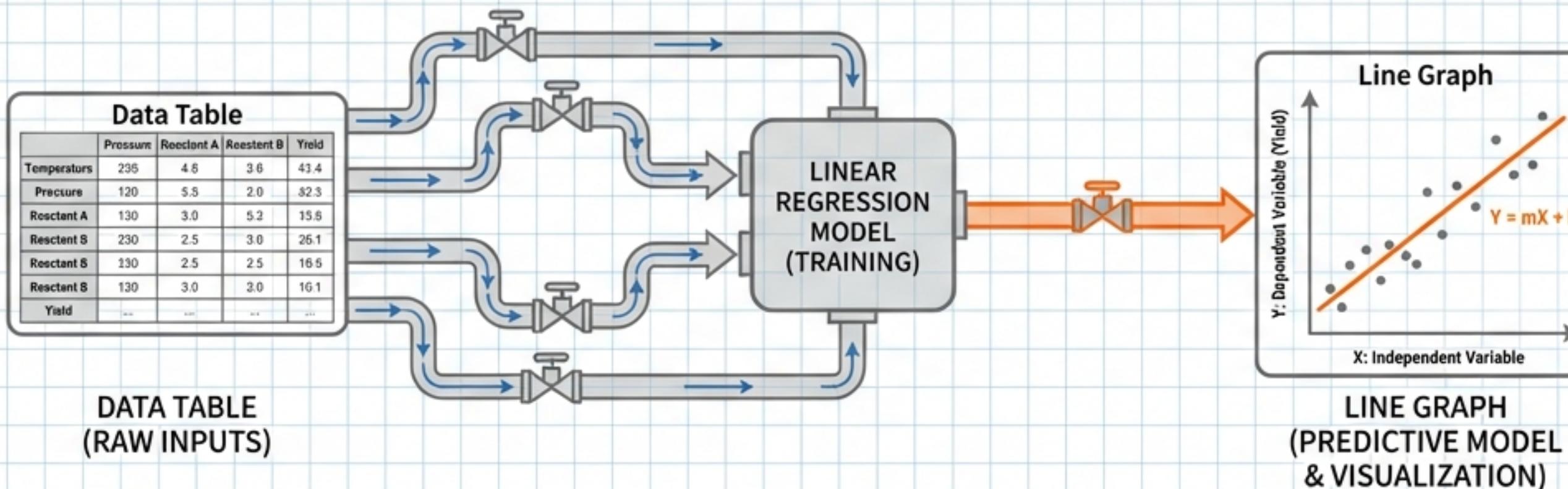


AI 在化工上的應用：Unit 10 線性回歸 (Linear Regression)

建立機器學習的基石：從最小二乘法到反應產率預測



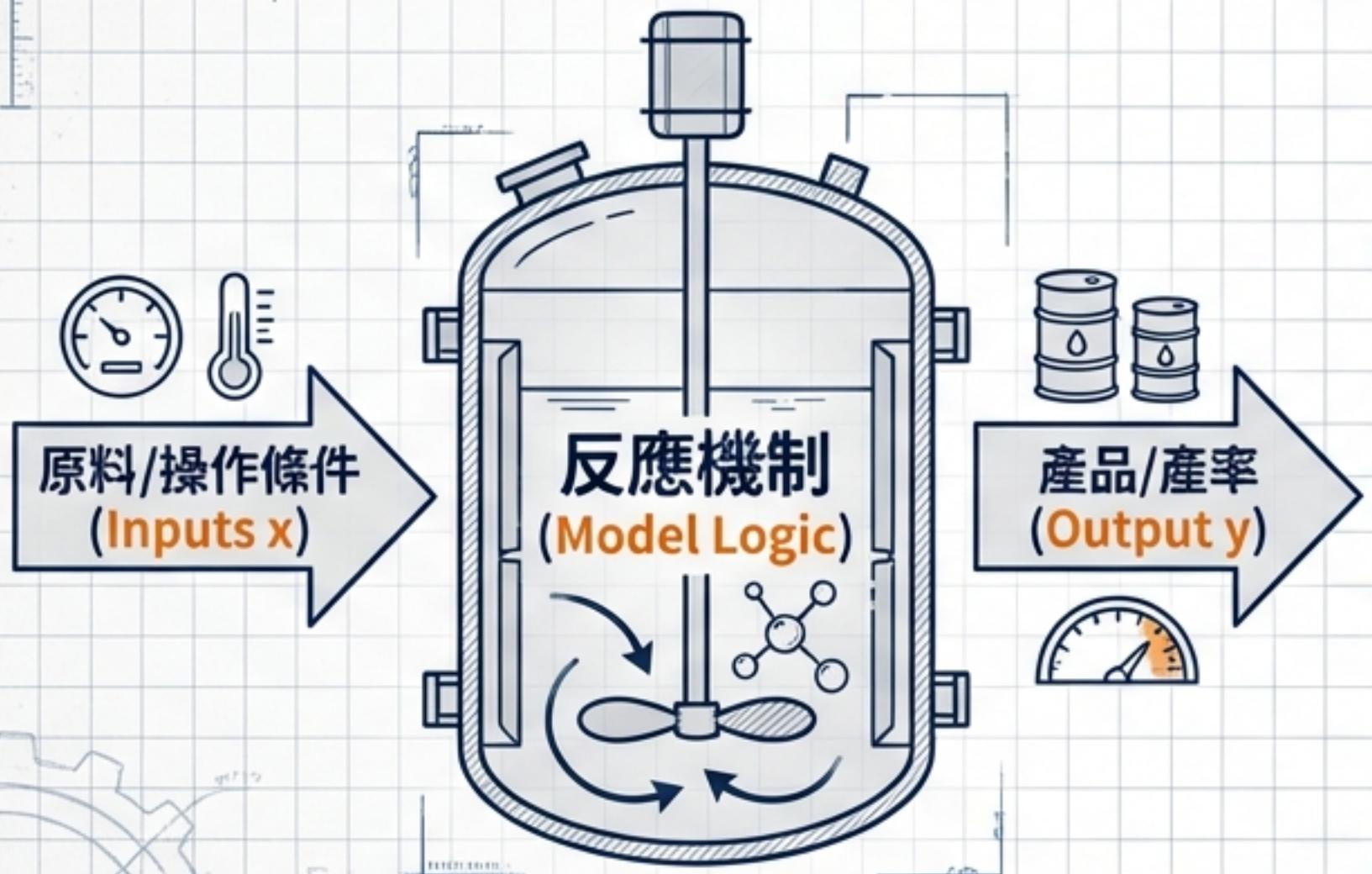
授課教師：莊曜楨 助理教授

課程單元：Unit 10

更新日期：2026-01-17 (Updated)

1. 基本概念：將數學視為製程單元

Physical Process:
Chemical Reactor (製程單元)



