การวิเคราะห์ข้อมูลเบื้องต้น

01418131

ดร. ธรรมกร แซ่ตั้ง

Core Processes in Data Science

- Statistical Analyses
- Data Visualization
- Data Summarization

- Cross-Validation
- Parameter Tuning
- Feature Selection
- Refining Algorithm

Understanding Problem → Data Acquisition → Data Preparation → Exploratory Data Analysis (EDA) → Model Building → Model Evaluation & Tuning → Model Deployment

- Data Cleaning
- Data Organizing
- Data Transformation

- Algorithm Selection
- Identify Data Pattern
- Model Training

- Designing Pipelines
- Updating Model with new data

การวิเคราะห์ข้อมูลเบื้องต้น

ในการที่จะบ่งบอกถึงลักษณะของข้อมูลได้นั้นต้องใช้การวัด (measure)

การวัดเบื้องต้นที่นิยมใช้มี 3 แบบคือ

I. **การวัดแนวโน้มเข้าสู่ส่วนกลาง (measures of central tendency)**

II. **การวัดตำแหน่ง (measure of location)**

III. **การวัดการกระจาย (measure of dispersion)**

I. การวัดแนวโน้มเข้าสู่ส่วนกลาง

คือการคำนวณค่ากลางของข้อมูลหรือจุดกึ่งกลางของข้อมูล เพื่อมองภาพรวมของข้อมูลว่ามีลักษณะเป็นอย่างไร

ค่าที่ใช้วัดแนวโน้มเข้าสู่ส่วนกลาง ได้แก่

1. **ค่าเฉลี่ยเลขคณิต (arithmetic mean)**
2. **มัธยฐาน (median)**
3. **ฐานนิยม (mode)**

1. ค่าเฉลี่ยเลขคณิต (arithmetic mean)

- มักถูกเรียกสั้นๆว่า ค่าเฉลี่ย (mean)
- เป็นค่าสถิติที่นิยมใช้มากที่สุด และจะใช้ได้ดีเมื่อข้อมูลมีการกระจายอย่างสม่ำเสมอหรือกระจายไม่มากนัก
- สำหรับค่าเฉลี่ยเลขคณิตของ ***ประชากร*** จะแทนด้วย μ ซึ่งโดยส่วนใหญ่เราจะไม่ทราบค่านี้
  ค่าเฉลี่ยเลขคณิตส่วนใหญ่จะคำนวณมาจาก ***ตัวอย่าง*** ซึ่งจะแทนด้วย $\bar{x}$

สูตรการคำนวณ:

สำหรับข้อมูลที่ไม่ได้มีการจัดหมวดหมู่ (ungrouped data)

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ เมื่อ n คือจำนวนตัวอย่างทั้งหมด

สำหรับข้อมูลที่มีการจัดหมวดหมู่ (grouped data) เช่น ตารางความถี่

$\bar{x} = \frac{\sum_{i=1}^{g} f_i x_i}{n}$ เมื่อ g คือ จำนวนหมวดหมู่,

n คือจำนวนตัวอย่างทั้งหมด,

f คือจำนวนตัวอย่างในแต่ละหมวดหมู่,

x คือค่ากลางในแต่ละหมวดหมู่

2. มัธยฐาน (median)

สัญลักษณ์แทนค่ามัธยฐานคือ Med, M_e หรือ $\tilde{x}$

คือค่าที่อยู่ตรงกลางของข้อมูลทั้งหมดเมื่อข้อมูล *ถูกเรียงลำดับ* แล้ว

ในกรณีที่ข้อมูลมีการกระจายมาก หรือมีการกระจายที่ผิดปกติ มัธยฐานจะเป็นค่ากลางของข้อมูลที่ดีกว่าค่าเฉลี่ยเลขคณิต

การหาค่ามัธยฐานสำหรับข้อมูลที่ไม่ได้จัดหมวดหมู่

1. เรียงลำดับค่าของข้อมูลจากน้อยไปหามาก (หรือจากมากไปหาน้อยก็ได้)
2. หาตำแหน่งของมัธยฐาน $\frac{n+1}{2}$ ค่าที่ตำแหน่งนี้คือค่ามัธยฐาน

2. ฐานนิยม (mode)

