MB634 Week 2
แบบจำลองสำหรับการพยากรณ์เศรษฐกิจและธุรกิจ

เฉลิมพงษ์ คงเจริญ

17 สิงหาคม 2568

ประเภทของแบบจำลอง

- Supervised Learning
  - ตัวแปรที่ต้องการอธิบาย (Dependent variable, Target)
  - ตัวแปรอธิบาย (Explanatory variables, Factor/Feature)
  - ปัญหา Regression และ Classification
- Unsupervised Learning ไม่มีตัวแปรเป้าหมาย จะเป็นการศึกษาเกี่ยวกับการจัดกลุ่ม (Classification)

ประเภทของแบบจำลองตามเป้าหมาย

- Descriptive Model
- Explanatory Model/Inference Model
  - $Y = f(x)$
- Predictive Model
  - $Y = f(x) + \varepsilon$
  - สร้างแบบจำลอง systematic formation ระหว่าง Y และ X
  - รูปแบบของ function f(x) – linear, nonlinear
  - ตัวแปร X

Predictive Model

- Regression Problem
  - ตัวแปร Y เป็น Quantitative variable เช่น ราคาบ้าน (อธิบายด้วย อายุ ขนาด ...)
- Classification Problem
  - ตัวแปร Y เป็น Qualitative variable เช่น ซื้อ/ไม่ซื้อ ดี/ไม่ดี

แบบจำลองมีความแม่นยำ?

- สมการจริง $y = f(x) + \varepsilon$
- แบบจำลองที่เราสร้างขึ้น $\hat{f}(x)$
- เปรียบเทียบระหว่างค่าจริงและค่าพยากรณ์

$$f(x) + \varepsilon - \hat{f}(x)$$

- ε เป็นค่าผิดพลาดที่ไม่สามารถลดได้ (irreducible error)
- สิ่งที่เราพยายามทำคือลด forecast error (reducible error)

$$e = f(x) - \hat{f}(x)$$

แบบจำลองมีความแม่นยำ?

- Forecast error อาจจะมีค่าเป็นบวกหรือลบ ดังนั้น ในการคำนวณความผิดพลาดรวมจะอยู่ในรูป Mean Square Errors (MSE)

$$MSE = E[(f(x) - \hat{f}(x))\text{^}2]$$

หรือ

Root Mean Square Errors

$$RMSE = \sqrt{MSE}$$

เป็น**ตัววัดความแม่นยำของการพยากรณ์**

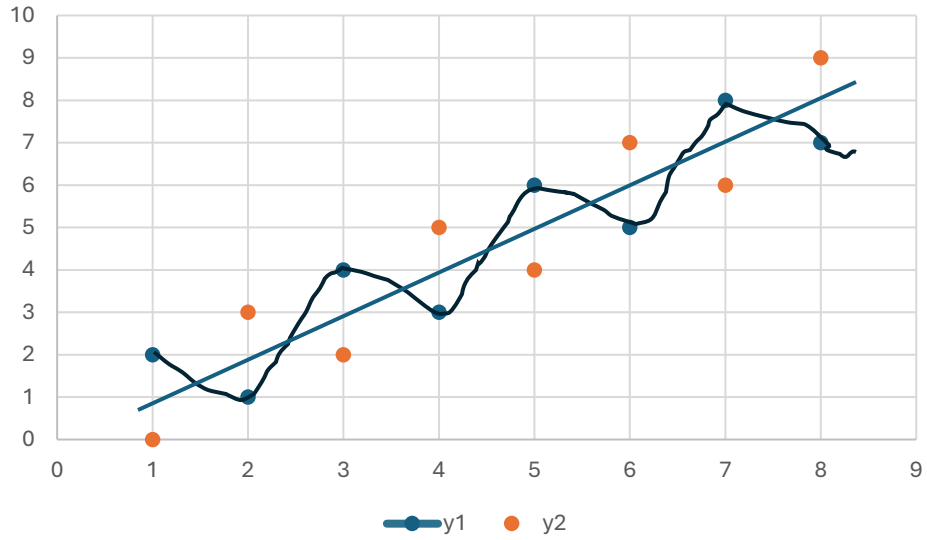
การสร้างแบบจำลอง

Target	Explanatory Variables		
Y_1	X_{11}	…	X_{K1}
Y_2	X_{12}		X_{K2}
…	…	…	…
Y_N	X_{1N}	…	X_{1N}

Predictor
$\hat{y}_1$
$\hat{y}_2$
…
$\hat{y}_N$

- หากเราเปรียบเทียบ in-sample forecast กับข้อมูล in-sample ด้วย in-sample RMSE จะได้ค่าที่ดูดีเกินไป และมีแนวโน้มที่จะ Overfit (อธิบายข้อมูลที่ใช้สร้างแบบจำลองได้ดีมาก แต่ใช้กับข้อมูลชุดใหม่มีความผิดพลาดมาก)