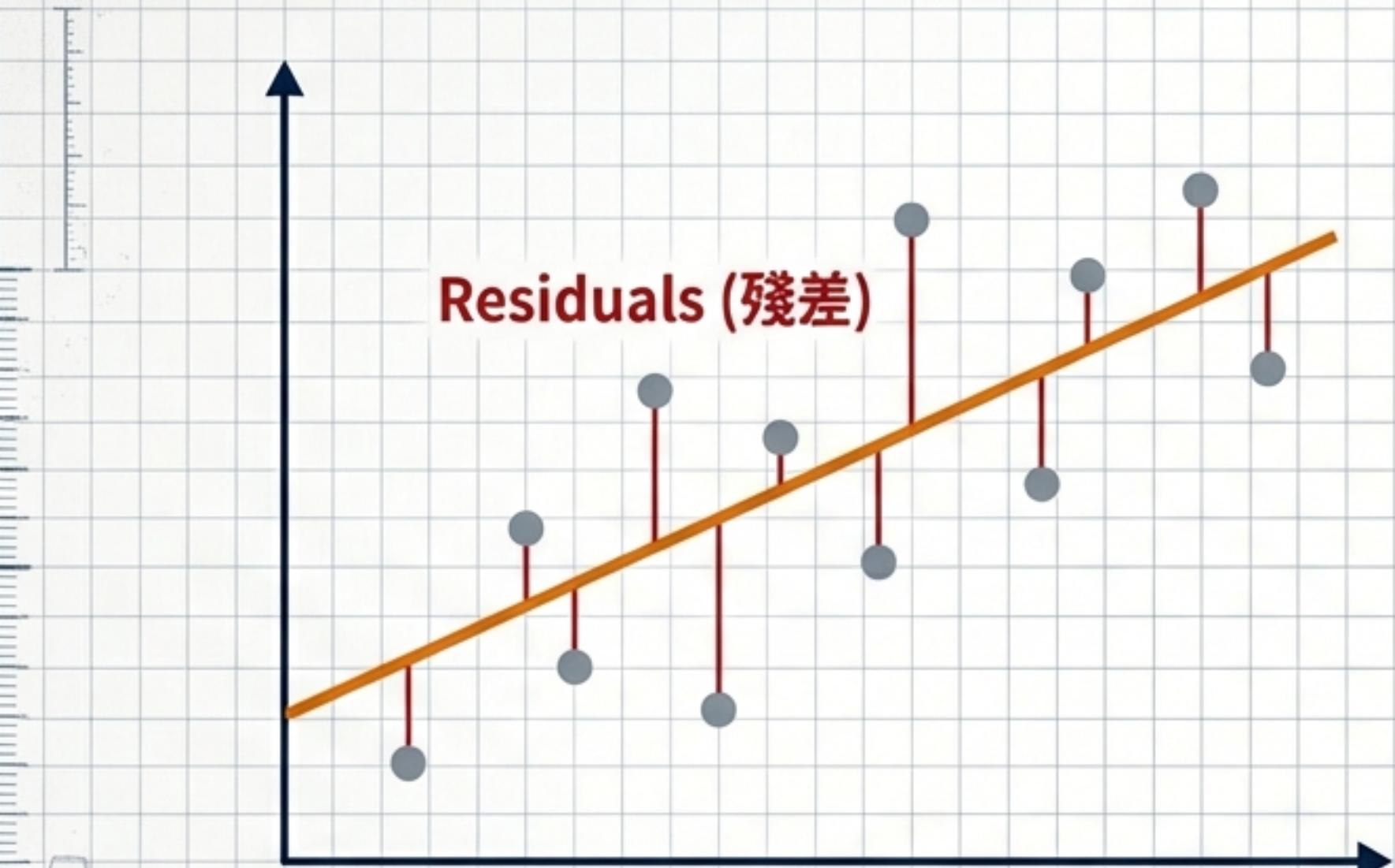
Mathematical Equation:
Linear Regression (線性迴歸模型)

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon$$
$$y = X\beta + \epsilon$$

- y (因變數)：目標變數 (如：產率、黏度)
- x (自變數)：特徵變數 (如：溫度、壓力)

- β (係數)：權重，代表該特徵對結果的影響力
- ϵ (誤差項)：雜訊，假設服從常態分佈 $N(0, \sigma^2)$

2. 核心原理：最小二乘法 (OLS)



Key Equation: $\text{RSS} = \sum (y_i - \hat{y}_i)^2$

Normal Equation: $\hat{\beta} = (X^T X)^{-1} X^T y$



目標是找到一組 β ，
使殘差平方和 (RSS)
最小化。

從幾何角度看，這是將目標向量 y
投影到特徵矩陣 X 的子空間上。

3. 操作限制：五大關鍵假設 (Assumptions)



安全檢查清單 (Safety Checklist)：啟動前審查 (Pre-startup Safety Review)

線性關係 (Linearity)：因變數與自變數需存在線性關聯 (檢查：殘差圖)。

獨立性 (Independence)：觀測值間互不影響 (無自相關)。

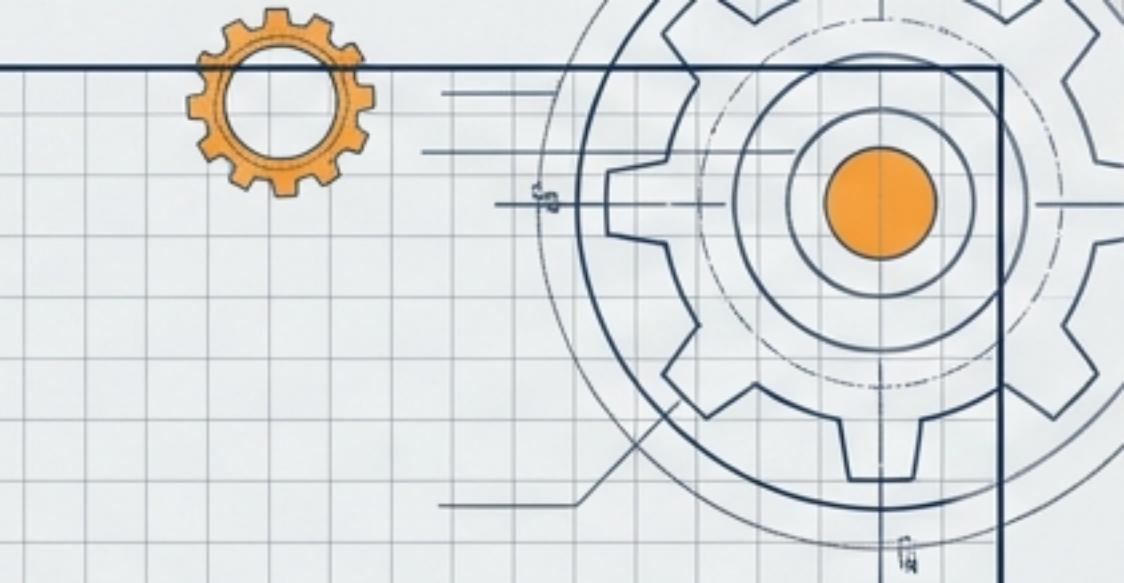
同質變異性 (Homoscedasticity)：殘差的變異數固定，不隨 x 變化 (變異均勻)。

