1}^{m} x_i$$

Example:

Training data set:
$m = 6$
Age in years (x):
$x_1, x_2, x_3, x_4, x_5, x_6$
Height in meters (y):
$y_1, y_2, y_3, y_4, y_5, y_6$

Age in years (x)	**Height in meters (y)**
3.04	0.91
3.64	1.01
4.61	1.09
5.57	1.11
6.74	1.20
7.77	1.30


Relationship between age and height
Height in meters, y
Age in years, x
1.4
1.35
1.3
1.25
1.2
1.15
1.1
1.05
1
0.95
0.9
3
3.5
4
4.5
5
5.5
6
6.5
7
7.5
8

Compute:

$$\sum_{i=1}^{m} x_i = 31.37 \qquad \sum_{i=1}^{m} y_i = 6.62$$

$$\sum_{i=1}^{m} {x_i}^2 = 180.569 \qquad \sum_{i=1}^{m} y_i x_i = 35.839$$

$$a = \frac{\sum_{i=1}^{m} y_i x_i \; - \; \frac{\left(\sum_{i=1}^{m} y_i\right)\left(\sum_{i=1}^{m} x_i\right)}{m}}{\sum_{i=1}^{m} {x_i}^2 \; - \; \frac{\left(\sum_{i=1}^{m} x_i\right)^2}{m}}$$

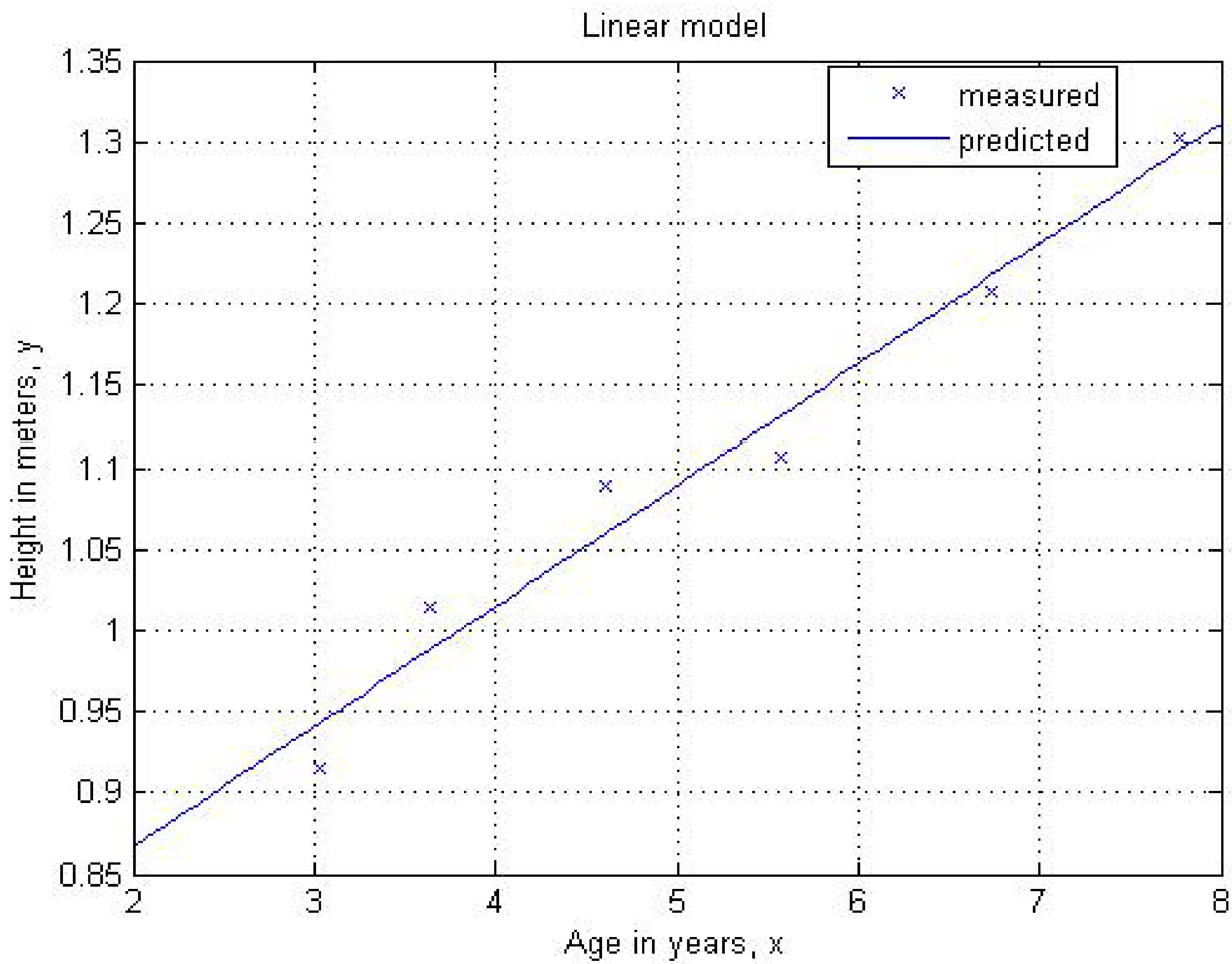
$$= \frac{35.839 \; - (6.62 * 31.37)/6}{180.569 - 31.37 * 31.37/6} = 0.0741$$

$$b = \frac{1}{m}\sum_{i=1}^{m} y_i \; - \frac{a}{m}\sum_{i=1}^{m} x_i$$

$$= \frac{6.62}{6} - \frac{0.0741 * 31.37}{6} = 0.716$$

$$y = 0.0741x + 0.716$$

Linear model for the relationship of age and height



Predict the height of a five-year old boy

$$y = 0.0741 \times 5 + 0.716 = 1.09$$

The predicted height is 1.09 m

End of
Simple Linear Regression Module