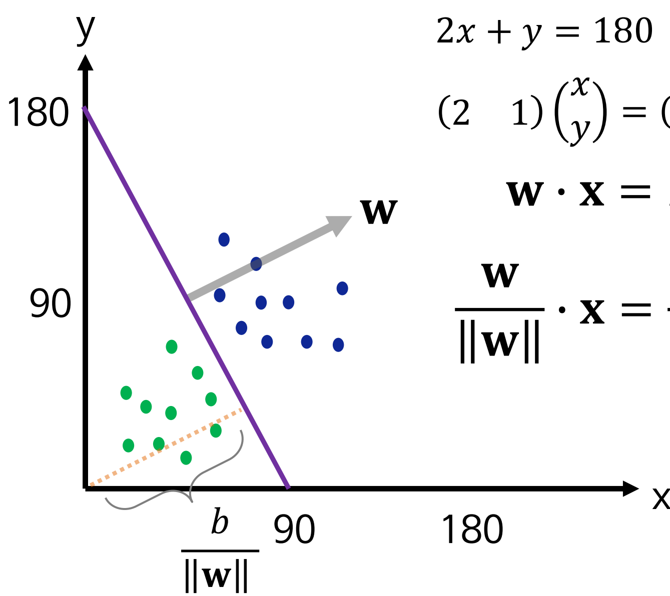
Linear regression

Vectors / Matrices in machine learning (classification)

- Training model and data are represented in vectors
  - To train a model, we utilize matrix and vector operations



$$2x + y = 180$$

$$\begin{pmatrix} 2 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = (180)$$

$$\mathbf{w} \cdot \mathbf{x} = b$$

$$\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x} = \frac{b}{\|\mathbf{w}\|}$$

Linear regression

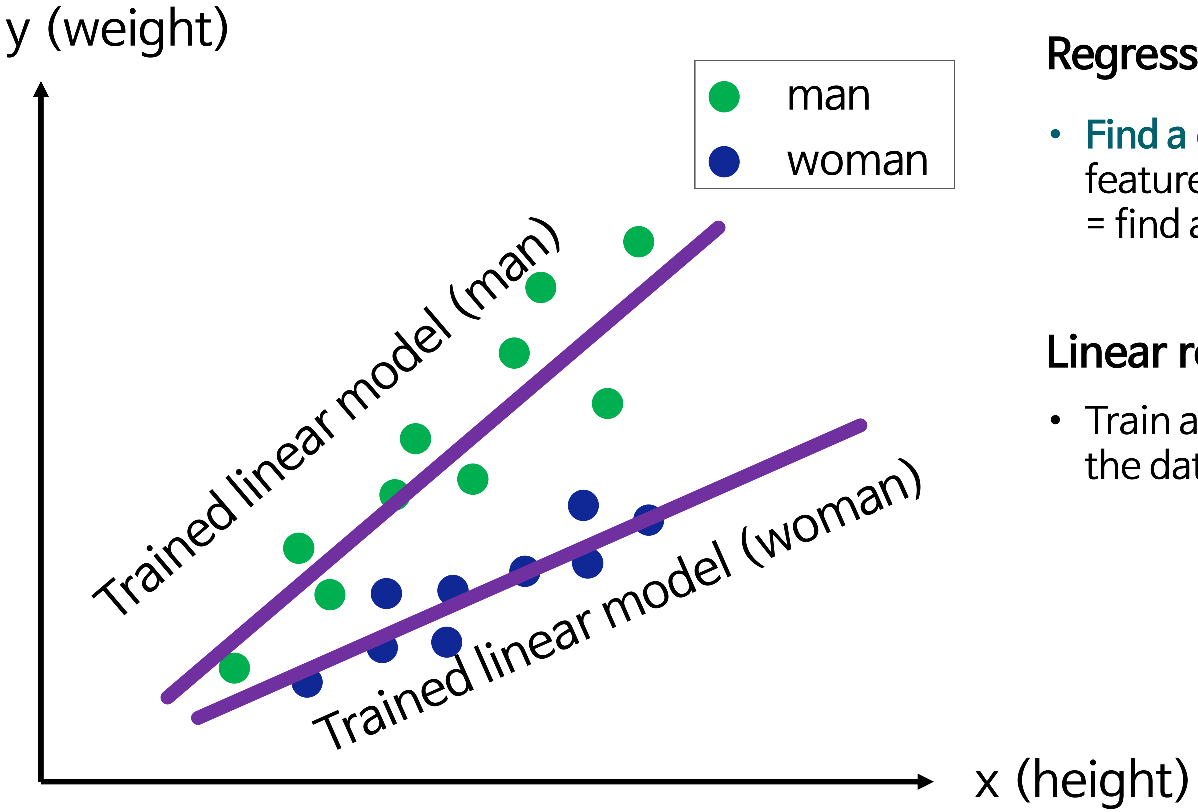
Supervised learning

- Linear regression
- Linear classification

Linear regression

Introduction to linear regression

- Linear regression (week #3)