常態性 (Normality)：誤差項 ϵ 需呈現常態分佈 (檢查：Q-Q 圖)。

無多重共線性 (No Multicollinearity)：自變數間不可高度相關 (檢查： $VIF < 5$)。

4. 實作工具：Scikit-learn Control Panel

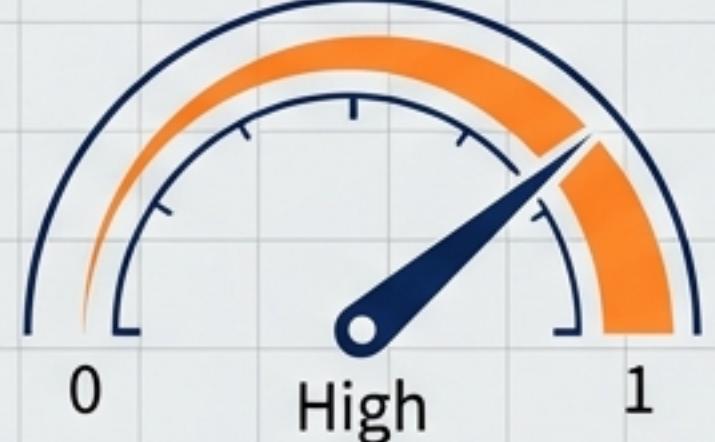
```
from sklearn.linear_model import LinearRegression  
  
# 1. 初始化模型（設定參數）  
model = LinearRegression(fit_intercept=True)  
  
# 2. 訓練模型（輸入 X, y）  
model.fit(X_train, y_train)  
  
# 3. 預測結果  
y_pred = model.predict(X_test)
```



關鍵參數 (Key Parameters)	
fit_intercept	是否計算截距 (預設 True)
n_jobs	平行運算核心數 (-1 為全開)

5. 品質控制：評估指標 (Metrics)

R^2 (決定係數)



模型解釋變異的比例 (1為完美)。數值越高越好，代表模型擬合度佳。

RMSE (均方根誤差)

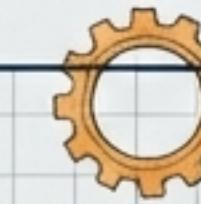
Formula: $\sqrt{\frac{\sum(y - \hat{y})^2}{m}}$

與 y 單位相同，直觀易懂 (例如：誤差為 $\pm 2.17\%$)。

MAE (平均絕對誤差)

誤差絕對值的平均。對異常值 (Outliers) 較不敏感。

6. 化工應用場景 (Applications)



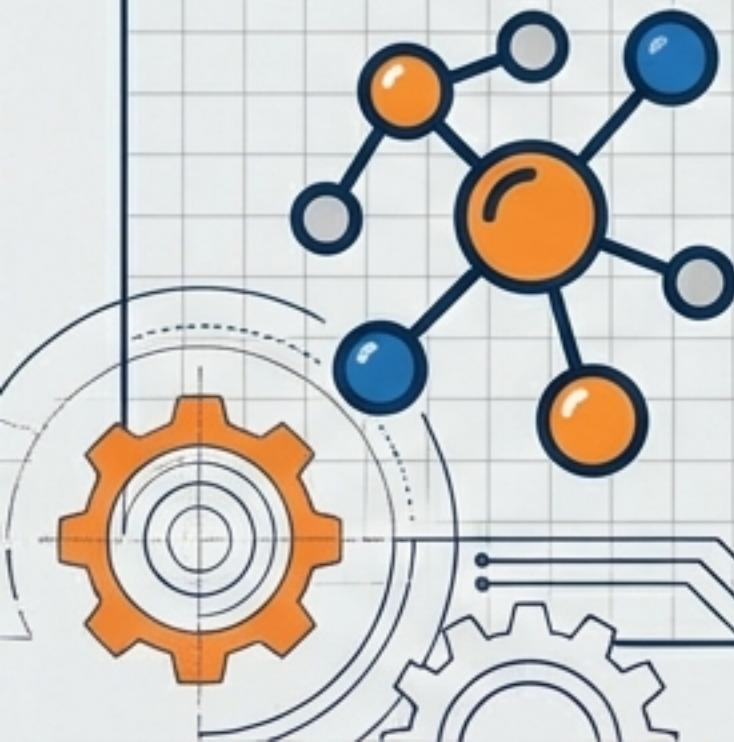
軟感測器 (Soft Sensors)

利用易測變數 (T , P) 預測難測指標 (黏度、純度)。



反應產率預測 (Yield Prediction)

建立操作條件與產率的定量關係 (本單元實作重點)。



物性估算 (Property Estimation)