สัญลักษณ์แทนค่ามัธยฐานคือ M_o

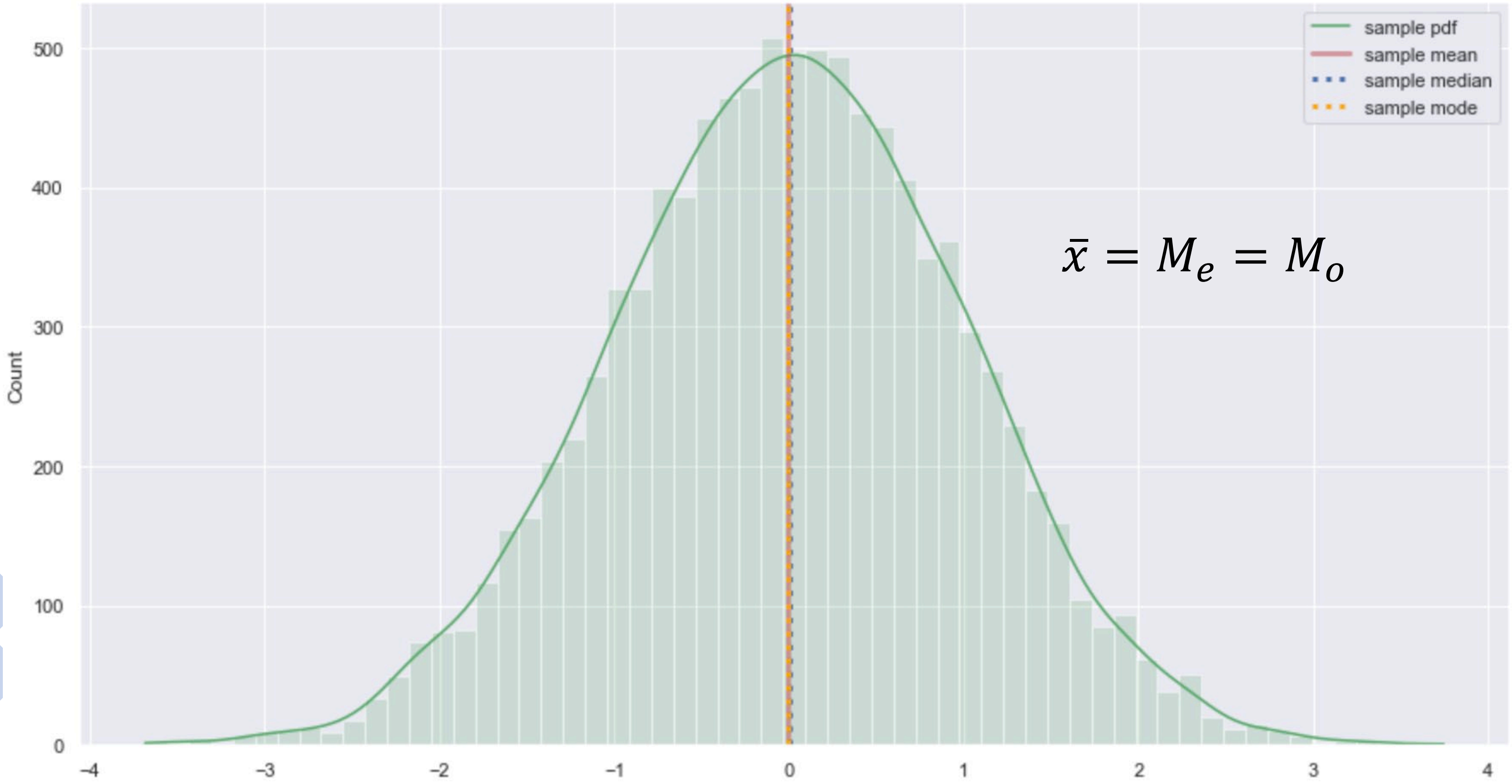
เป็นค่าของข้อมูลที่มีจำนวนซ้ำกันมากที่สุด ฐานนิยมสามารถใช้ทั้งข้อมูลเชิงปริมาณและข้อมูลเชิงคุณภาพ

การหาค่าฐานนิยมสำหรับข้อมูลที่ไม่ได้จัดหมวดหมู่

M_o = ค่าของข้อมูลที่มีจำนวนซ้ำกันมากที่สุด

ความสัมพันธ์ระหว่างค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม

1. ข้อมูลมีการแจกแจงแบบโค้งปกติ (normal distribution)
   ค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม จะมีค่าเท่ากันหรือใกล้เคียงกันมาก



