

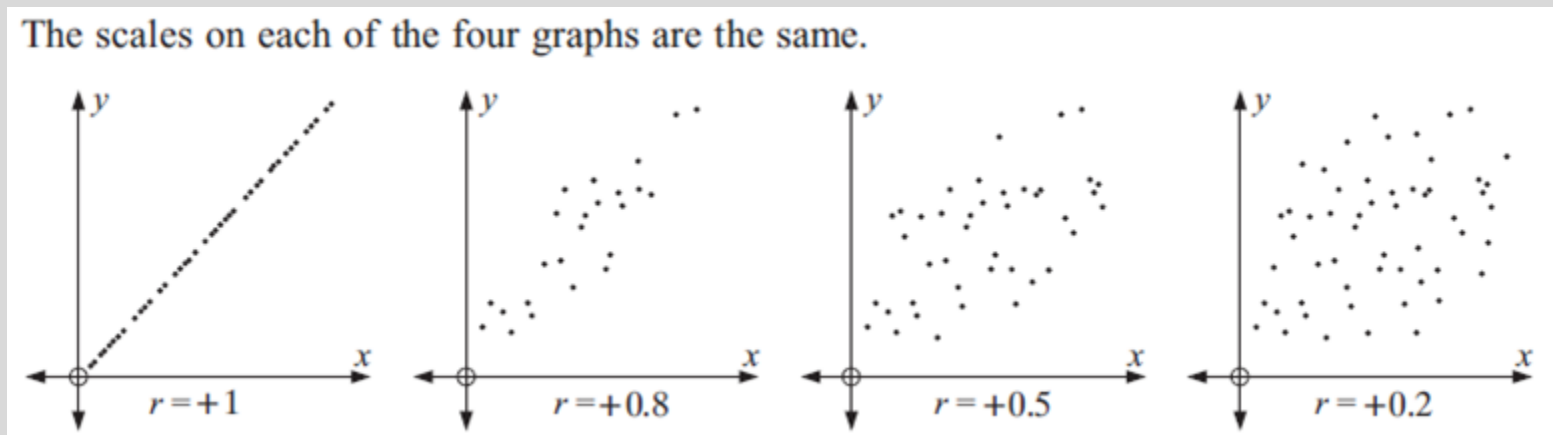


Correlation & Regression

วัดความสัมพันธ์เชิงเส้นตรงของ 2 ตัวแปร

$$\text{Cor}(x,y) = \text{Cor}(y,x)$$

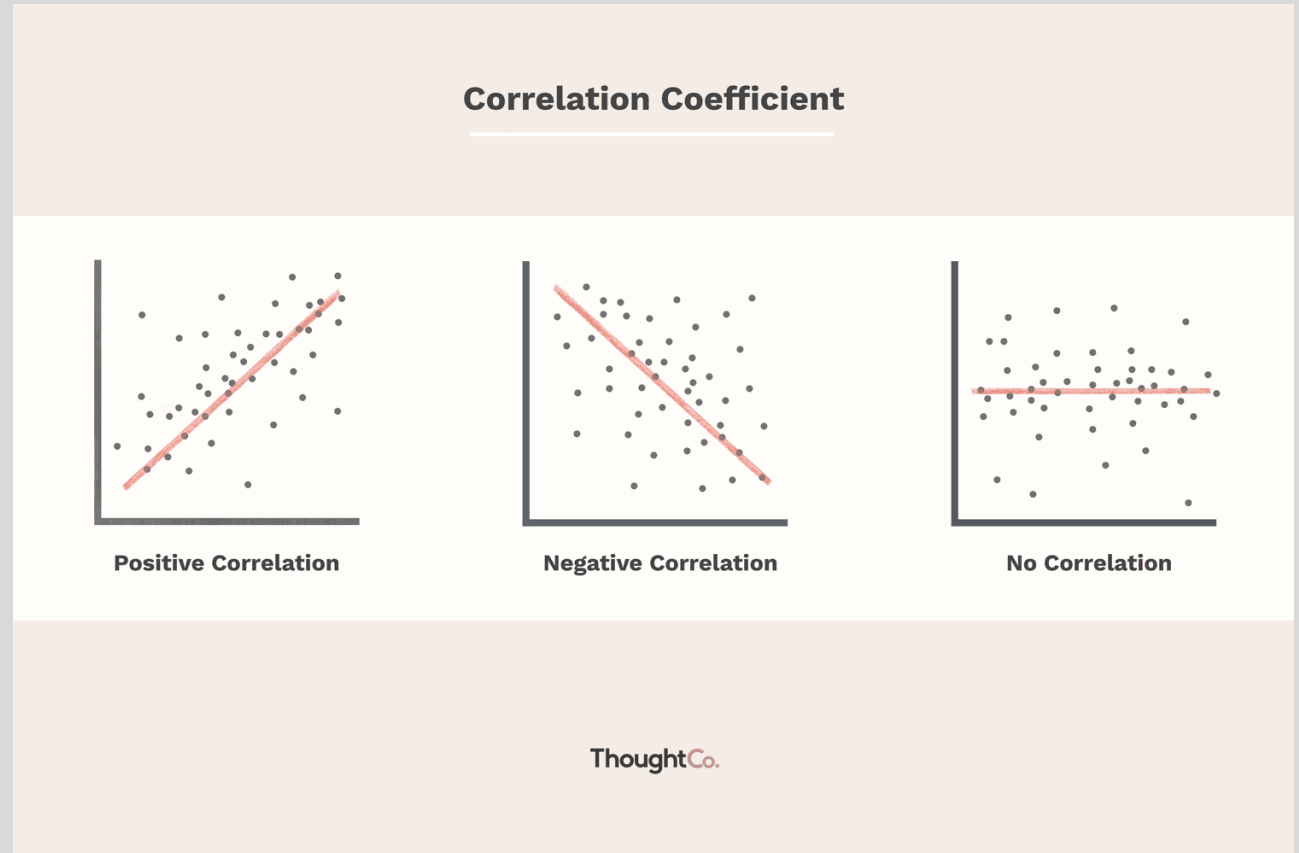
1. สัมพันธ์เชิงบวก Positive relationship
2. สัมพันธ์เชิงลบ negative relationship
3. ไม่มีความสัมพันธ์ No relationship