An example of linear regression

Regression problem

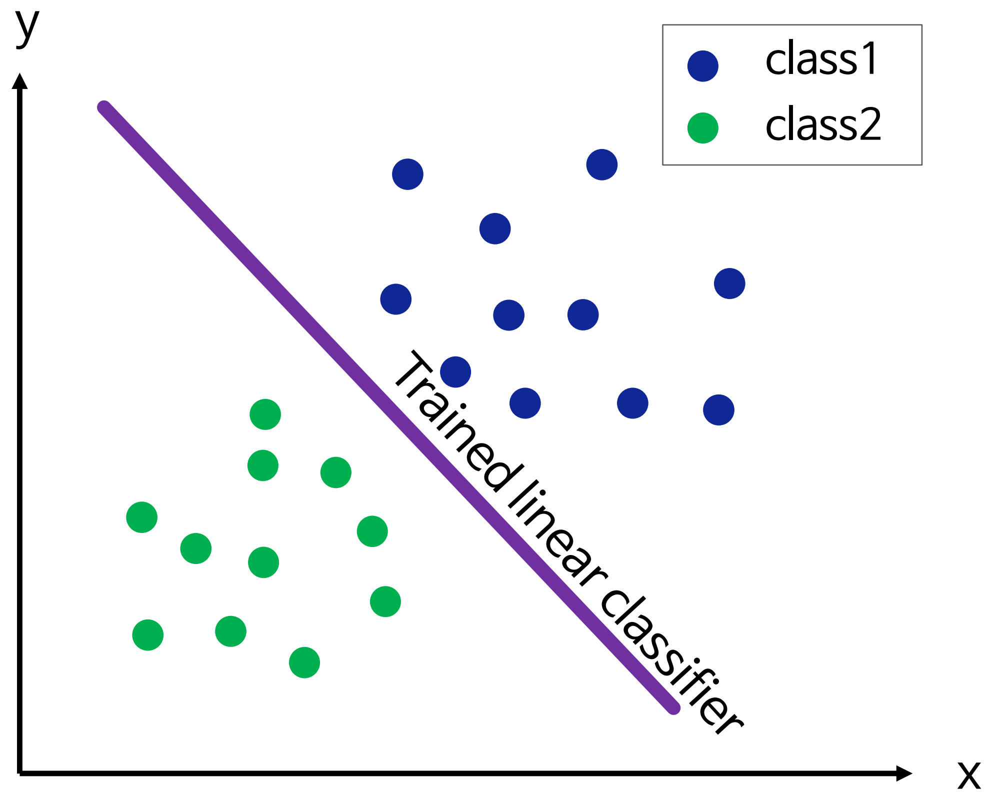
- **Find a correlation** between features
  = find a distribution of dataset

Linear regression problem

- Train a linear model that explains the data distribution

Introduction to linear regression

- Linear classification (week #4)

