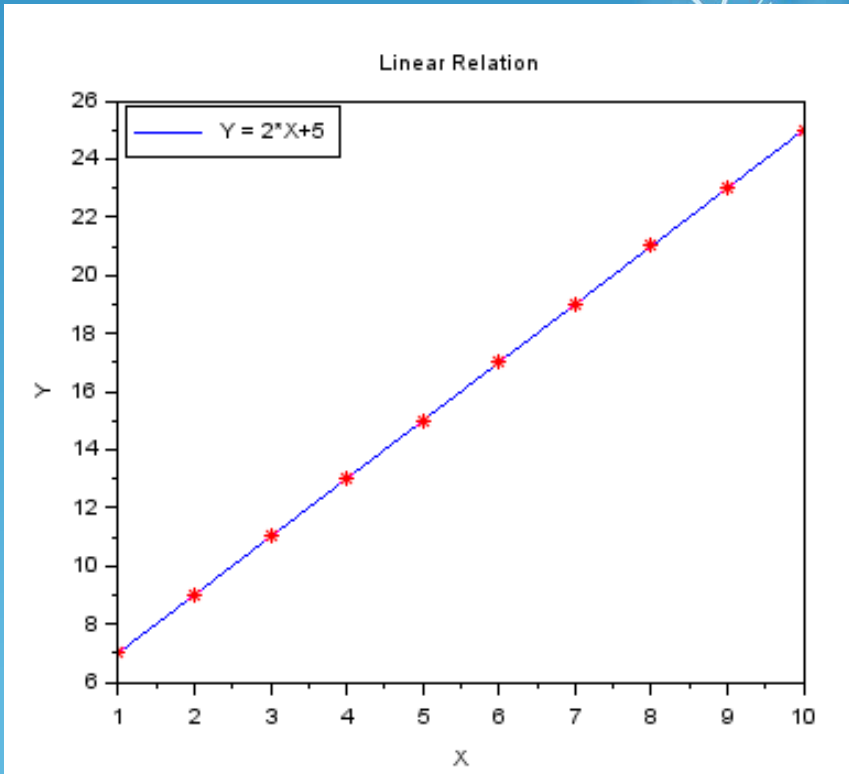
Linear Regression

ภาควิชาวิทยาการคอมพิวเตอร์ มหาวิทยาลัยเชียงใหม่

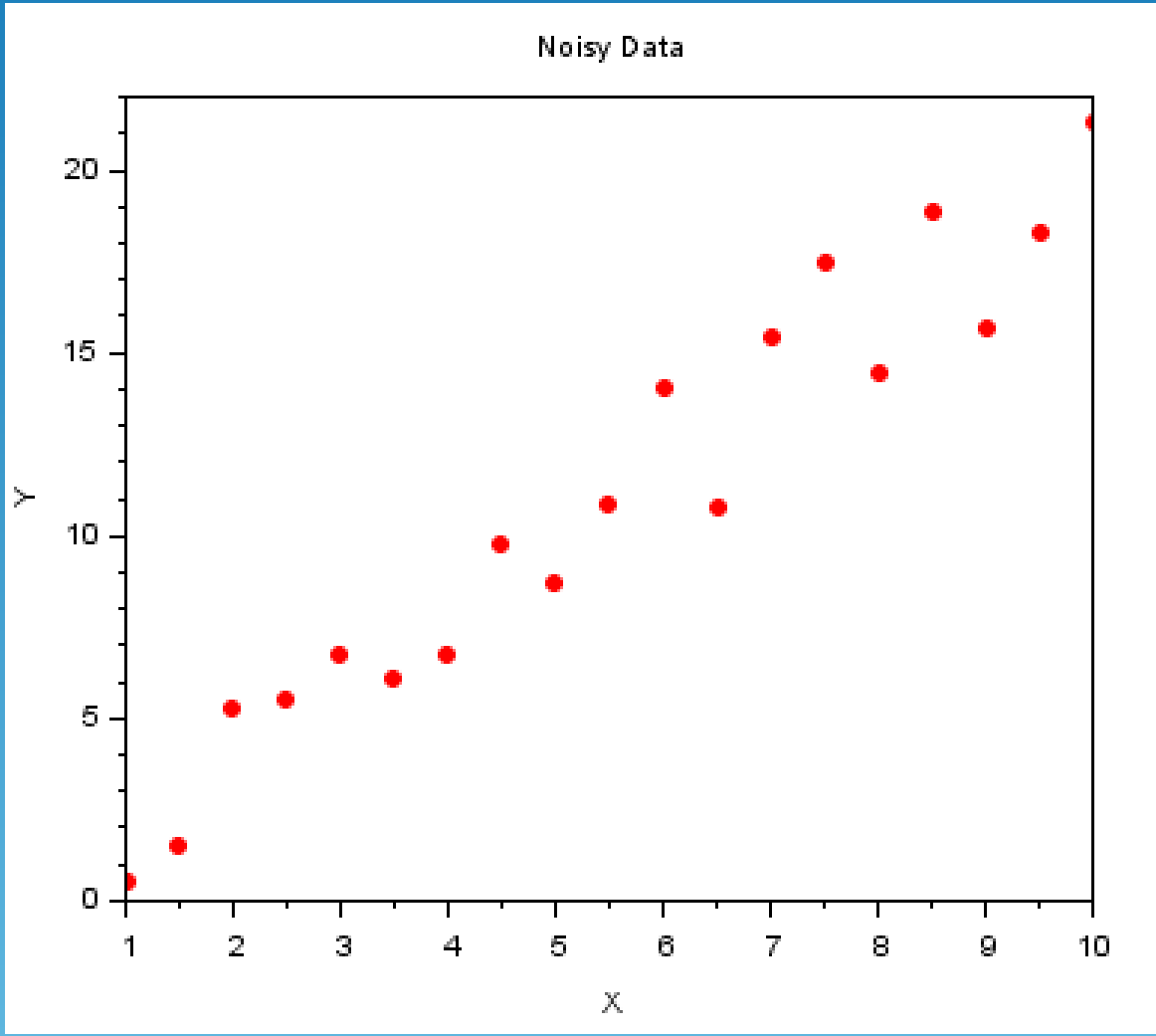
การหาความสัมพันธ์ระหว่างสองตัวแปร

- บางที เราต้องการจะหาความสัมพันธ์ระหว่าง ตัวแปรหรือชุดข้อมูล 2 ชุด