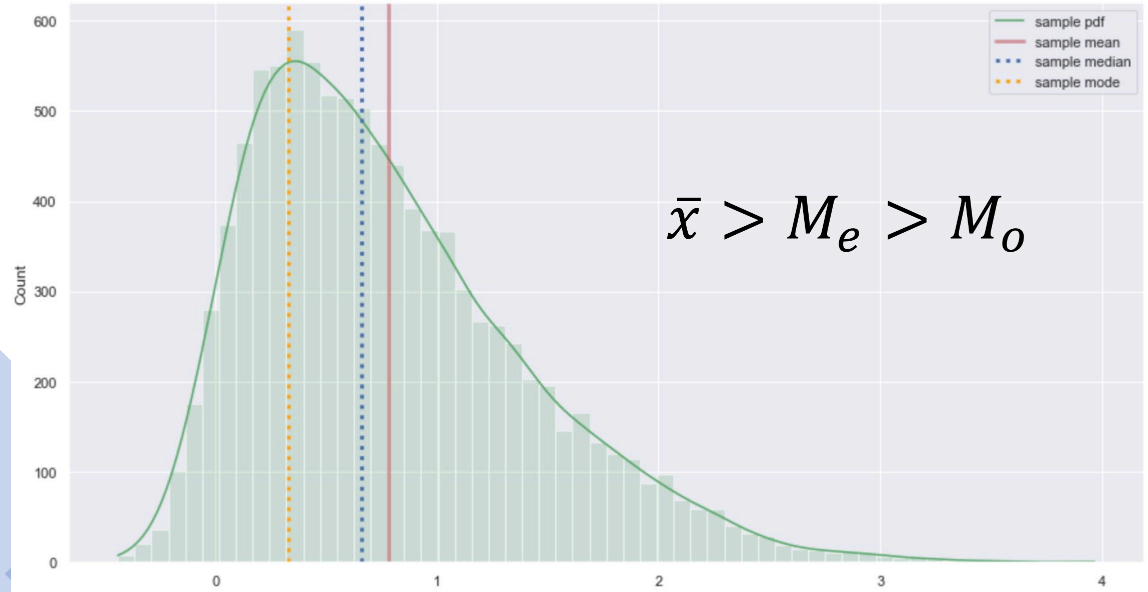
$$\bar{x} = M_e = M_o$$

see Viz_mean_median_mode.ipynb

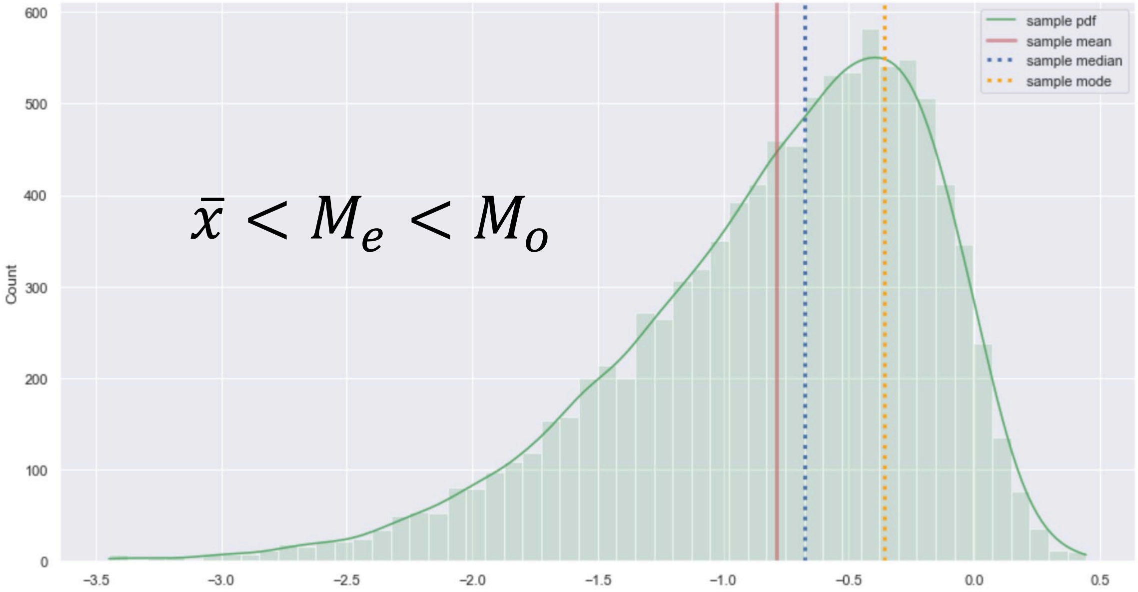
ความสัมพันธ์ระหว่างค่าเฉลี่ยเลขคณิต มัธยฐาน และฐานนิยม

2. ข้อมูลมีการแจกแจงแบบไม่ใช่โค้งปกติมีความเบ้ของข้อมูล (skewed) เช่นมีการเบ้ขวาหรือเบ้ซ้าย

แจกแจงเบ้ขวา



แจกแจงเบ้ซ้าย



II. การวัดตำแหน่งของข้อมูล (Measure of Location)

นอกจากมัธยฐานซึ่งเป็นค่าที่อยู่ ณ ตำแหน่งตรงกลางข้อมูล ค่าสถิติอื่นที่ใช้วัดตำแหน่งของข้อมูล ได้แก่ ควอร์ไทล์ (Quartiles, Q_r), เดไซล์ (Deciles, D_r) และเปอร์เซ็นไทล์ (Percentiles, P_r)

ค่าสถิติ	จำนวนการแบ่งส่วนข้อมูล	สัญลักษณ์	ช่วงของค่า r
ควอร์ไทล์	4	Q_r	1-4
เดไซล์	10	D_r	1-10
เปอร์เซ็นไทล์	100	P_r	1-100

เปรียบเทียบตำแหน่ง ควอร์ไทล์, เดไซล์ และเปอร์เซ็นไทล์

เปอร์เซ็นไทล์	P_{25}	P_{50}	P_{75}	P_{100}
เดไซล์	$D_{2.5}$	D_{5}	$D_{7.5}$	D_{10}
ควอร์ไทล์	Q_1	Q_2	Q_3	Q_4
ข้อมูลที่เรียงลำดับแล้ว แบ่งตำแหน่งเป็น 4 ส่วนอย่างละ 25%	25%	25%	25%	25%

มัธยฐาน

III. การวัดการกระจาย (Measure of Dispersion)

หากต้องการพิจารณาภาพรวมของข้อมูลว่ามีความแตกต่างมากน้อยแค่ไหน ค่าสถิติที่นิยมใช้วัดคือ