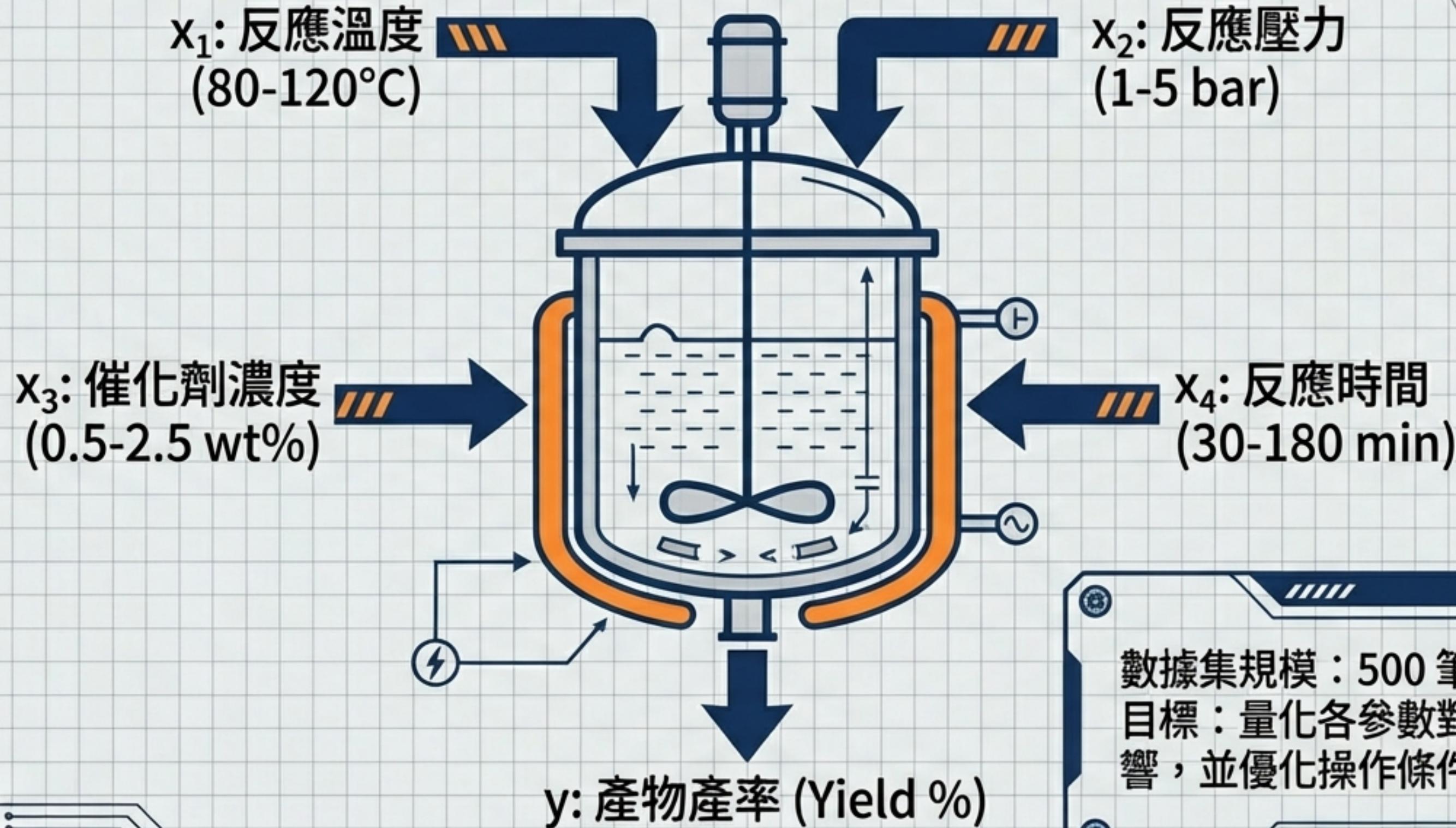
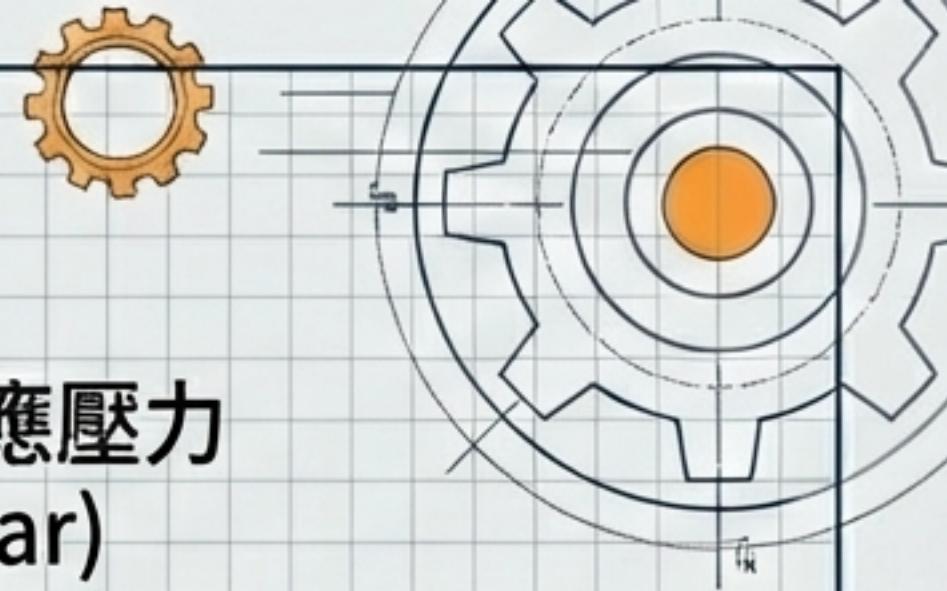
根據分子結構特徵預測沸點、溶解度。



品質控制 (Quality Control)