วัดความสัมพันธ์เชิงเส้นตรงของ 2 ตัวแปร

$$\text{Cor}(x,y) = \text{Cor}(y,x)$$

1. สัมพันธ์เชิงบวก Positive relationship
2. สัมพันธ์เชิงลบ negative relationship
3. ไม่มีความสัมพันธ์ No relationship



รูปจาก

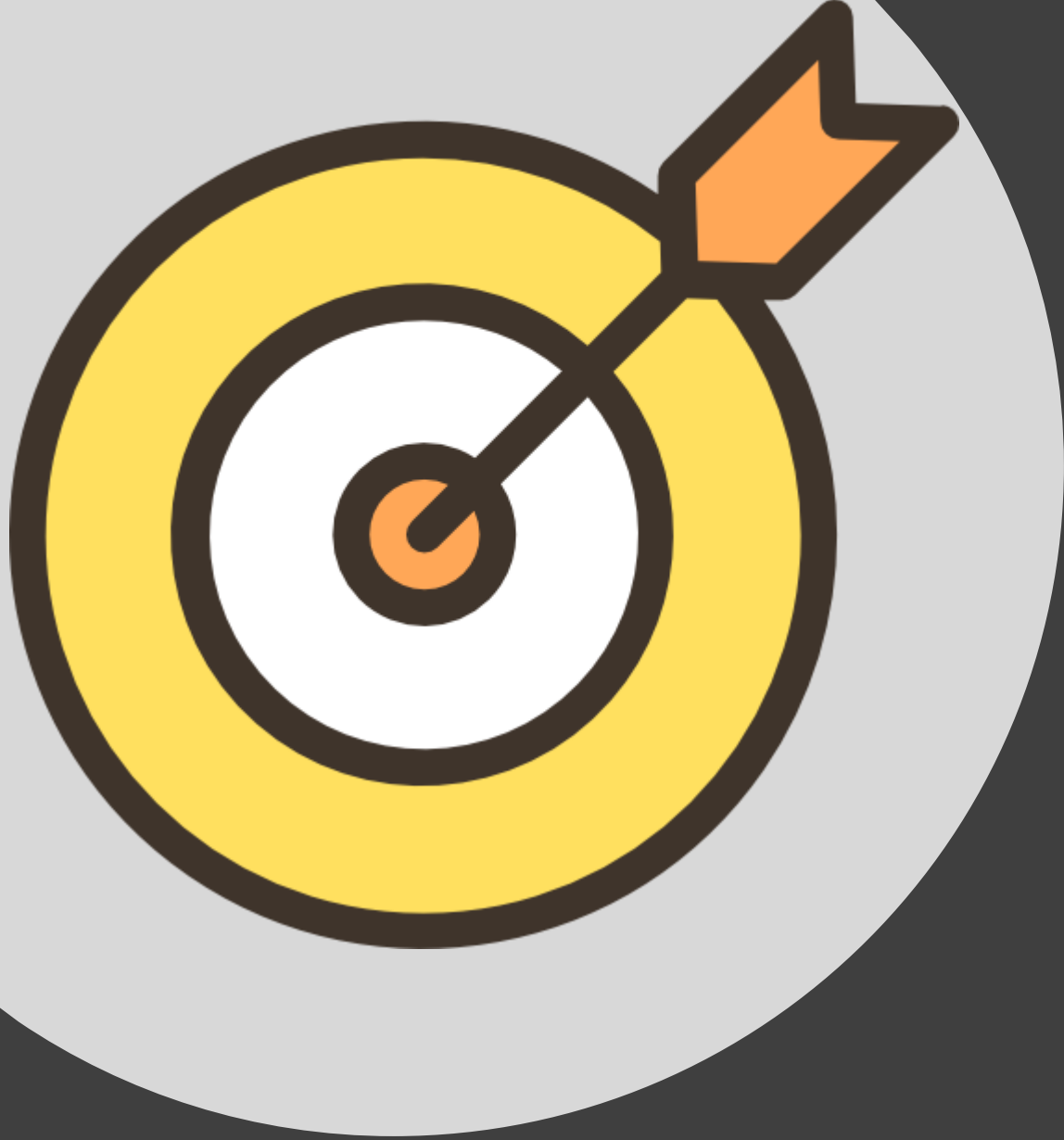
<https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp>