1. **พิสัย (range)**
2. **ส่วนเบี่ยงเบนควอร์ไทล์ (quartiles deviation)**
3. **ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)**
4. **ความแปรปรวน (variance)**
5. **สัมประสิทธิ์ของการแปรผัน (coefficient of variation)**

1. พิสัย (range)

คือการวัดการกระจายของข้อมูลแบบคร่าว

สำหรับข้อมูลที่ไม่ได้จัดหมวดหมู่

พิสัย = ข้อมูลที่มีค่าสูงสุด (max) - ข้อมูลที่มีค่าต่ำสุด (min)

สำหรับข้อมูลที่จัดหมวดหมู่

พิสัย = ขีดจำกัดบนที่แท้จริงของชั้นที่ข้อมูลมีค่าสูงสุด - ขีดจำกัดล่างที่แท้จริงของชั้นที่ข้อมูลมีค่าน้อยสุด

2. ส่วนเบี่ยงเบนควอร์ไทล์ (quartiles deviation, Q.D.)

จะใช้ค่า Q_1 และ Q_3 ในการคำนวณ

$$Q.D. = \frac{Q_3 - Q_1}{2}$$

3. ส่วนเบี่ยงเบนมาตรฐาน (standard deviation)

เป็นการวัดการกระจายที่นิยมใช้ เป็นการเปรียบเทียบว่าค่าต่างๆ ในชุดข้อมูลกระจายตัวออกไปมากน้อยเท่าใด หากข้อมูลส่วนใหญ่อยู่ใกล้ค่าเฉลี่ยมาก ค่าส่วนเบี่ยงเบนมาตรฐานก็จะมีค่าน้อย ในทางกลับกัน ถ้าข้อมูลแต่ละจุดอยู่ห่างไกลจากค่าเฉลี่ยเป็นส่วนมาก ค่าส่วนเบี่ยงเบนมาตรฐานก็จะมีค่ามาก

สัญลักษณ์ของส่วนเบี่ยงเบนมาตรฐาน

s ส่วนเบี่ยงเบนมาตราฐานของตัวอย่าง

σ ส่วนเบี่ยงเบนมาตรฐานของประชากร (โดยส่วนมากจะไม่ทราบ)

สูตรคำนวณ

สำหรับข้อมูลที่ไม่ได้จัดเป็นหมวดหมู่

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \text{ หรือ } s = \sqrt{\frac{\sum x^2 - n\bar{x}^2}{n-1}}$$

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

สำหรับข้อมูลที่จัดเป็นหมวดหมู่

$$s = \sqrt{\frac{\sum fx^2 - n\bar{x}^2}{n-1}}$$

เมื่อ n คือจำนวนตัวอย่างทั้งหมด

fx^2 คือความถี่คูณด้วยจุดกึ่งกลางกำลังสองของแต่ละชั้น

4. ความแปรปรวน (variance)

คือค่ายกกำลังสองของส่วนเบี่ยงเบนมาตรฐาน

สัญลักษณ์ของความแปรปรวน

s^2	ความแปรปรวนของตัวอย่าง
σ^2	ความแปรปรวนของประชากร (โดยส่วนมากจะไม่ทราบ)

5. สัมประสิทธิ์ของการแปรผัน (coefficient of variation, c.v.)

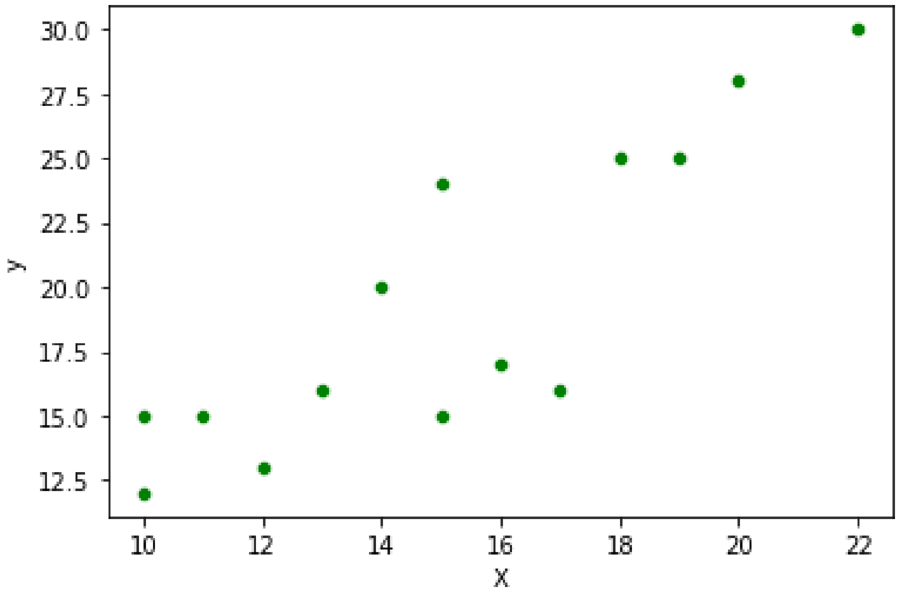
ใช้ในการเปรียบเทียบการกระจายของข้อมูลตั้งแต่ 2 ชุดขึ้นไป ซึ่งมีหน่วยวัดที่ต่างกัน ค่าสัมประสิทธิ์ของการแปรผันจะ*มีหน่วยเป็นเปอร์เซ็นต์* ข้อมูลชุดใดที่มีค่าสัมประสิทธิ์ของการแปรผันมากกว่า แสดงว่าข้อมูลชุดนั้นมีการกระจายมากกว่า