預測聚合物拉伸強度、斷裂伸長率。

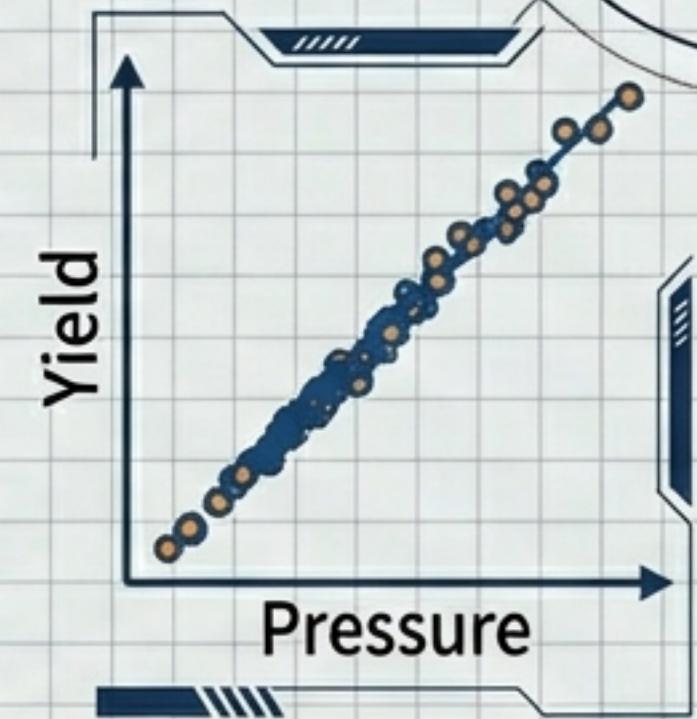
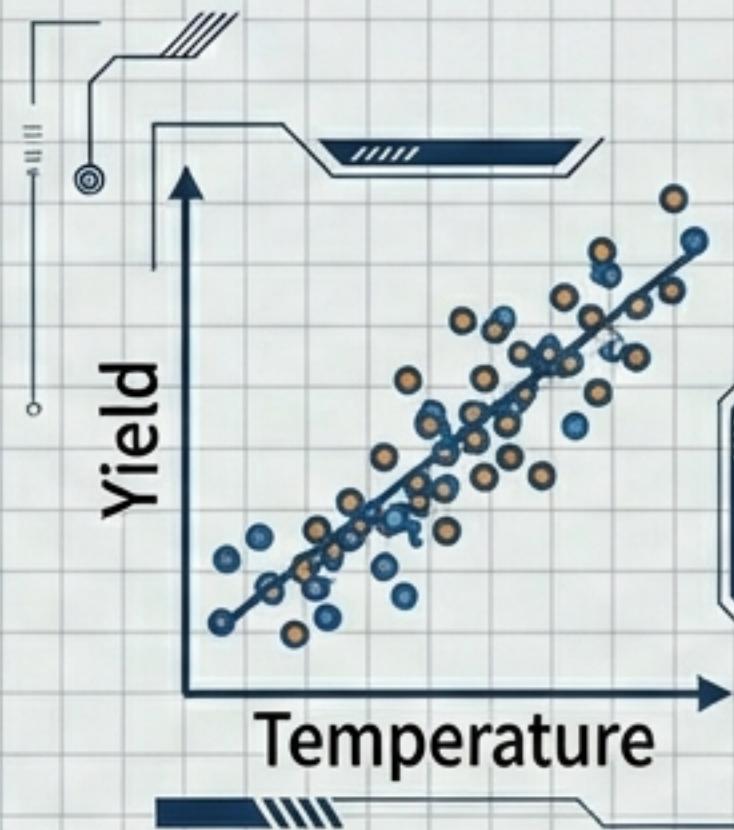
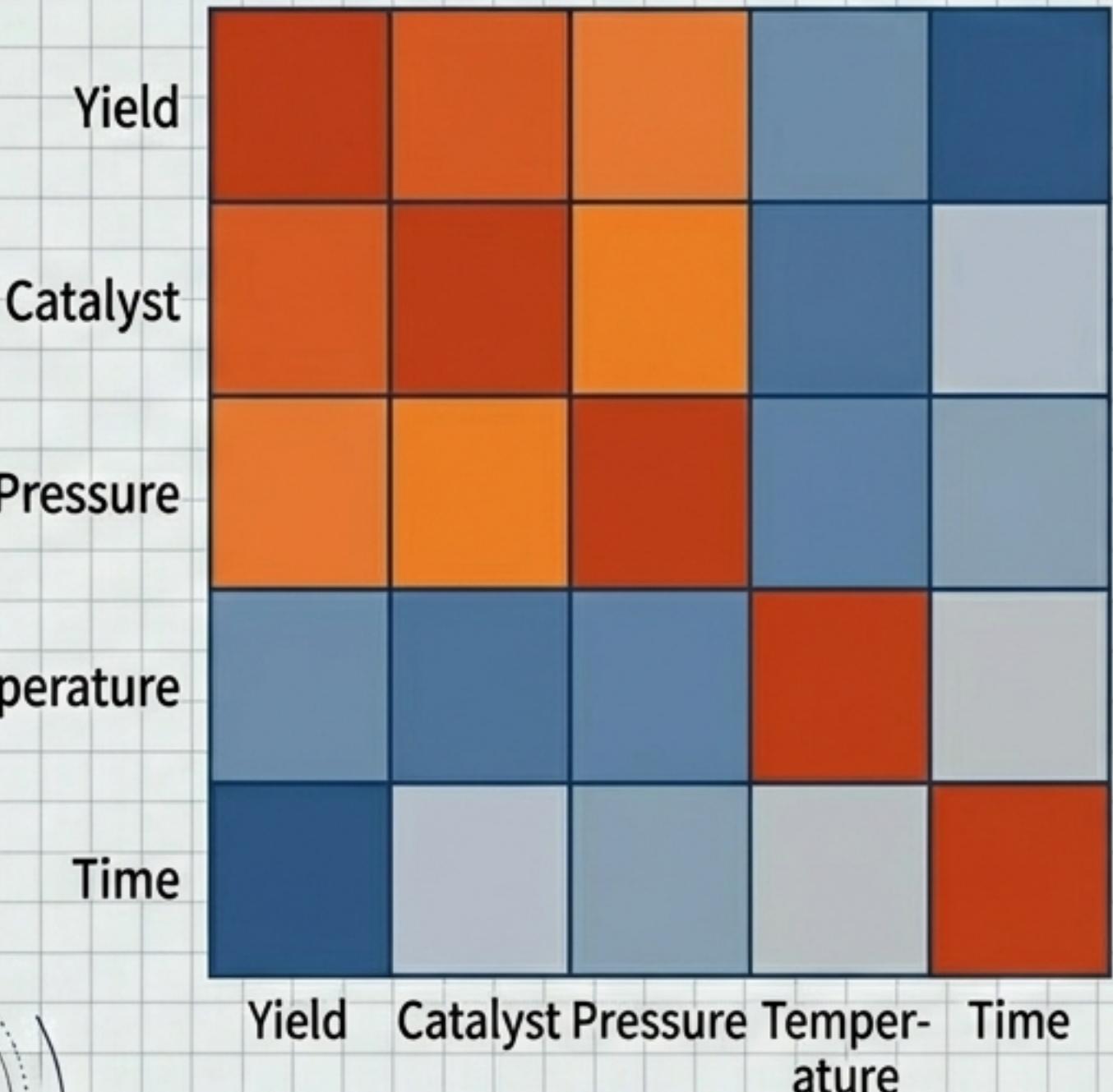
[實作案例] 化學反應產率預測：流程與變數