correlation

- ความสัมพันธ์ของทั้ง 2 ตัวแปร
- $\text{Cor}(x, y) = \text{Cor}(y, x)$
- ค่าออกมาเป็นตัวเลขเดียว
- เช่น $\text{Cor}(x, y) = 0.5$

Linear Regression

- ความสัมพันธ์ของทั้ง x ส่งผลถึง y (หรือสลับกัน)
- $\text{lm}(x, y)$ ไม่เท่า $\text{lm}(y, x)$
- ค่าออกมาเป็นสมการ
- เช่น $y = 0.1 + 0.5x$



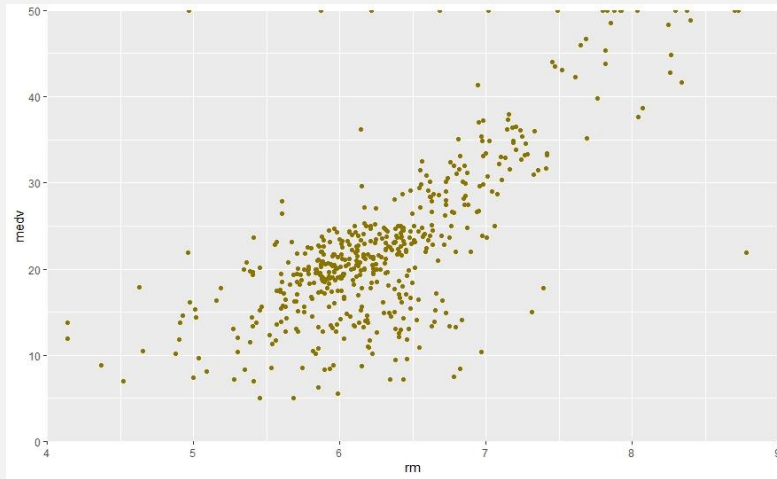
Regression Modeling Objective

Regression Modeling มีเป้าหมายหลัก 2 อย่าง

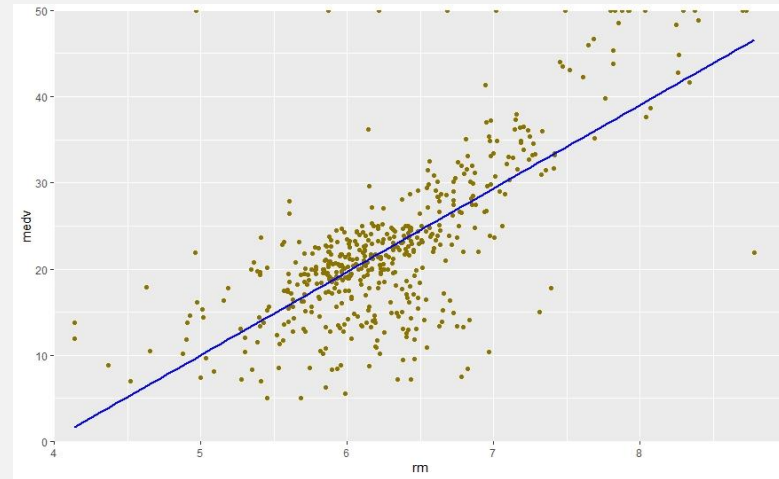
- 1. ทำนายข้อมูล (quantitative / ตัวเลข) ที่เราไม่เคยเห็นมาก่อน เช่น