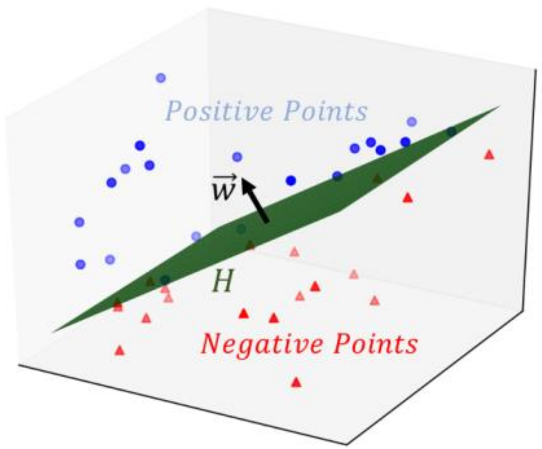


Classification problem

- Train a classifier that classifies the data

Linear Classification problem

- Train a linear classifier

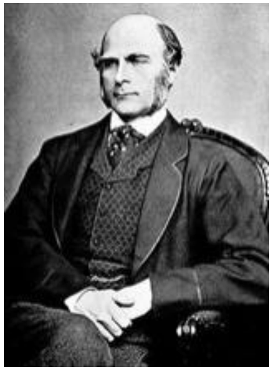


An example of linear classifier

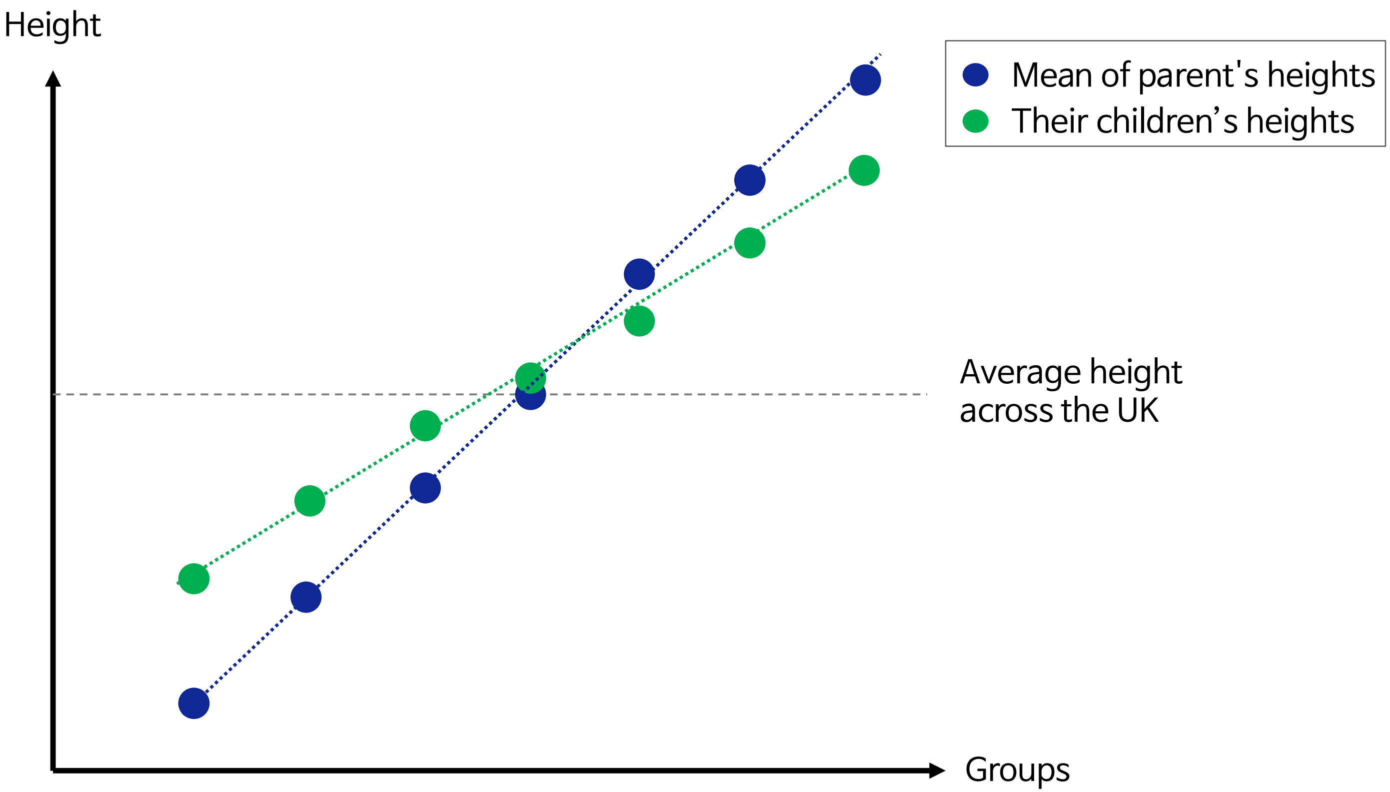
Linear regression

Introduction to linear regression

- Why 'regression' ?
  : The height of the children is linearly related to the average height of the parents
  (Assumption) It **has a tendency to return (=regress) to the average height of the population.**



Galton, Statistician
19c, England

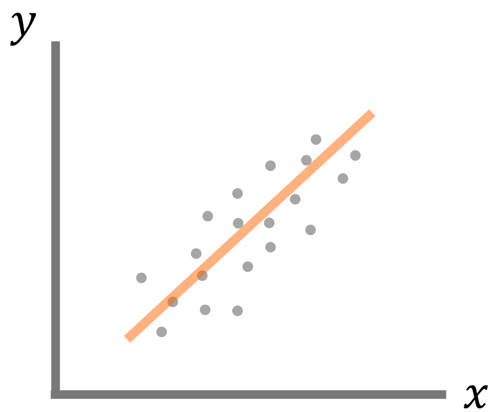


From the paper "**Regression** towards Mediocrity in Hereditary Stature"

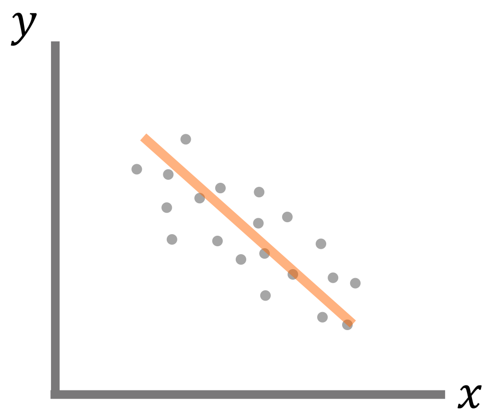
Linear regression

Correlation

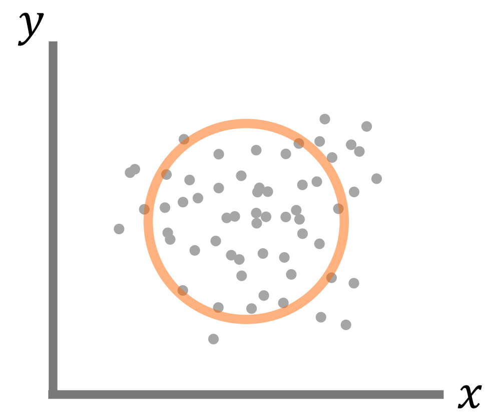
- Correlation is a statistic that measures the degree to which **two variables move in relation** to each other



Positive correlation



Negative correlation



No correlation

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

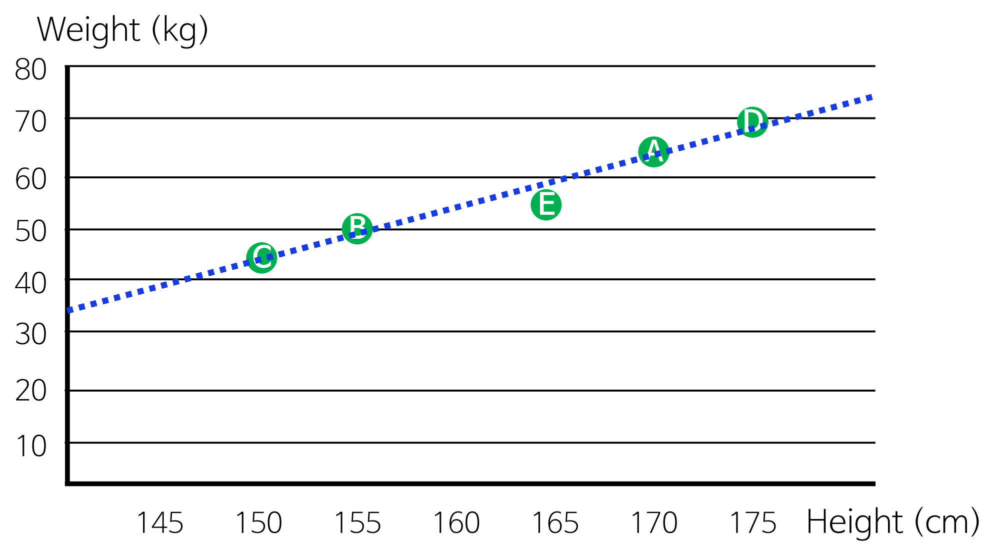
Linear regression

Training a linear regression model

- Let's train a linear regression model that analyzes the correlation between height and weight
  : (Assumption) Height and weight will have **a positive correlation**