สูตรคำนวณ

$$\mathrm{c.v.} = \frac{s}{\bar{x}} \times 100$$

ความสัมพันธ์ของตัวแปรเชิงปริมาณ

- **ตัวแปรอิสระ (Independent variable) ใช้สัญลักษณ์ X**
  - สามารถทำการวิเคราะห์โดยใช้ตัวแปรอิสระเพียงตัวเดียวหรือมากกว่า 1 ตัว
  - มีอีกชื่อว่า ตัวแปรต้น, ตัวแปรเหตุ, feature data
- **ตัวแปรตาม (Dependent variable) ใช้สัญลักษณ์ y**
  - ในการวิเคราะห์มักสนใจตัวแปรตามเพียงแค่ 1 ตัว (จึงใช้ y ตัวเล็ก)
  - มีอีกชื่อว่าตัวแปรผล, ตัวแปรตาม

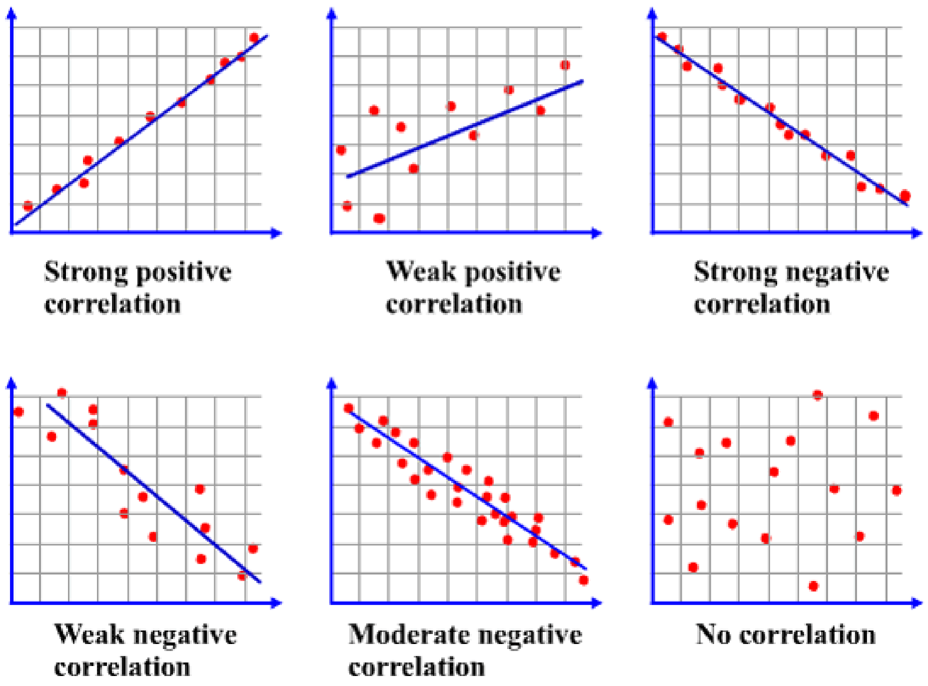


การวิเคราะห์สหสัมพันธ์และการถดถอยอย่างง่าย

- ศึกษาตัวแปร 2 ตัว (X 1 ตัว, y 1 ตัว) ว่ามีรูปแบบความสัมพันธ์อย่างไร ทิศทางใด และมี ขนาดมากน้อยเพียงใด
- ศึกษาอิทธิพลของปัจจัยต่าง ๆ (X ทีละตัว) ต่อผลที่เกิดขึ้น (y)
- สามารถทำนายว่าปริมาณของตัวแปรตาม (y) มีปริมาณเท่าใด ถ้าทราบค่าของปริมาณของตัวแปรอิสระ (X) โดยพยายามให้ค่าที่ประมาณหรือค่าที่พยากรณ์ได้มีความคาดเคลื่อนน้อย หรือมีค่าใกล้เคียงกับความเป็นจริงมากที่สุด

แผนภาพการกระจาย (Scatter Plot)

- แผนภาพการกระจาย (Scatter Plot) เป็นการนำข้อมูลตัวอย่างมาสร้างกราฟเพื่อแสดงดูรูปแบบความสัมพันธ์ของข้อมูล โดยเป็นกราฟของ 2 ตัวแปร คือ ตัวแปรอิสระ X และตัวแปรตาม y
- ใช้ดูลักษณะความสัมพันธ์ของทั้ง X และ y ว่า*มีความสัมพันธ์ในเชิงเส้นตรงหรือไม่* ในการวิเคราะห์สหสัมพันธ์และการถดถอยอย่างง่าย
- ลักษณะของ scatter plot เมื่อ X และ y มีความสัมพันธ์ของข้อมูลแบบต่างๆ:



x-axis: ตัวแปรอิสระ X

y-axis: ตัวแปรตาม **y**

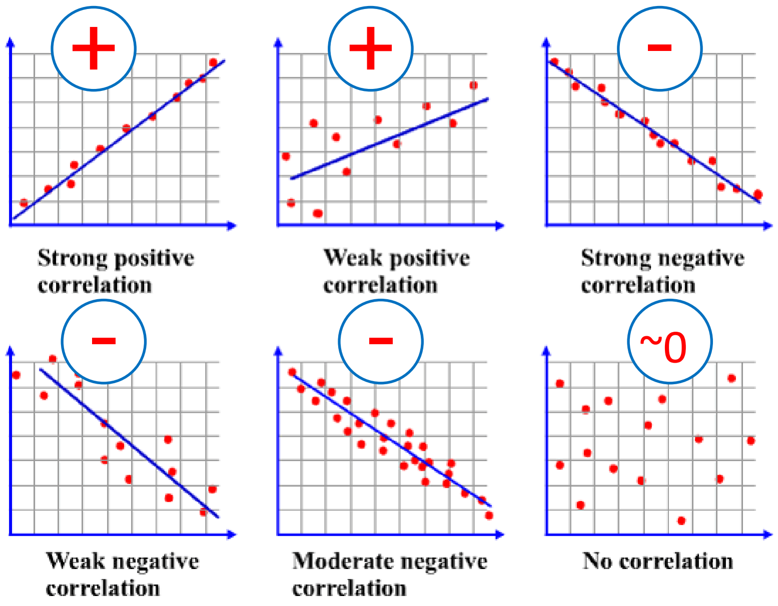
สัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Correlation Coefficient)

- เป็นการวัดว่าความสัมพันธ์ของ x และ y มีขนาดและทิศทางอย่างไร
- สำหรับ correlation coefficient ของประชากรจะใช้สัญลักษณ์ ρ เมื่อ $-1 \leq \rho \leq 1$
- สำหรับ correlation coefficient ของตัวอย่างจะใช้สัญลักษณ์ r เมื่อ $-1 \leq r \leq 1$ (โดยส่วนมากแล้วค่า correlation coefficient คำนวณจากตัวอย่าง)

<u>การคำนวณ:</u>

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

ค่า r ที่ได้



ใช้ corr() จาก pandas

หาก df คือ pandas dataframe จะสามารถหาสัมประสิทธิ์สหสัมพันธ์ในแนวคอลัมน์ โดยใช้ .corr()

```
df.corr()
```

ใช้ pearsonr() หรือ spearmanr() จาก scipy

X และ y เป็นได้ทั้ง list หรือ np.array ผลลัพธ์จะให้ค่าสัมประสิทธิ์สหสัมพันธ์พร้อมกับค่า p-value

```
r, p_val = stats.pearsonr(x, y)
```

**pearsonr() ใช้ได้กับข้อมูลที่มีการแจกแจงแบบปกติเท่านั้น หากข้อมูลไม่ใช่การแจกแจงแบบปกติจะใช้ spearmanr()

การทดสอบสมมติฐานเกี่ยวกับสัมประสิทธิ์สหสัมพันธ์

การทดสอบสมมติฐานความสัมพันธ์คือการทดสอบว่าตัวแปร x และ y มีความสัมพันธ์เชิงเส้นหรือไม่

$\mathrm{H}_0: \rho = 0$ (x กับ y ไม่มีความสัมพันธ์เชิงเส้น)

$\mathrm{H}_1: \rho \neq 0$ (x กับ y มีความสัมพันธ์เชิงเส้น)

สถิติทดสอบ:

$$t = \frac{r}{S_r}, \qquad S_r = \sqrt{\frac{1-r^2}{n-2}}, \qquad \nu = n-2$$

ปฏิเสธ H_0 เมื่อ

$t < -\left|t_{\frac{\alpha}{2},\nu=n-2}\right|$ หรือ $t > \left|t_{\frac{\alpha}{2},\nu=n-2}\right|$

****เป็น two-tailed test**

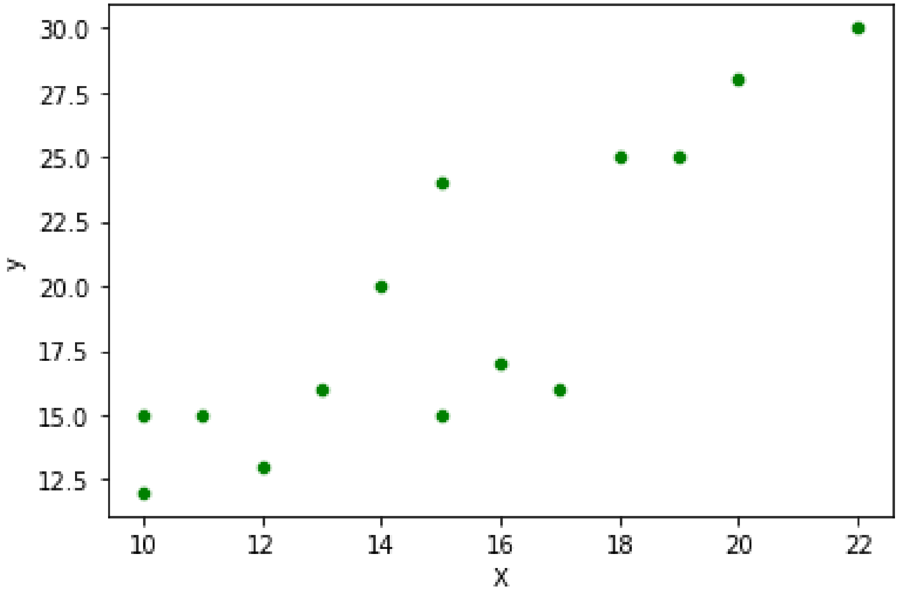
การวิเคราะห์การถดถอย (Regression Analysis)