 - 1.1 ทำนายยอดขายของ เพื่อที่บริษัทจะได้จัดการ Inventory ถูก
 - 1.2 ทำนายราคาวัตถุดิบ เพื่อที่จะประเมิน ต้นทุนเบื้องต้นได้
- 2. อธิบายความสัมพันธ์ทางธุรกิจ
 - 2.1 ดูว่าระหว่าง ปริมาณฝุ่น มีผลต่อยอดขายของบริษัทไหม
 - 2.2 ดูว่าระหว่าง ราคาขายโลก มีผลต่อ ราคาวัตถุดิบที่เราต้องการไหม

Linear Regression Model (Fundamental)

- สร้าง เส้นตรง ที่ fit ข้อมูลได้ดีที่สุด
- นำไป predict ข้อมูลที่ไม่เคยเห็นมาก่อน



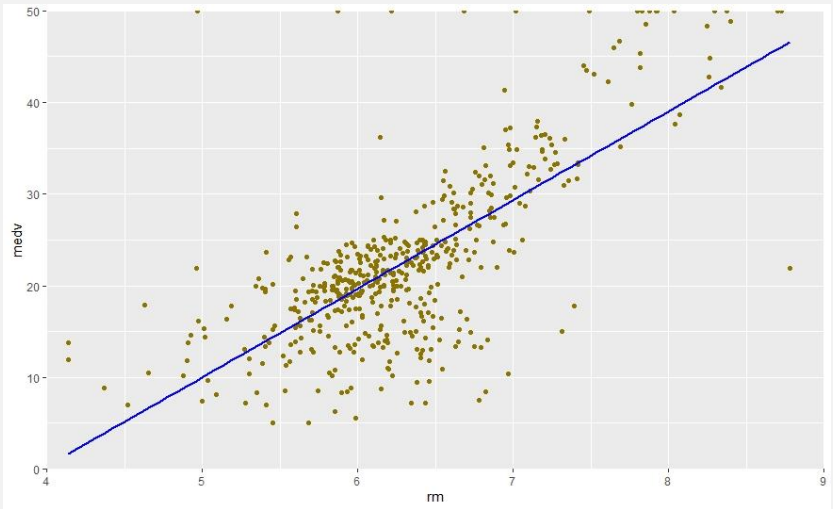
Data ข้อมูลของเรา



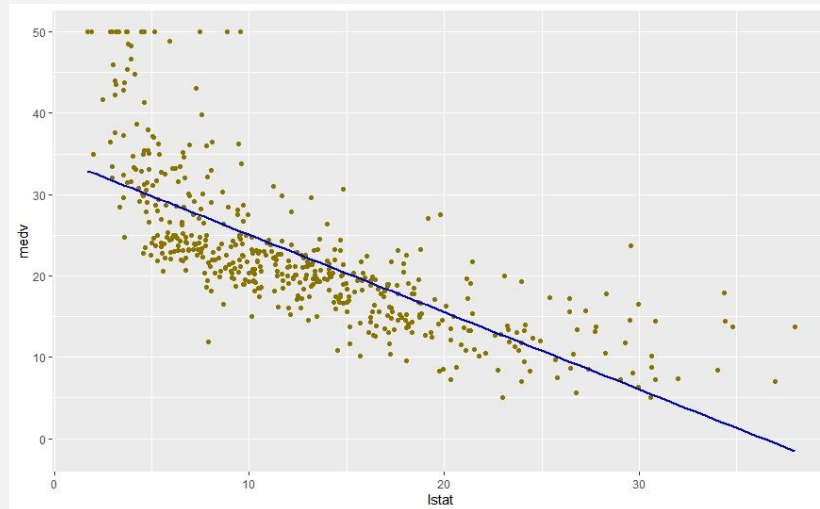
Linear Regression model

Linear Regression Model (Fundamental)