Training dataset : 5 people's heights, weights

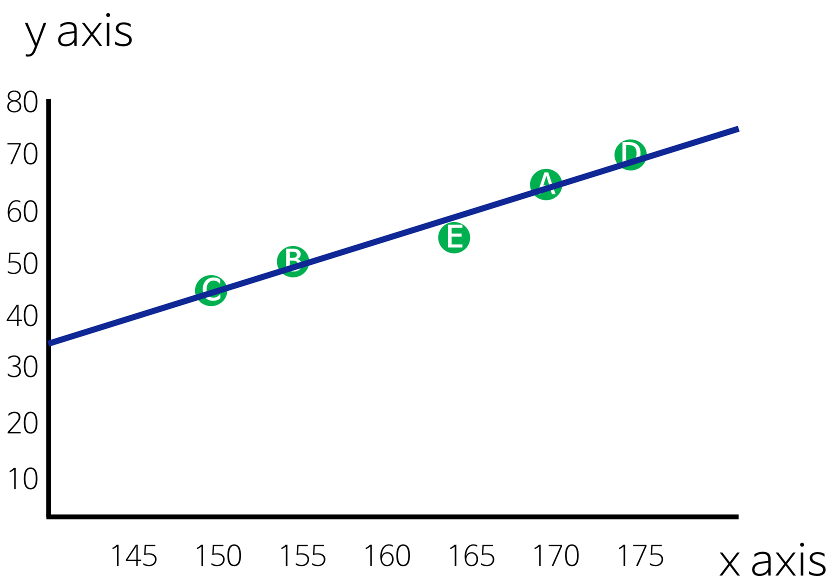
index	height (cm)	weight (kg)
A	170	65
B	155	50
C	150	45
D	175	70
E	165	55



Linear regression

Training a linear regression model

- What should we train?
  : Linear function that represents the relationship between x and y linearly (straight line)
  Parameters: slope (a), y- intercept (b)



A trained linear model

$$y = ax + b$$

- a: slope of a line
- b: y intercept

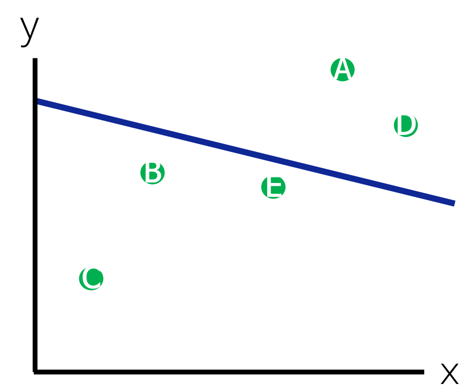
The parameters of linear regression model

$\theta: a, b$

Training a linear regression model

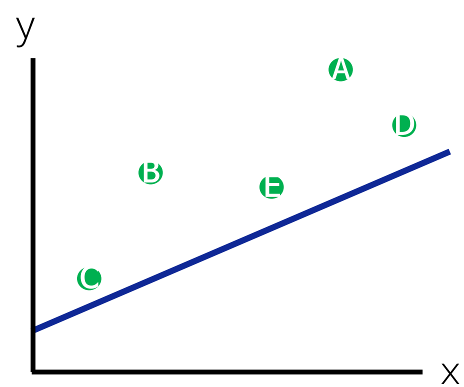
- Training objectives
  :Train a linear model to best describe (or predict) given data samples

Result 1



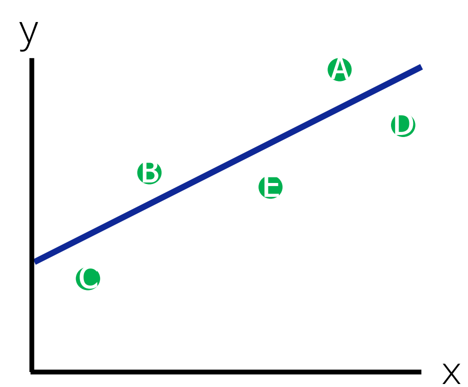
$$y = -0.3x + 80$$

Result 2



$$y = 0.3x + 10$$

Result 3



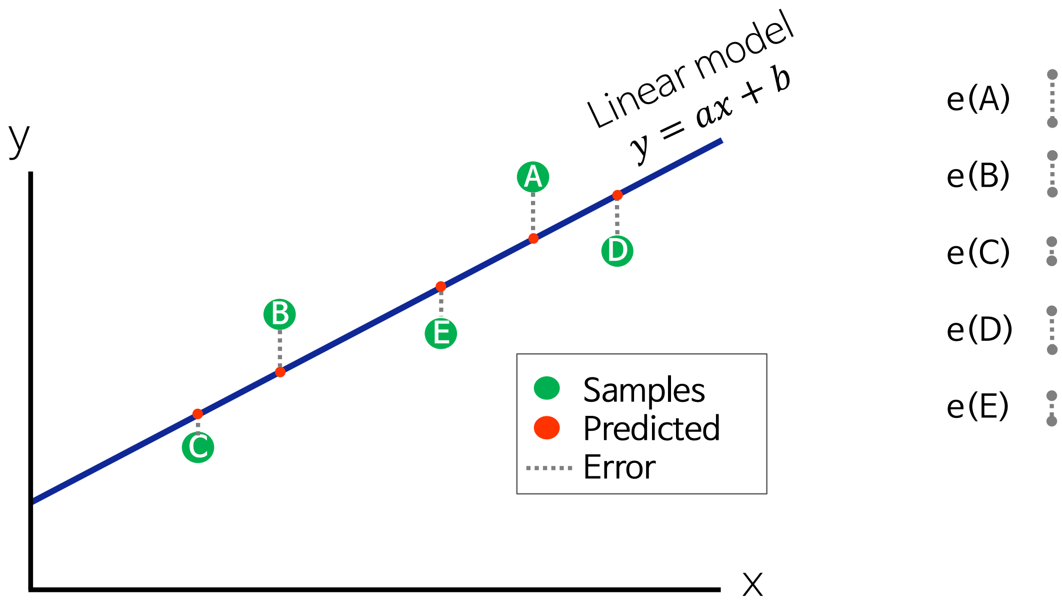
$$y = 0.3x + 35$$

Which one is the best?