  - จำนวนชั่วโมงที่ใช้ท่องหนังสือกับคะแนนสอบ
  - ปริมาณแอลกอฮอล์ที่บริโภคกับอายุขัย (ปี)
  - การศึกษา (ปี) กับรายได้



Noisy Data

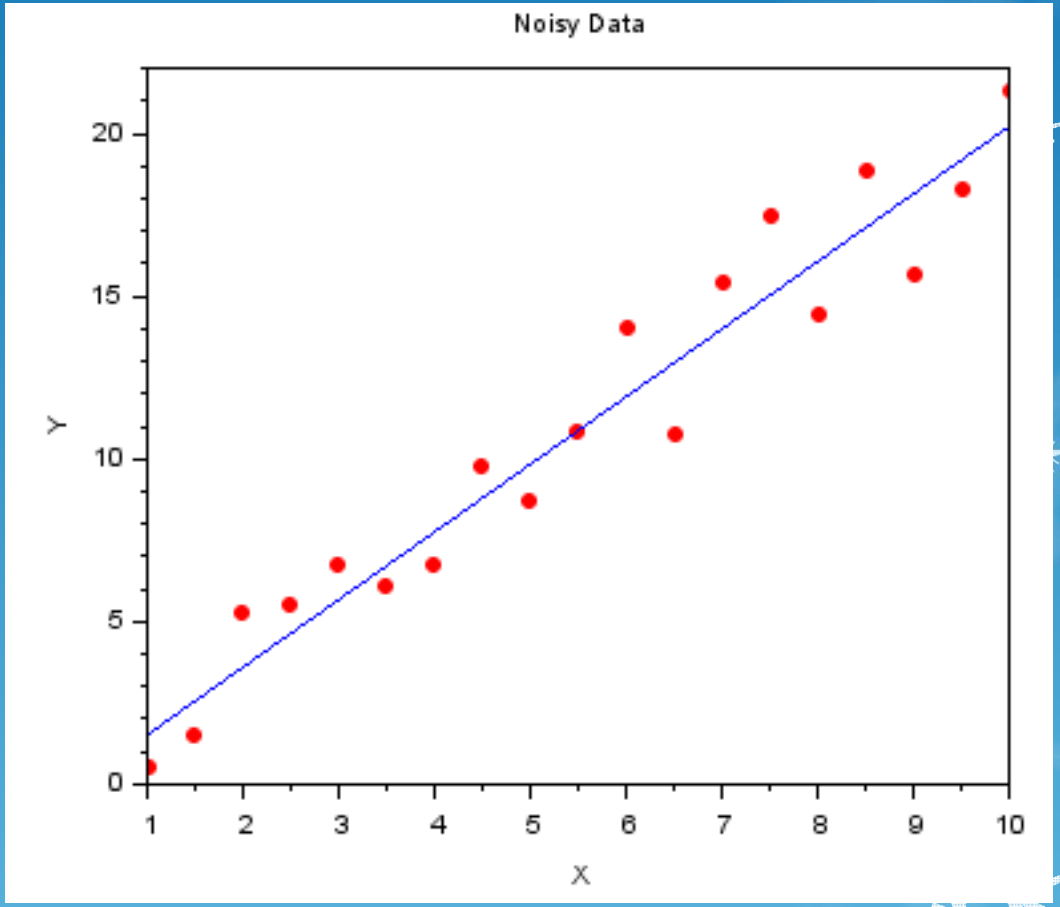
- แต่ว่า ข้อมูลจริงมักจะไม่มีความสัมพันธ์ชัดเจน