數據集規模：500 筆實驗數據
目標：量化各參數對產率的影響，並優化操作條件。

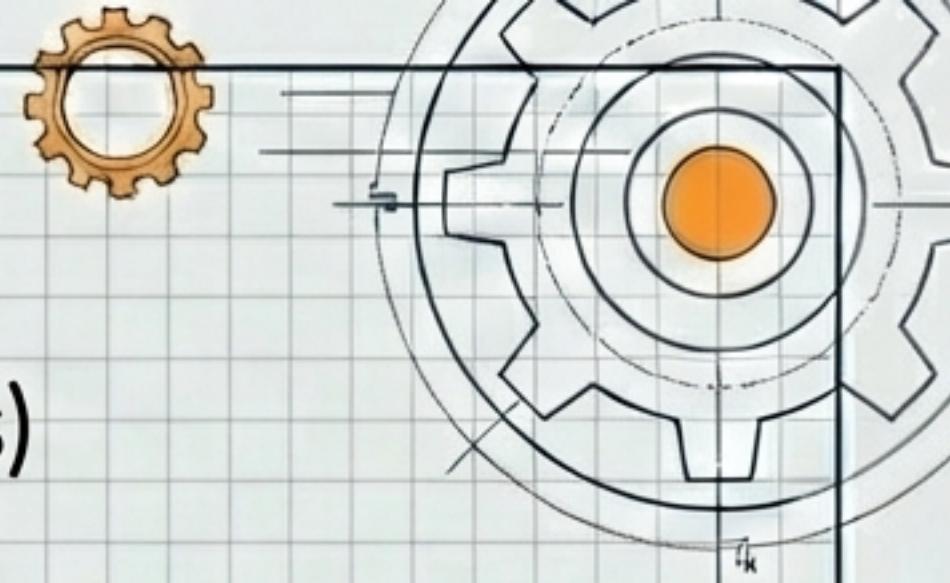
[實作案例] 數據探索與特徵分析

Correlation Heatmap

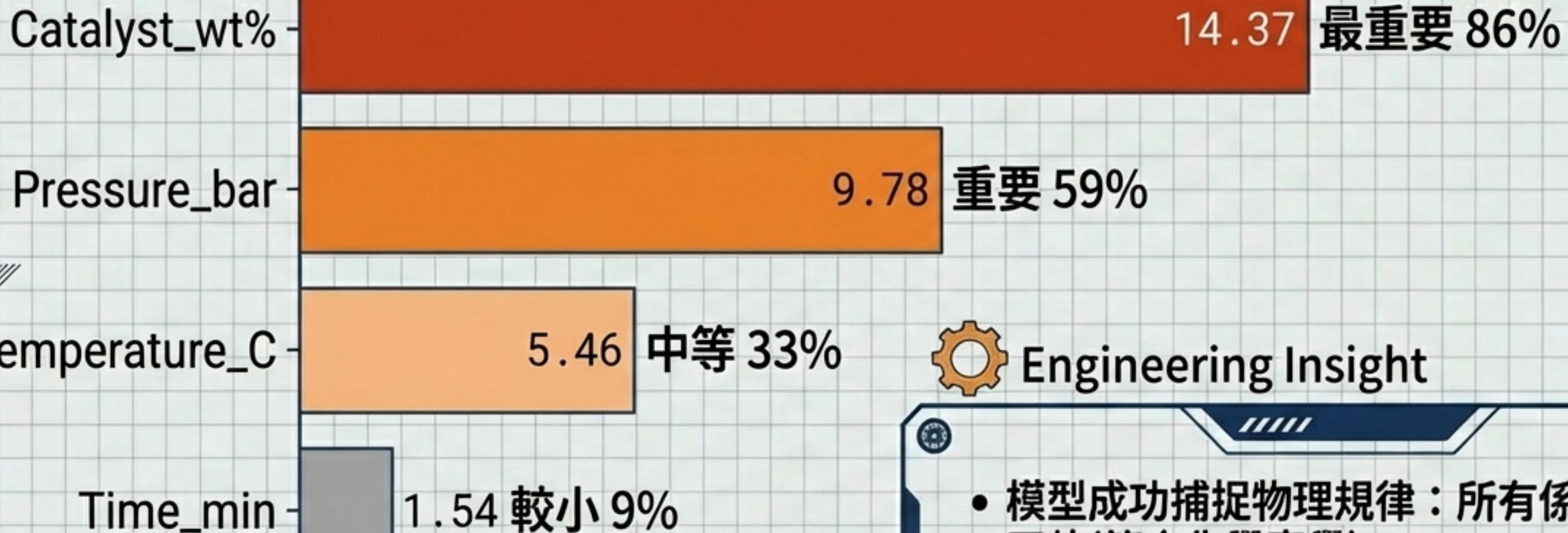


平均產率：94.45% (常態分佈)
相關性排序：催化劑 > 壓力 > 溫度 > 時間
Insight : 數據檢查確認特徵間無嚴重多重共線性 (VIF OK)，適合線性回歸。

[實作案例] 模型結果：物理意義解讀



Feature Importance (Standardized Coefficients)