Linear regression

Training a linear regression model

- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals



Find optimal **a, b**
that minimize
sum of error

$$\underset{a,b}{\operatorname{argmin}} \sum error$$

Linear regression

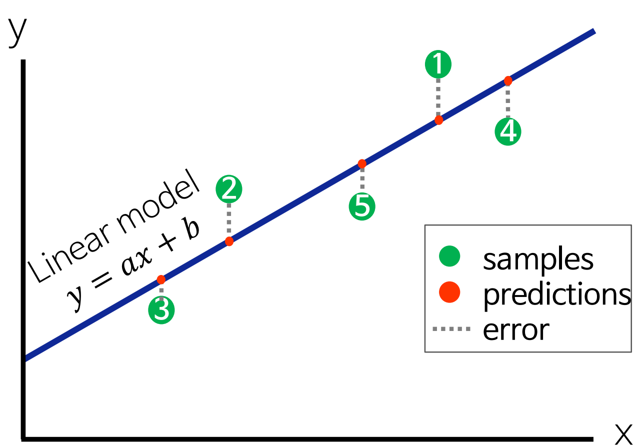
Training a linear regression model

- How can we train a linear regression model?
- Least square method

  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

model $y = ax + b$

Sample (i)	Height (x_i)	Weight (y_i)	Prediction $(\bar{y}_i)$	Error $(\bar{y}_i - y_i)$
1	170	65	a × 170 + b	a × 170 + b − 65
2	155	50	a × 155 + b	a × 155 + b − 50
3	150	45	a × 150 + b	a × 150 + b − 45
4	175	70	a × 175 + b	a × 175 + b − 70
5	165	55	a × 165 + b	a × 165 + b − 55



$$\text{Sum of error} = \sum_{i=1}^{5} |\bar{y}_i - y_i|$$

$$\text{Sum of squares} = \sum_{i=1}^{5} (\bar{y}_i - y_i)^2$$

Linear regression

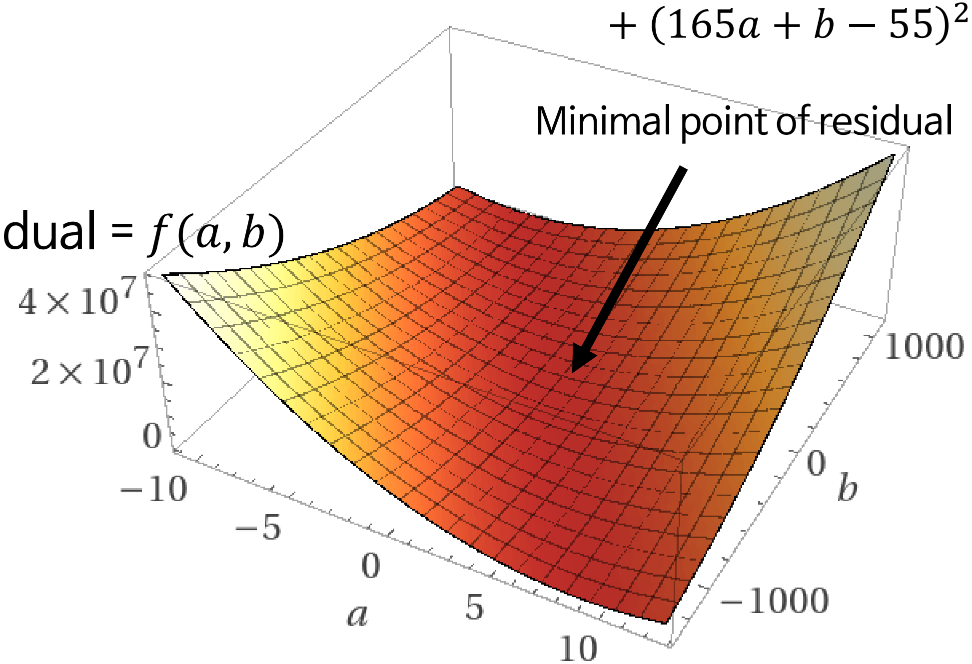
Training a linear regression model

- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

Sample (i)	Error $(\bar{y}_i - y_i)$
1	a × 170 + b − 65
2	a × 155 + b − 50
3	a × 150 + b − 45
4	a × 175 + b − 70
5	a × 165 + b − 55

$$\text{Residual} = \sum_{i=1}^{5} (\bar{y}_i - y_i)^2 = (170a + b - 65)^2 + (155a + b - 50)^2 + (150a + b - 45)^2 + (175a + b - 70)^2 + (165a + b - 55)^2$$

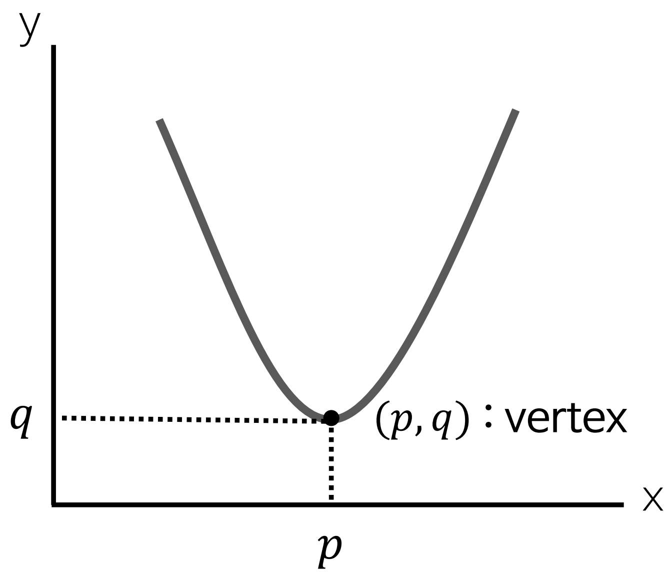
Residual = $f(a, b)$



Linear regression

Training a linear regression model

- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals



A graph of
Quadratic function

$$y = a(x-p)^2 + q, \qquad a > 0$$

$$y = ax^2 + bx + c$$

Q) If a〉0 , then which x minimizes the function?
: The point where the derivative = 0