- สร้าง เส้นตรง ที่ fit ข้อมูลได้ดีที่สุด
- นำไป predict ข้อมูลที่ไม่เคยเห็นมาก่อน
- เส้นตรงของเรา มีสมมุติฐาน (assumption) มาจาก ข้อมูลของเรามีความสัมพันธ์เชิงเส้นตรง



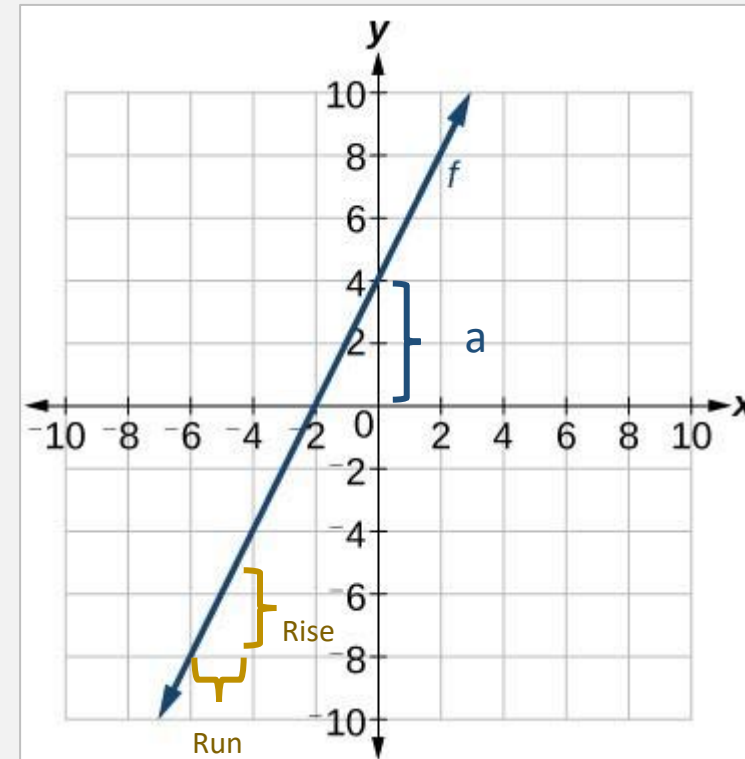
ความสัมพันธ์เชิงบวก (Positive relationship)



ความสัมพันธ์เชิงลบ (negative relationship)

Linear Regression Model (Fundamental)

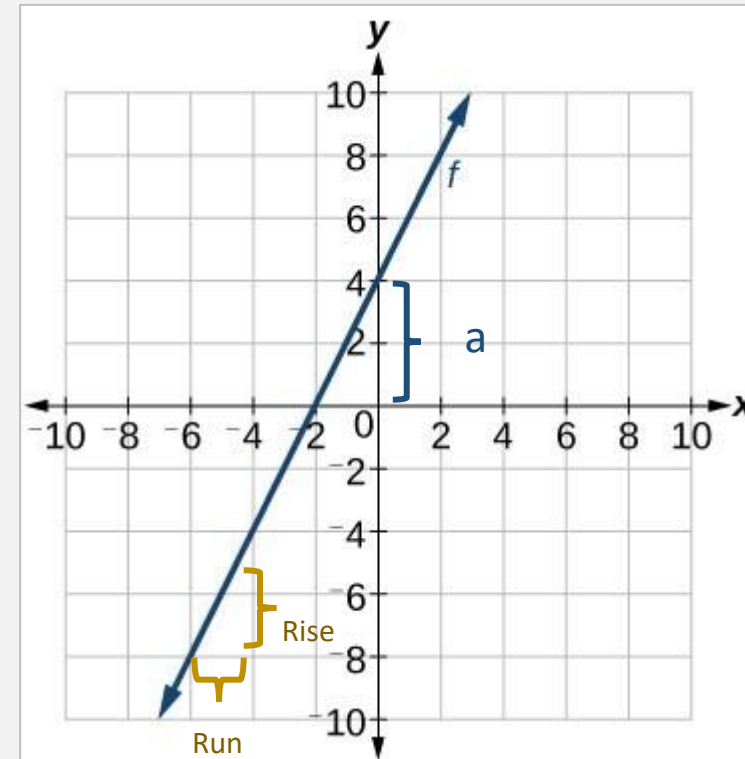
- สมการ เส้นตรงทั่วไป : $y = a + bx$
- a : intercept
- b : slope (Rise \div Run)