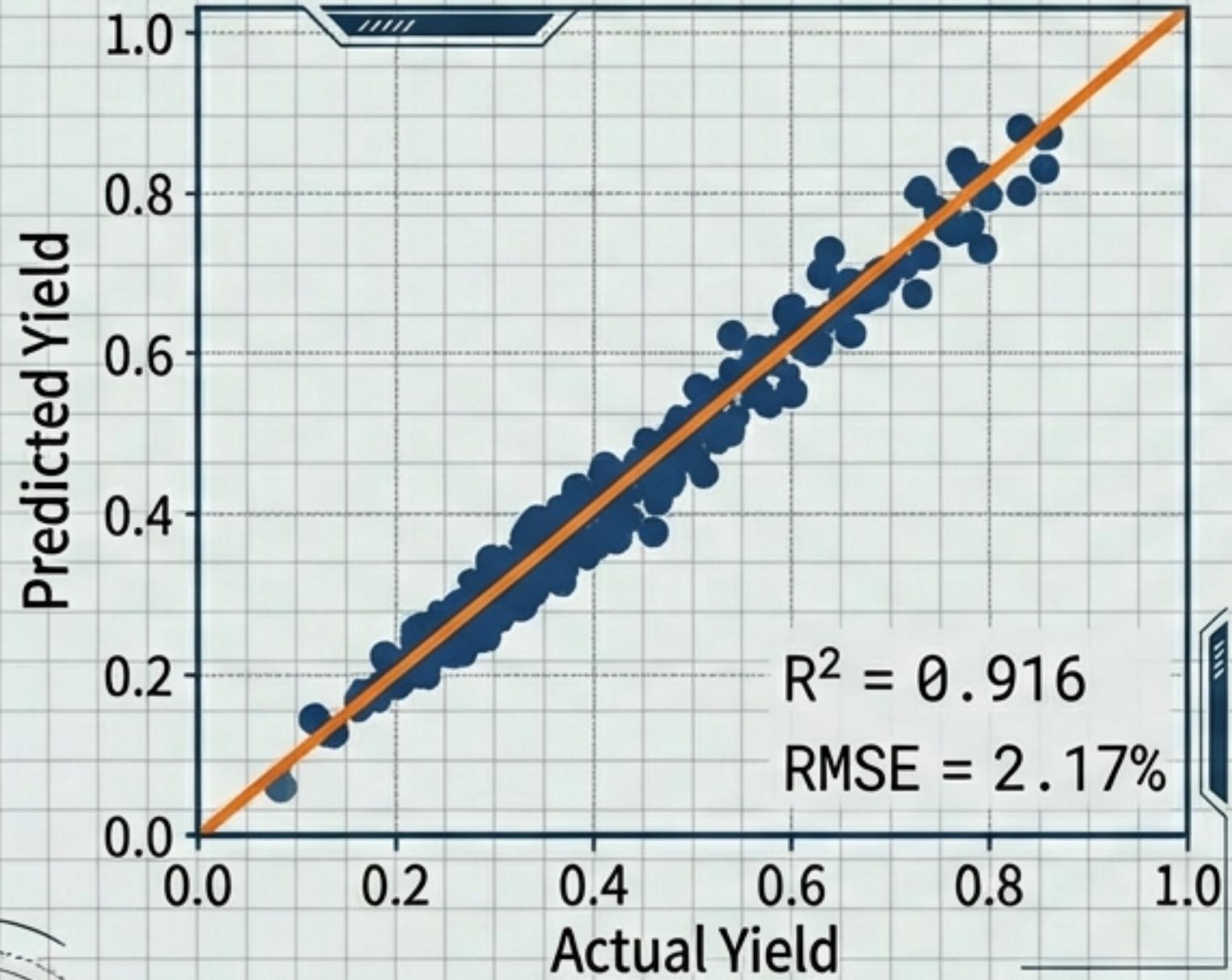
Engineering Insight

- 模型成功捕捉物理規律：所有係數均為正值(符合化學直覺)。
- 操作建議：優先調整催化劑與壓力以最大化產率效益。

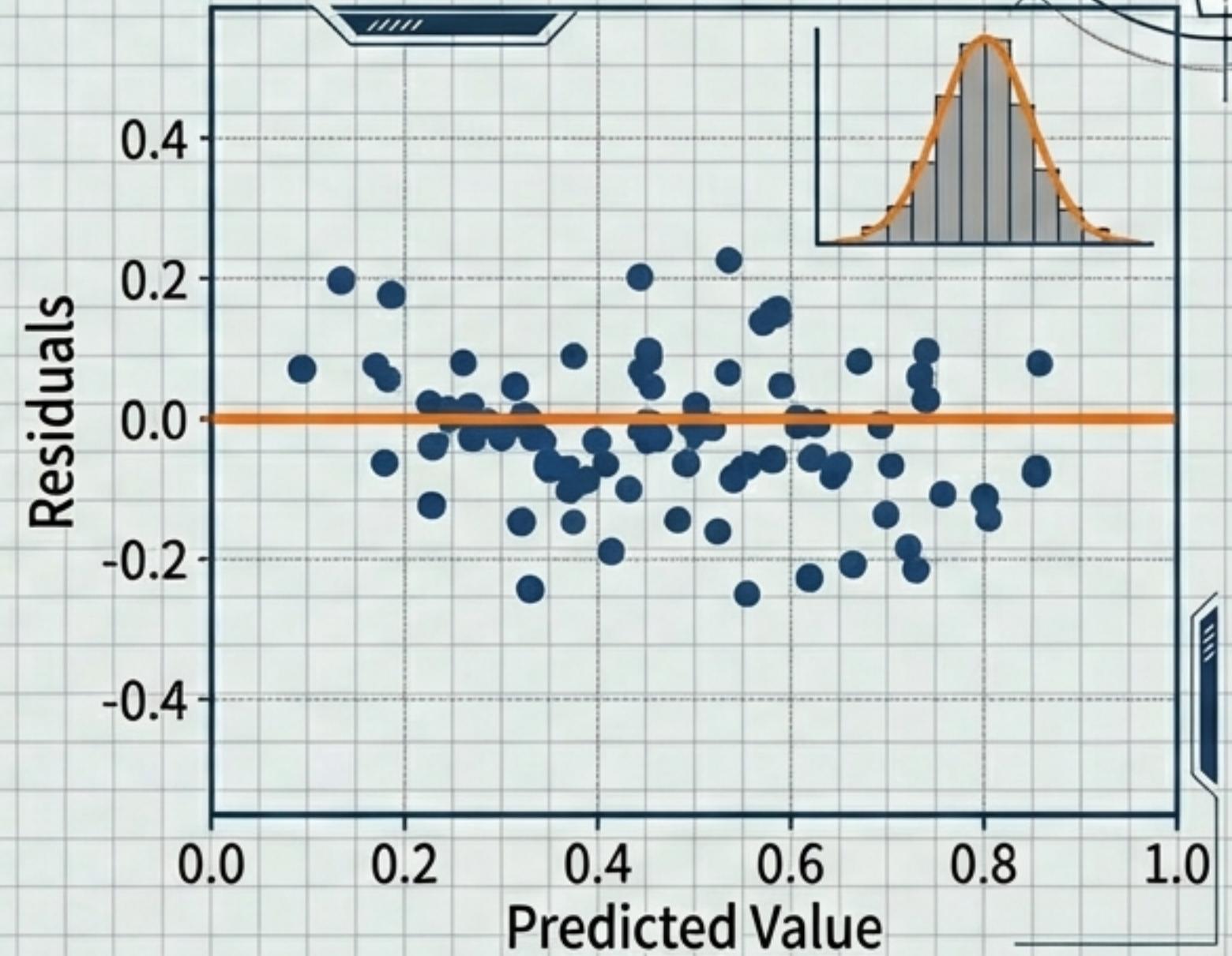
[實作案例] 模型診斷與驗證



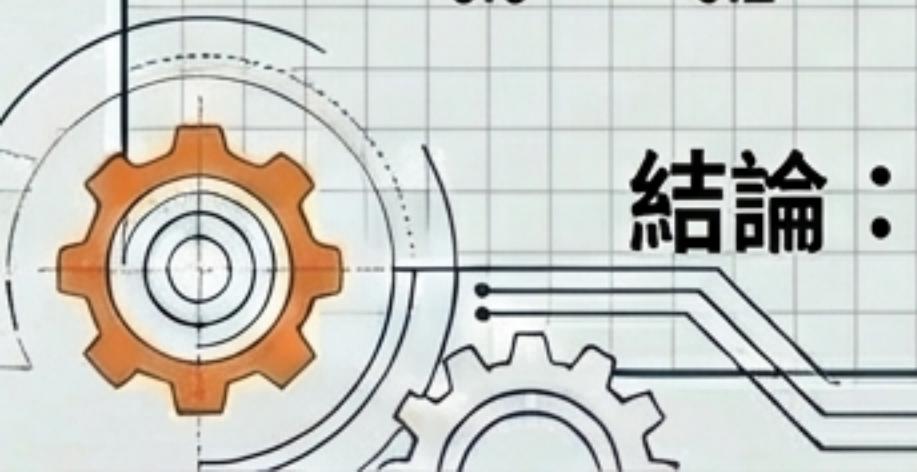
Parity Plot (預測值 vs 實際值)



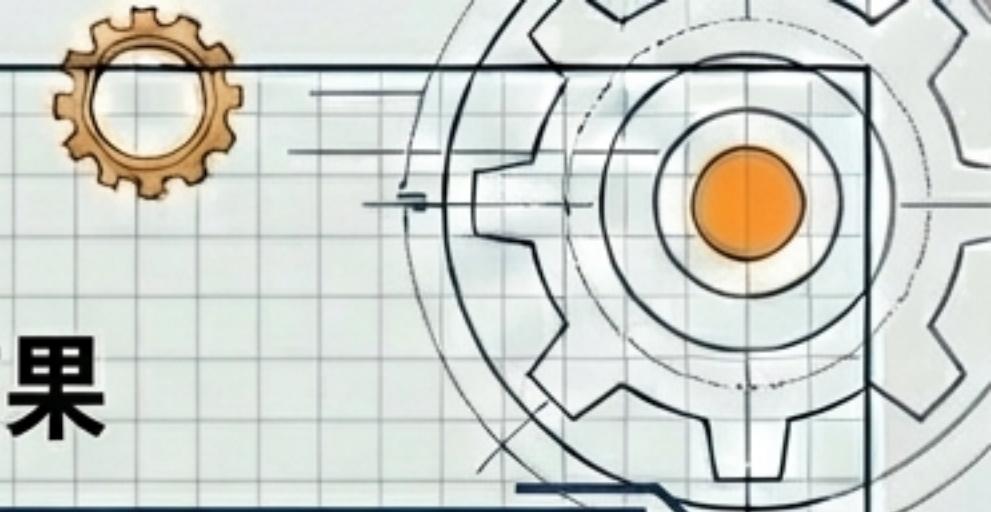
Residual Plot (殘差分佈)