$$\frac{dy}{dx} = \frac{d}{dx}\left(a(x-p)^2 + q\right) = 0$$

$$\frac{d}{dx}\left(a(x-p)^2 + q\right) = 2a(x-p) = 0$$

$\therefore\ x = p$ is a minimal value of the function

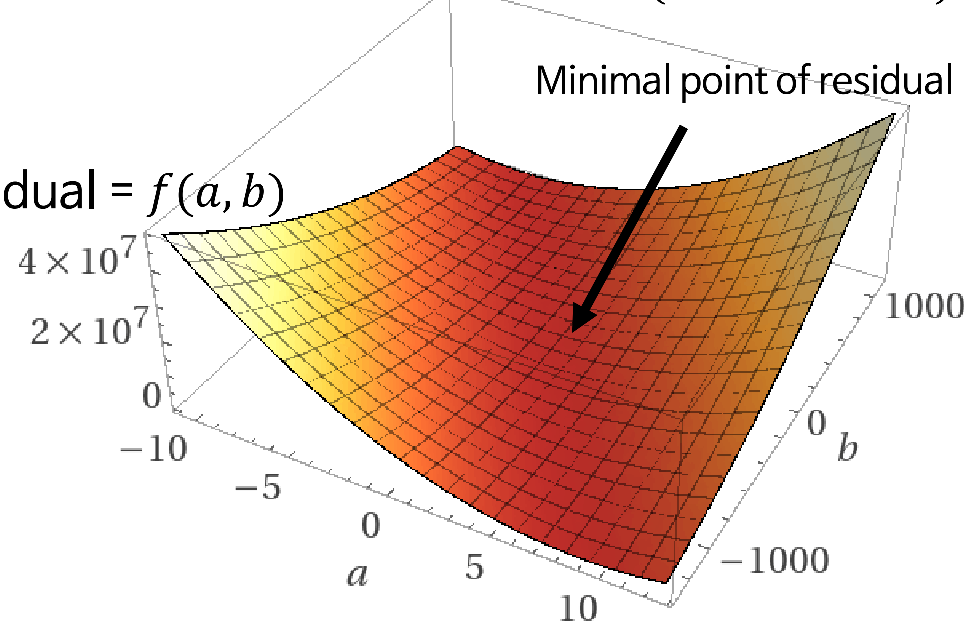
Training a linear regression model

- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

Sample (i)	Error $(\bar{y}_i - y_i)$
1	a × 170 + b − 65
2	a × 155 + b − 50
3	a × 150 + b − 45
4	a × 175 + b − 70
5	a × 165 + b − 55

$$\text{Residual} = \sum_{i=1}^{5} (\bar{y}_i - y_i)^2 = (170a + b - 65)^2 + (155a + b - 50)^2 + (150a + b - 45)^2 + (175a + b - 70)^2 + (165a + b - 55)^2$$

Residual = $f(a, b)$



Linear regression

Optimization

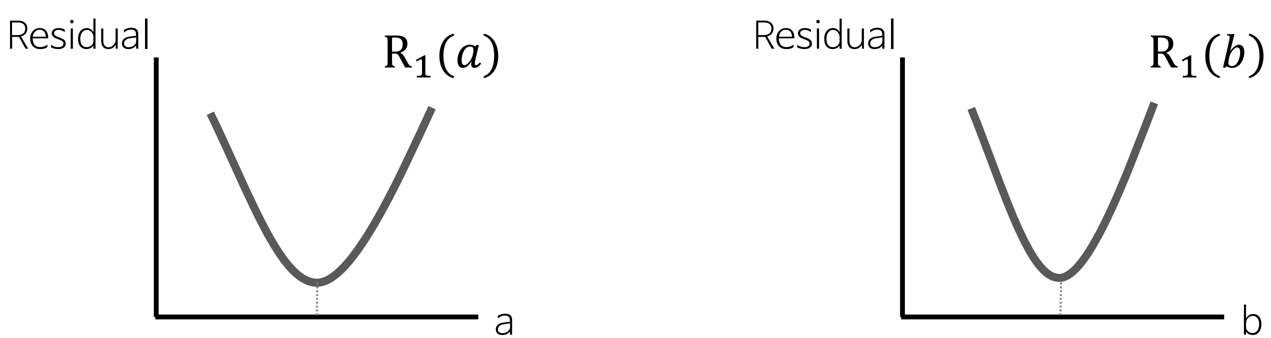
- How can we train a linear regression model?
- Least square method

  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

$$
\begin{aligned}
R = &(170a + b - 65)^2 \\
&+ (155a + b - 50)^2 \\
&+ (150a + b - 45)^2 \\
&+ (175a + b - 70)^2 \\
&+ (165a + b - 55)^2
\end{aligned}
$$

$\mathrm{R}_1(a) = 170^2a^2 + (2 \cdot 170b - 2 \cdot 170 \cdot 65)a + b^2 - 130b + 65^2$ Function of a - Eq (1)

$\mathrm{R}_1(b) = b^2 + (340a - 130)b + 170^2a^2 - 170 \cdot 130a + 65^2$ Function of b - Eq (2)