เป็นการศึกษาและวิเคราะห์ตัวแบบความสัมพันธ์ของตัวแปรเชิงปริมาณตั้งแต่สองตัวขึ้นไป ซึ่งประกอบด้วย

- **ตัวแปรอิสระ (Independent variable) ใช้สัญลักษณ์ X**
  - สามารถทำการวิเคราะห์โดยใช้ตัวแปรอิสระเพียงตัวเดียว: การวิเคราะห์ถดถอยและสหสัมพันธ์อย่างง่าย (simple / univariate regression analysis) หรือมากกว่า 1 ตัว: การวิเคราะห์ถดถอยแบบพหุคูณ (multiple / multivariate regression analysis)
- **ตัวแปรตาม (Dependent variable) ใช้สัญลักษณ์ y**
  - ในการวิเคราะห์มักสนใจตัวแปรตามเพียงแค่ 1 ตัว (จึงใช้ y ตัวเล็ก)

สิ่งที่ศึกษา:

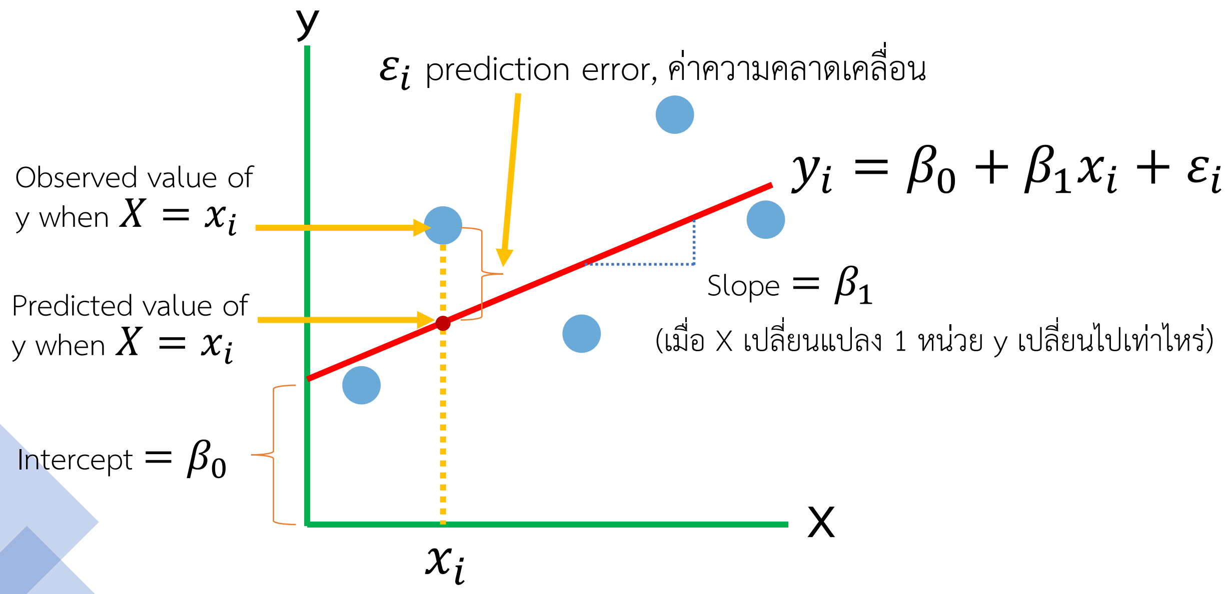
- X และ y มีความสัมพันธ์ในรูปแบบใด
- X และ y สัมพันธ์มากน้อยเพียงใด
- หาสมการเพื่อใช้สำหรับพยากรณ์หรือประมาณค่า y เมื่อทราบค่า x



การวิเคราะห์การถดถอยอย่างง่าย (Simple Regression Analysis)

เมื่อตัวแปรอิสระ (X) และตัวแปรตาม (y) มีความสัมพันธ์ลักษณะเชิงเส้นตรงแล้ว สามารถสร้างสมการถดถอยเพื่อใช้พยากรณ์ค่า y โดยใช้ค่า X

ตัวแบบความสัมพันธ์เชิงเส้นตรงสำหรับข้อมูลประชากร: $\boxed{y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$