Linear Regression Model (Fundamental)

- สมการ เส้นตรงทั่วไป : $y = a + bx$
- a : intercept
- b : slope (Rise \div Run)

$$y = b_0 + b_1 * x_1$$



Linear regression theory (Equation)

Simple Linear
Regression

$$y = b_0 + b_1 * x_1$$



Multiple Linear
Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

Linear regression theory (Dataset Introduction)

```
library(mlbench)
data("BostonHousing")
str(BostonHousing)
```

```
'data.frame':      506 obs. of  14 variables:
 $ crim      : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn        : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ indus     : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ chas      : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ nox       : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm        : num  6.58 6.42 7.18 7 7.15 ...
 $ age       : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ dis       : num  4.09 4.97 4.97 6.06 6.06 ...
 $ rad       : num  1 2 2 3 3 3 5 5 5 5 ...
 $ tax       : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio   : num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ b         : num  397 397 393 395 397 ...
 $ lstat     : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv      : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

ราคาบ้านโดยเฉลี่ย (1,000 USD) (medv) | Y
| Respond variable | Outcome variable

จำนวนห้องโดยเฉลี่ย (rm) | X | Factor |
Features | predictor variable

Linear regression theory (Simple Linear Regression)

Simple Linear
Regression

$$y = b_0 + b_1 \cdot x_1$$



| ราคาบ้านโดยเฉลี่ย (1,000 USD) (medv) | จำนวนห้องโดยเฉลี่ย (rm) |
|---|----------------------------|
| 24 | 6.575 |
| 21.6 | 6.421 |
| 34.7 | 7.185 |
| 33.4 | 6.998 |
| 36.2 | 7.147 |
| 28.7 | 6.43 |
| 22.9 | 6.012 |
| 27.1 | 6.172 |
| 16.5 | 5.631 |
| 18.9 | 6.004 |
| 15 | 6.377 |
| 18.9 | 6.009 |
| 21.7 | 5.889 |
| 20.4 | 5.949 |
| 18.2 | 6.096 |
| | |
| | |

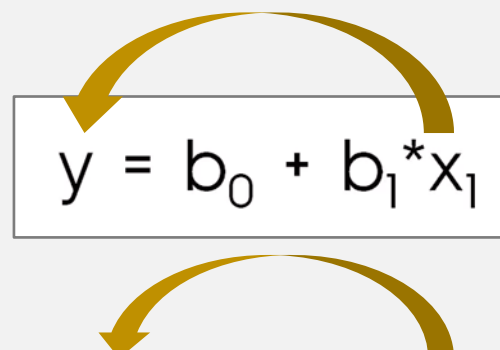
$$\text{medv} = b_0 + b_1 \cdot \text{Rm}$$

ราคาบ้านโดยเฉลี่ย (1,000 USD) (medv) | Y |
Respond variable | Outcome variable

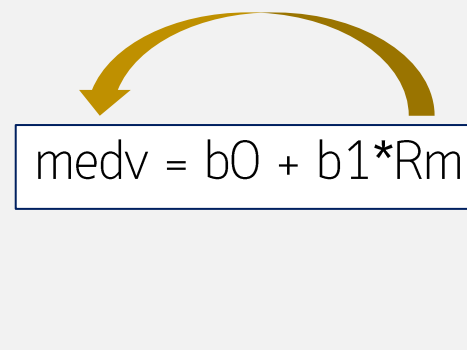
จำนวนห้องโดยเฉลี่ย (rm) | X | Factor | Features |
predictor variable

Linear regression theory (Simple Linear Regression)

Simple Linear
Regression


$$y = b_0 + b_1 * x_1$$

| ราคาบ้านโดยเฉลี่ย (1,000 USD) (medv) | จำนวนห้องโดยเฉลี่ย (rm) |
|---|----------------------------|
| 24 | 6.575 |
| 21.6 | 6.421 |
| 34.7 | 7.185 |
| 33.4 | 6.998 |
| 36.2 | 7.147 |
| 28.7 | 6.43 |
| 22.9 | 6.012 |
| 27.1 | 6.172 |
| 16.5 | 5.631 |