Linear regression

Optimization

- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

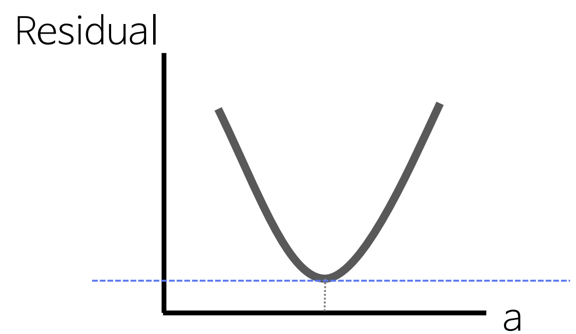
$$
\begin{aligned}
R = &(170a + b - 65)^2 \\
&+ (155a + b - 50)^2 \\
&+ (150a + b - 45)^2 \\
&+ (175a + b - 70)^2 \\
&+ (165a + b - 55)^2
\end{aligned}
$$

$$\mathrm{R}_1(a) = 170^2 a^2 + (2 \cdot 170b - 2 \cdot 170 \cdot 65)a + b^2 - 130b + 65^2$$ Eq (1)

$$\frac{\partial \mathrm{R}_1(a)}{\partial a} = 2 \cdot 170^2 a + (2 \cdot 170b - 2 \cdot 170 \cdot 65) = 0$$

Do the same process in $\mathrm{R}_2, \mathrm{R}_3, \mathrm{R}_4, \mathrm{R}_5$

$\therefore 133275a + 816b = 46875$ Eq (3)



$$\frac{\partial \mathrm{R}_1(a)}{\partial a} = 0$$

Linear regression

Optimization

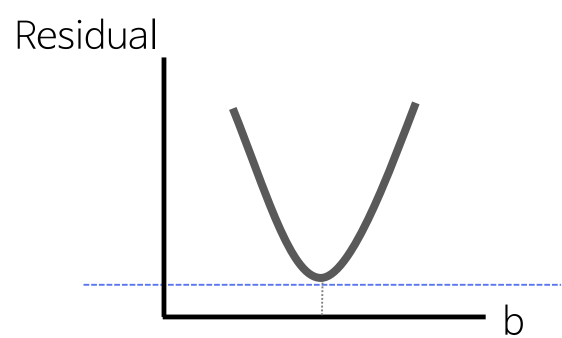
- How can we train a linear regression model?
- Least square method
  : A statistical procedure **to find the best fit for a set of data** by minimizing **the sum of the offsets** or residuals

$$
\begin{aligned}
R = &(170a + b - 65)^2 \\
&+ (155a + b - 50)^2 \\
&+ (150a + b - 45)^2 \\
&+ (175a + b - 70)^2 \\
&+ (165a + b - 55)^2
\end{aligned}
$$

$$\mathrm{R}_1(b) = b^2 + (340a - 130)b + 170^2a^2 - 170 \cdot 130a + 65^2 \quad \cdots\cdots \quad \text{Eq (2)}$$

$$\frac{\partial \mathrm{R}_1(b)}{\partial b} = 2b + (340a - 130) = 0$$

$$\therefore 815a + 5b = 285 \quad \cdots\cdots \quad \text{Eq (4)}$$



$$\frac{\partial \mathrm{R}_1(b)}{\partial b} = 0$$

b에 대한 2차 함수

Linear regression

Optimization

- Detailed solving process

$$R = (170a + b - 65)^2 + (155a + b - 50)^2 + (150a + b - 45)^2 + (175a + b - 70)^2 + (165a + b - 55)^2$$

$$\frac{\partial R}{\partial a} = 170 \cdot 2(170a \cdot b - 65) + 155 \cdot 2(155a \cdot b - 50) + 150 \cdot 2(150a \cdot b - 45) + 175 \cdot 2(175a \cdot b - 70) + 165 \cdot 2(165a \cdot b - 55) = 0$$

$$\Longrightarrow \; 170(170a \cdot b - 65) + 155(155a \cdot b - 50) + 150(150a \cdot b - 45) + 175(175a \cdot b - 70) + 165(165a \cdot b - 55) = 0$$

$$\Longrightarrow \; 133275a + 815b = 46874$$ Eq(3)

$$\frac{\partial R}{\partial b} = 2b + 2(170a - 65) + 2b + 2(155a - 50) + 2b + 2(150a - 45) + 2b + 2(175a - 70) + 2b + 2(165a - 55) = 0$$

$$\Longrightarrow \; 5b + (170a - 65) + (155a - 50) + (150a - 45) + (175a - 70) + (165a - 55) = 0$$

$$\Longrightarrow \; 815a + 5b = 285$$ Eq(4)

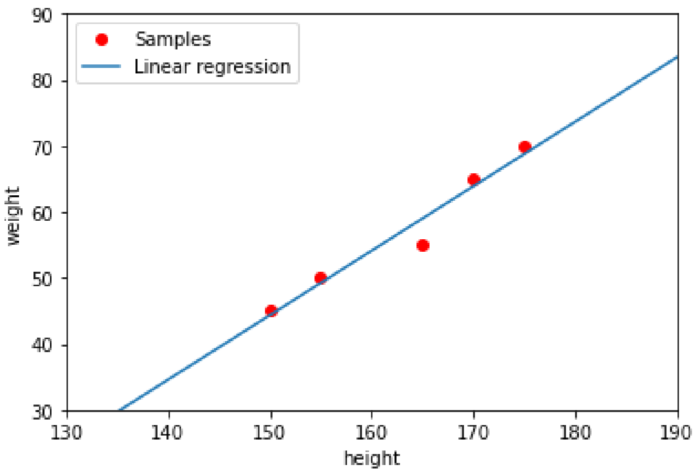
Linear regression

Optimization

- Optimization results:

a = 0.976
b = −102.209

$$133275a + 815b = 46875 \quad \text{Eq (3)}$$
$$815a + 5b = 285 \quad \text{Eq (4)}$$




```
from matplotlib import pyplot as plt
plt.plot([170,155,150,175,165],[65,50,45,70,55],'ro')
plt.plot([0,190],[-102.209,83.421])
plt.show()
```

Linear regression

Python programming for linear regression

- Numpy package
- Sympy package

Training dataset : 5 people's heights, weights

index	height (cm)	weight (kg)
A	170	65
B	155	50
C	150	45
D	175	70
E	165	55

```
In [24]: training_data = [[170, 65],
                          [155, 50],
                          [150, 45],
                          [175, 70],
                          [165, 55]]
```

```
In [25]: print(training_data)

[[170, 65], [155, 50], [150, 45], [175, 70], [165, 55]]
```