結論：殘差符合常態性與同質性假設，模型有效且穩健。



[實作案例] 虛擬試驗與預測

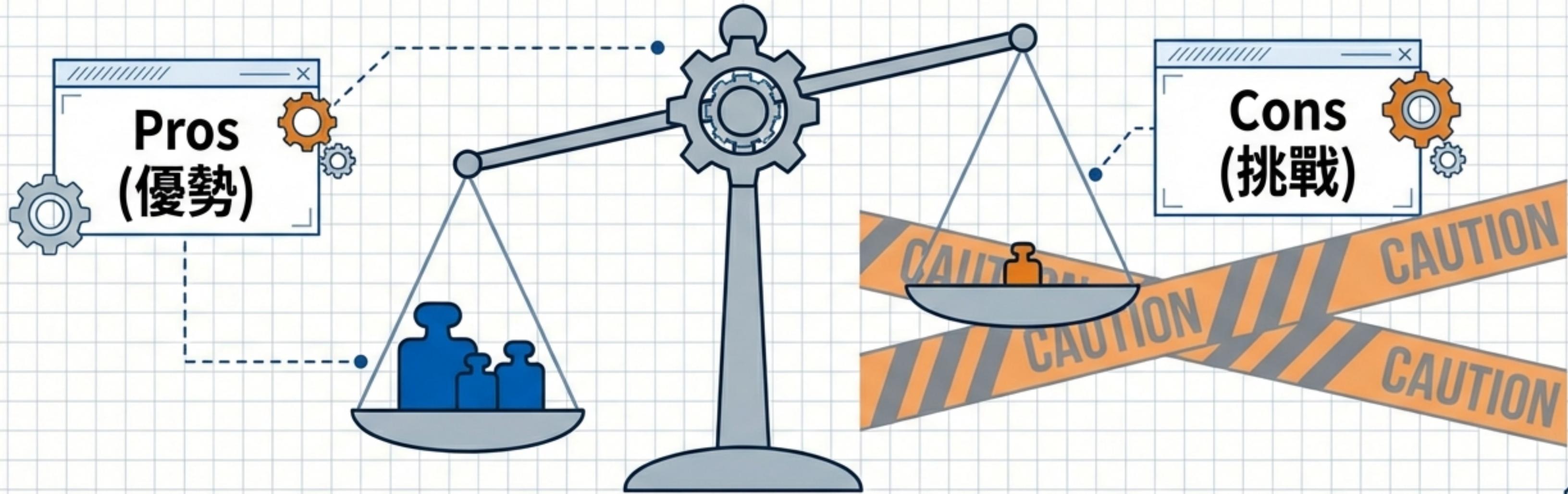


透過模型預測不同操作策略下的產率結果

Scenario	Temp (°C)	Pressure (bar)	Catalyst (wt%)	Time (min)	Predicted Yield (%)
1 (Conservative)	90	2.0	1.0	60	75.50
2 (Standard)	100	3.5	1.5	120	93.75
3 (Aggressive)	110	4.5	2.0	150	111.75

Value Proposition: 效益：減少試驗次數，快速鎖定最佳操作窗口
(High Yield & Economic Feasibility)。

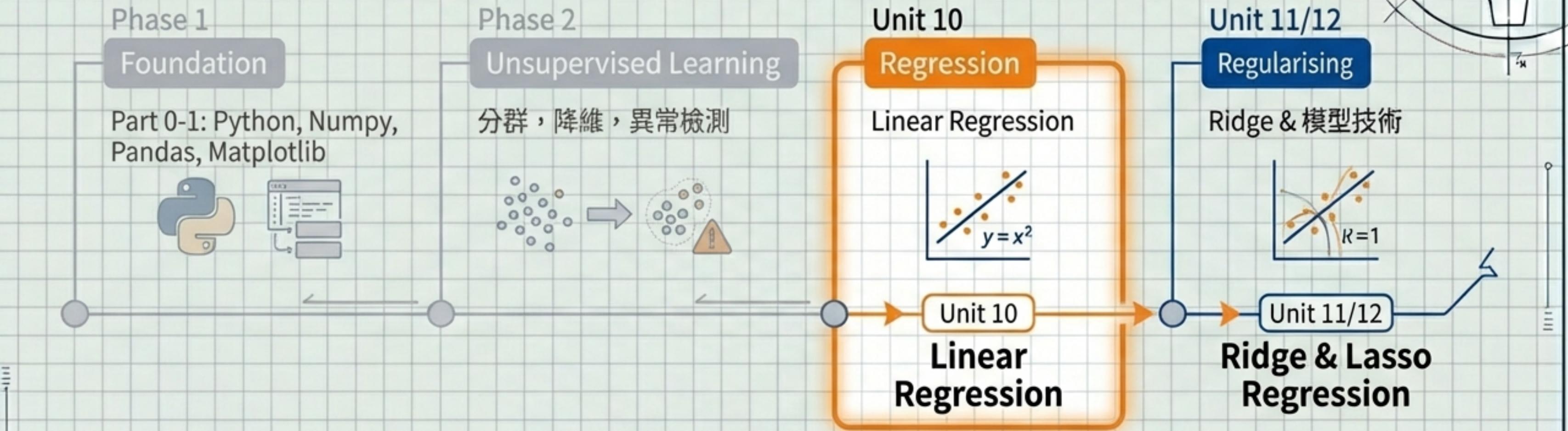
現實世界的挑戰：優勢與限制



- 可解釋性高：白箱模型，工程師能理解 Why
- 計算極快：適合即時控制 (Real-time control)
- 簡單穩健：抗雜訊能力強

- 線性限制：無法模擬複雜反應動力學
- 異常值敏感：設備故障數據可能扭曲模型
- 外推風險：切勿在訓練範圍外進行預測
(如：溫度 $> 150^{\circ}\text{C}$)

課程地圖與下一步



本單元建立了回歸基礎。

下一步：當遇到「多重共線性」或需要「特徵選擇」時，我們將引入正則化技術 (Regularization)。

**“簡單的模型往往是
最穩健的起點。』**

雖然深度學習強大，但線性回歸提供了化工數據
最需要的「可解釋性」與「物理意義」。