Linear Regression

- จึงจะต้องหาความสัมพันธ์ที่ใกล้เคียงที่สุด
- เราจะมาใช้เทคนิคชื่อ linear regression ซึ่งจะหาความสัมพันธ์ในรูปของสมการเชิงเส้น

$$y = ax + b$$



reglin() Function ใน SciLab

- ฟังก์ชัน reglin() ใช้ในการแก้ linear regression
- รูปแบบการเรียกใช้ reglin()

```
[a b sig] = reglin(X, Y)
```

Input:

- X คือ vector ที่มีค่า x_1 ถึง x_n อยู่
- Y คือ vector ที่มีค่า y_1 ถึง y_n อยู่

Output:

- a และ b เป็นค่าสัมประสิทธิ์ ในสมการ $y = ax + b$
- sig คือ ค่าเบี่ยงเบนมาตรฐานของค่าความต่างของแต่ละค่าใน vector กับค่าเฉลี่ย

- ลำดับขั้นตอนการใช้ reglin()
  1. กำหนด vector สำหรับค่า x
  2. กำหนด vector สำหรับค่า y
  3. เรียก reglin()

การใช้ค่า a, b ที่ได้จาก reglin()

- ประมาณค่าจากสมการเชิงเส้นถดถอย
  - มีค่า *x* เข้ามาใหม่
  - ให้ *z* เป็นการประมาณค่า *y* จากสมการที่ได้จาก linear regression:
  - $z = a \cdot x + b$
- พลอตกราฟเส้นตรงจากสมการเพื่อเปรียบเทียบกับข้อมูล
  - จำเป็นที่จะต้องมีค่า $z = a \cdot x + b$ เพียงสองจุดคือ
  - $z_1 = a \cdot x_1 + b$ และ $z_n = a \cdot x_n + b$
  - เมื่อได้ z_1 กับ z_n ก็สามารถวาดกราฟเส้นตรงจาก (x_1, z_1) ไปยัง (x_n, z_n) ได้

การหาค่า *r*

- ค่า *r* คือสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ของ *X* และ *Y*
  - มีค่าตั้งแต่ -1 ถึง 1
    - r = 0 แสดงว่าไม่มีความสัมพันธ์เชิงเส้นระหว่าง *X* กับ *Y*
    - r < 0 แสดงว่ามีความสัมพันธ์เชิงเส้นในทางลบระหว่าง *X* กับ *Y*
      - เมื่อ X ลดลง y จะเพิ่มขึ้น
      - เมื่อ X เพิ่มขึ้น y จะลดลง
    - r > 0 แสดงว่ามีความสัมพันธ์เชิงเส้นในทางบวกระหว่าง *X* กับ *Y*
      - เมื่อ X เพิ่มขึ้น y จะเพิ่มขึ้นตาม
      - เมื่อ X ลดลง y จะลดลง ตาม

$$r = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

ตัวอย่าง (1)

- **จากแบบฝึกหัดที่ 5.7 หน้า 234**
- **“จากข้อมูลอายุต้นไม้และความสูงของต้นไม้ชนิดหนึ่งจำนวน 8 ต้น ได้ข้อมูลดังนี้**

ต้นที่	1	2	3	4	5	6	7	8
อายุ (ปี)	1	2	3	4	5	6	7	8
ความสูง (ฟุต)	3.6	3.7	7.8	7.6	10.3	14.0	14.6	18.1

จากข้อมูลนี้จงสร้างสมการเพื่อประมาณความสูงของต้นไม้จากอายุ และ ประมาณความสูงของต้นไม้ ถ้าต้นไม้มีอายุ 9 ปี”

ตัวอย่าง (2) — การหา Linear Regression ด้วย SciLab

1. สร้าง vector สำหรับค่า x และ y

```
X = 1:8;
Y = [3.6 3.7 7.8 7.6 10.3 14.0 14.6 18.1];
```

2. ใช้ reglin() เพื่อหาค่า a และ b

```
[a b sig] = reglin(X, Y);
```

ตัวอย่าง (3) — การหาสมการด้วย SciLab

3. วาดกราฟเปรียบเทียบ

```
plot(X, Y, '.r');
[m n] = size(X);
X_lin = [X(1) X(n)];
Z = a*X_lin + b;
plot(X_lin, Z, '-b');
```

4. ประมาณค่าเมื่อ x = 9

```
x = 9
z =a*x + b;
```

ตัวอย่าง (4) — การหาค่า r

3. หาค่า r

```
mean_X = mean(X);
mean_Y = mean(Y);
sum_diff_XY = sum((X-mean_X).*(Y-mean_Y));
sum_diff_Xsq = sum((X-mean_X).^2);
sum_diff_Ysq = sum((Y-mean_Y).^2);
r = sum_diff_XY/(sqroot(sum_diff_Xsq*sum_diff_Ysq));
```

$\leftarrow \bar{X}$

$\leftarrow \bar{Y}$

$\Leftarrow \sum(X-\bar{X})(Y-\bar{Y})$

$\leftarrow \sum(X-\bar{X})^2$

$\leftarrow \sum(Y-\bar{Y})^2$

$$\Leftarrow r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2 \sum(Y-\bar{Y})^2}}$$