Linear regression

Python programming for linear regression

- Numpy package
- Sympy package

In [26]:
```
import sympy as sym

a = sym.Symbol('a')
b = sym.Symbol('b')

Residual = 0

for i in range(len(training_data)):

    Residual += (training_data[i][0]*a + b - training_data[i][1]) **2
```

In [27]:
```
Residual
```

Out[27]: $(150a + b - 45)^2 + (155a + b - 50)^2 + (165a + b - 55)^2 + (170a + b - 65)^2 + (175a + b - 70)^2$

$$R = \sum_{i=1}^{5} (\overline{y_i} - y_i)^2$$

Linear regression

Python programming for linear regression

- Numpy package
- Sympy package

In [28]:
```
R_Diff_a = sym.diff(Residual,a)
R_Diff_a
```

Out[28]: $266550a + 1630b - 93750$

In [29]:
```
R_Diff_b = sym.diff(Residual,b)
R_Diff_b
```

Out[29]: $1630a + 10b - 570$

Linear regression

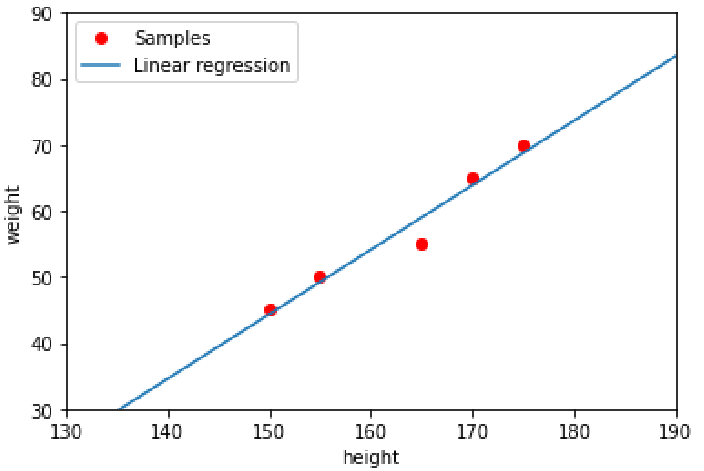
Python programming for linear regression

- Numpy package
- Sympy package

```
import numpy as np

A = [[266550, 1630],
     [1630, 10]]

B = [93750, 570]

inv_A = np.linalg.inv(A)
X = np.dot(inv_A,B)
```

```
print(X)
```

```
[   0.97674419 -102.20930233]
```

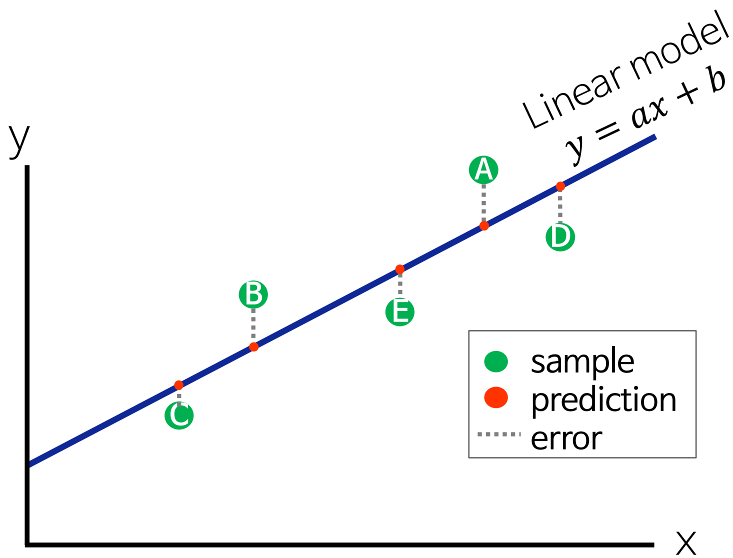



```
from matplotlib import pyplot as plt
plt.plot([170,155,150,175,165],[65,50,45,70,55],'ro')
plt.plot([0,190],[-102.209,83.421])
plt.show()
```

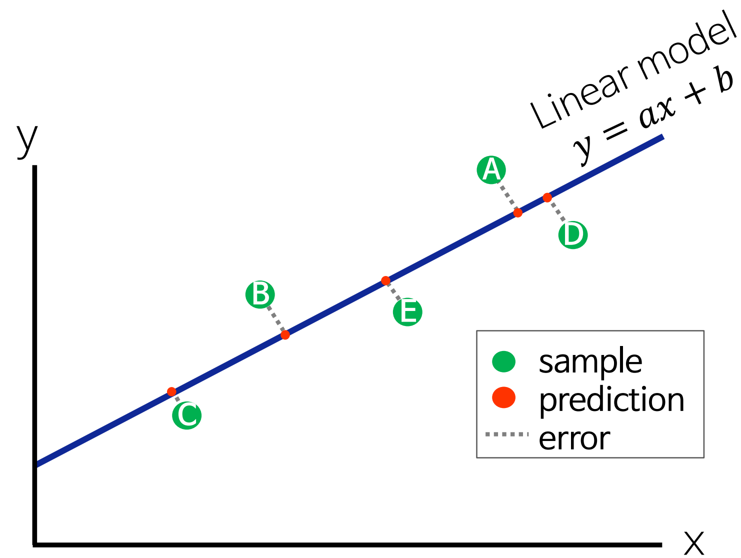
Linear regression

Discussion

- Let's improve the residual



VS.



Error : Y-axis distance between sample and predicted value

$$error = |(ax_1 + by_1 + c) - y_1|$$

Error : Euclidean distance between sample and predicted value

$$error = \frac{|ax_1 + by_1 + c|}{\sqrt{a^2 + b^2}}$$