| 18.9 | 6.004 |
| 15 | 6.377 |
| 18.9 | 6.009 |
| 21.7 | 5.889 |
| 20.4 | 5.949 |
| 18.2 | 6.096 |
| | |
| | |


$$\text{medv} = b_0 + b_1 * \text{Rm}$$

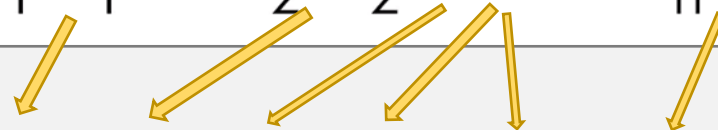
ราคาบ้านโดยเฉลี่ย (1,000 USD) (medv) | Y |
Respond variable | Outcome variable

จำนวนห้องโดยเฉลี่ย (rm) | X | Factor | Features |
predictor variable

Linear regression theory (Multiple Linear Regression)

Multiple Linear Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$



| medv | rm | age | dis | ptratio | b | lstat |
|------|-------|------|--------|---------|--------|-------|
| 24 | 6.575 | 65.2 | 4.09 | 15.3 | 396.9 | 4.98 |
| 21.6 | 6.421 | 78.9 | 4.9671 | 17.8 | 396.9 | 9.14 |
| 34.7 | 7.185 | 61.1 | 4.9671 | 17.8 | 392.83 | 4.03 |
| 33.4 | 6.998 | 45.8 | 6.0622 | 18.7 | 394.63 | 2.94 |
| 36.2 | 7.147 | 54.2 | 6.0622 | 18.7 | 396.9 | 5.33 |
| 28.7 | 6.43 | 58.7 | 6.0622 | 18.7 | 394.12 | 5.21 |
| 22.9 | 6.012 | 66.6 | 5.5605 | 15.2 | 395.6 | 12.43 |
| 27.1 | 6.172 | 96.1 | 5.9505 | 15.2 | 396.9 | 19.15 |
| 16.5 | 5.631 | 100 | 6.0821 | 15.2 | 386.63 | 29.93 |
| 18.9 | 6.004 | 85.9 | 6.5921 | 15.2 | 386.71 | 17.1 |
| 15 | 6.377 | 94.3 | 6.3467 | 15.2 | 392.52 | 20.45 |
| 18.9 | 6.009 | 82.9 | 6.2267 | 15.2 | 396.9 | 13.27 |
| 21.7 | 5.889 | 39 | 5.4509 | 15.2 | 390.5 | 15.71 |
| 20.4 | 5.949 | 61.8 | 4.7075 | 21 | 396.9 | 8.26 |
| 18.2 | 6.096 | 84.5 | 4.4619 | 21 | 380.02 | 10.26 |
| ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... |

$$\text{medv} = b_0 + b_1 * \text{Rm} + b_2 * \text{age} + b_3 * \text{dis} + b_4 * \text{ptratio} + b_5 * b + b_6 * \text{lstat}$$

ราคาบ้านโดยเฉลี่ย (1,000 USD) | Y | Respond variable | Outcome variable

(rm , age, dis , ptratio , b , lstat) | X | Factor | Features | predictor variable

Linear regression theory (Linear Regression R code)

```
linear_model <- lm(medv ~ rm , data = BostonHousing)
summary(linear_model)
```

Call:

```
lm(formula = medv ~ rm, data = BostonHousing)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|--------|
| -23.346 | -2.547 | 0.090 | 2.986 | 39.433 |

Coefficients:

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|----------|------------|---------|------------|
| (Intercept) | -34.671 | 2.650 | -13.08 | <2e-16 *** |
| rm | 9.102 | 0.419 | 21.72 | <2e-16 *** |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.616 on 504 degrees of freedom

Multiple R-squared: 0.4835, Adjusted R-squared: 0.4825

F-statistic: 471.8 on 1 and 504 DF, p-value: < 2.2e-16

Column (Y) ~ Column (X)

$$Y = b_0 + b_1 X_1$$

$$\text{Medv} = -34.671 + 9.102 \cdot \text{rm}$$