** β_0 และ β_1 เป็นพารามิเตอร์ของการถดถอย

หาก

$\beta_1 = 0$ แสดงว่า X และ y ไม่มีความสัมพันธ์เชิงเส้นตรงเลย

$\beta_1 > 0$ แสดงว่า X และ y มีความสัมพันธ์เชิงเส้นตรงเชิง +

$\beta_1 < 0$ แสดงว่า X และ y มีความสัมพันธ์เชิงเส้นตรงเชิง -

สมการถดถอยของตัวอย่าง

ในทางปฏิบัติไม่สามารถหาค่าพารามิเตอร์ β_0 และ β_1 ได้ จึงต้องสุ่มข้อมูลตัวอย่างเพื่อประมาณค่าพารามิเตอร์ด้วยค่าสถิติ b_0 และ b_1 ดังสมการ:

$$\boxed{\hat{y} = b_0 + b_1 x_i}$$

การประมาณค่าพารามิเตอร์ในที่นี้ใช้วิธีกำลังสองน้อยที่สุด (Method of Least Square)

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$ หรือ $$\boxed{b_0 = \frac{\sum y - b_1 \sum x}{n}}$$

$$\boxed{b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}}$$ หรือ $$\boxed{b_1 = \frac{n\sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2}}$$ หรือ $$\boxed{b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2}}$$

ค่า b_0 และ b_1

ค่า b_0 เป็นค่าที่บอกให้ทราบว่า ถ้าค่า X เป็น 0 แล้วค่า y จะมีค่าเป็น b_0 โดยเฉลี่ย

- ถ้า $b_0 = 0$ แสดงว่าเส้นถดถอยตัดแกน y ที่จุดกำเนิด (Origin)

- ถ้า $b_0 < 0$ แสดงว่าเส้นถดถอยตัดแกน y ต่ำกว่าจุดกำเนิด

- ถ้า $b_0 > 0$ แสดงว่าเส้นถดถอยตัดแกน y เหนือจุดกำเนิด

ค่า b_1 เป็นค่าที่บอกอัตราการเพิ่มหรือลดลงของ y เมื่อ X มีค่าเพิ่มขึ้น 1 หน่วย

- ถ้า $b_1 > 0$ แสดงว่าตัวแปร X กับ y จะมีความสัมพันธ์ไปในทางเดียวกัน
  นั่นคือเมื่อ X มีค่าเพิ่มขึ้น y จะมีค่าเพิ่มขึ้น และถ้า X มีค่าลดลง y จะมีค่าลดลง
- ถ้า $b_1 < 0$ แสดงว่าตัวแปร X กับ y มีความสัมพันธ์กันในทางตรงกันข้าม
  นั่นคือเมื่อ X มีค่าเพิ่มขึ้น y จะมีค่าลดลง และถ้า X มีค่าลดลง y จะมีค่าเพิ่มขึ้น

ความคลาดเคลื่อนมาตรฐานในการประมาณค่า (Standard Error of the Estimate)

ความคลาดเคลื่อนมาตรฐานในการประมาณ (S) เป็นค่าวัดความแตกต่างระหว่างเส้นถดถอยที่ประมาณได้กับค่าของตัวอย่าง

$$S = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$$
$$= \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n-2}}$$

การทดสอบสมมติฐานค่า β_1 โดยการวิเคราะห์ความแปรปรวน

$\mathrm{H_0}: \beta_1 = 0$ (ตัวแปร X และ y ไม่มีความสัมพันธ์เชิงเส้นตรง)
$\mathrm{H_1}: \beta_1 \neq 0$ (ตัวแปร X และ y มีความสัมพันธ์เชิงเส้นตรง)

ค่าสถิติทดสอบ: $\mathrm{F} = \dfrac{\mathrm{MSR}}{\mathrm{MSE}}$

โดย $\mathrm{MSR} = \dfrac{\mathrm{SSR}}{1}$, $\mathrm{SSR} = b_0 \sum y + b_1 \sum xy - n\bar{y}^2$ (SSR: ความแปรผันจากค่าถดถอย)

$\mathrm{MSE} = \dfrac{\mathrm{SSE}}{n-2}$, $\mathrm{SSE} = \mathrm{SST} - \mathrm{SSR}$, $\mathrm{SST} = \sum y^2 - n\bar{y}^2$ (SSE: ความคลาดเคลื่อน, SST: ความแปรผันรวม)

ปฏิเสธ $\mathrm{H_0}$ เมื่อ

$\mathrm{F} > F_{\alpha,\ v_1=1,\ v_2=n-2}$ (เป็น right-tailed test)

สัมประสิทธิ์การตัดสินใจ (Coefficient of Determination)

สัมประสิทธิ์การตัดสินใจคือค่าที่ใช้อธิบายความผันแปรของ y ที่เกิดขึ้นว่าเป็นผลมาจากความผันแปรของตัวแปร X มากน้อยเพียงใด ใช้สัญลักษณ์แทนคือ r^2

$$\boxed{r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}}$$

r^2 มีค่าในช่วง 0 ถึง 1

- r^2 เข้าใกล้ 1 หมายความว่า ความผันแปรของตัวแปรตาม (y) ได้รับอิทธิพลมาจากความผันแปรของตัวแปรอิสระ (X) เท่ากับ 100%
- r^2 เข้าใกล้ 0 หมายความว่า ความผันแปรของ y ไม่ได้เกิดจากความผันแปรของ X เลย