Overfit



- หากสร้างแบบจำลองจากข้อมูล (sample) สีน้ำเงิน แทนด้วยเส้นสีน้ำเงิน
- $bias_{insample} = E(y - \hat{y}) = 0$
- หากนำไปพยากรณ์ข้อมูลสีส้ม Variance จะสูง

$$Variance_{outofsample} = E(y - \hat{y})^2$$

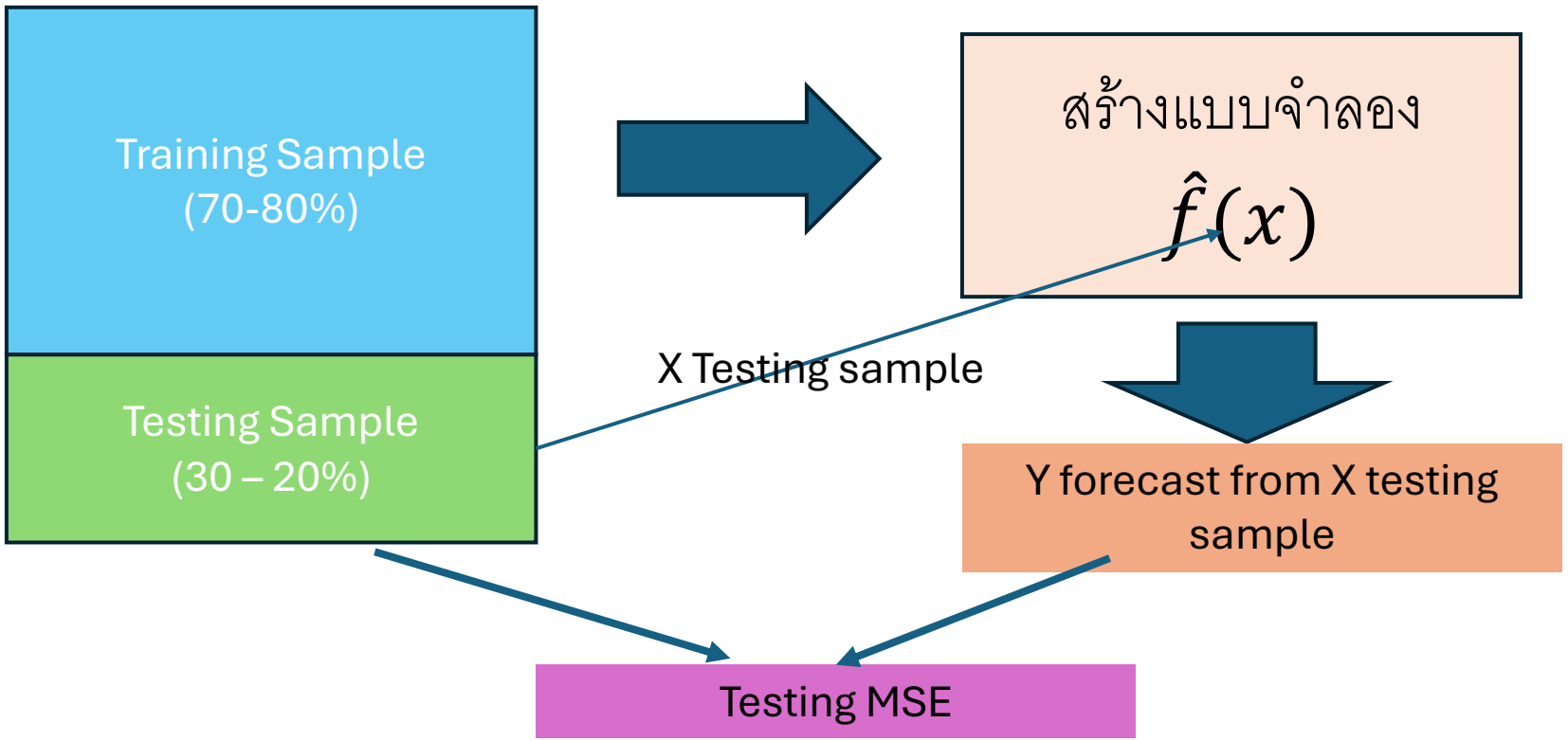
หากใช้เส้นตรงสีเขียว $bias_{insample} > 0$

แต่ $Variance_{outofsample}$ จะน้อยกว่า

- เราจะพิจารณา Out-of-Sample MSE

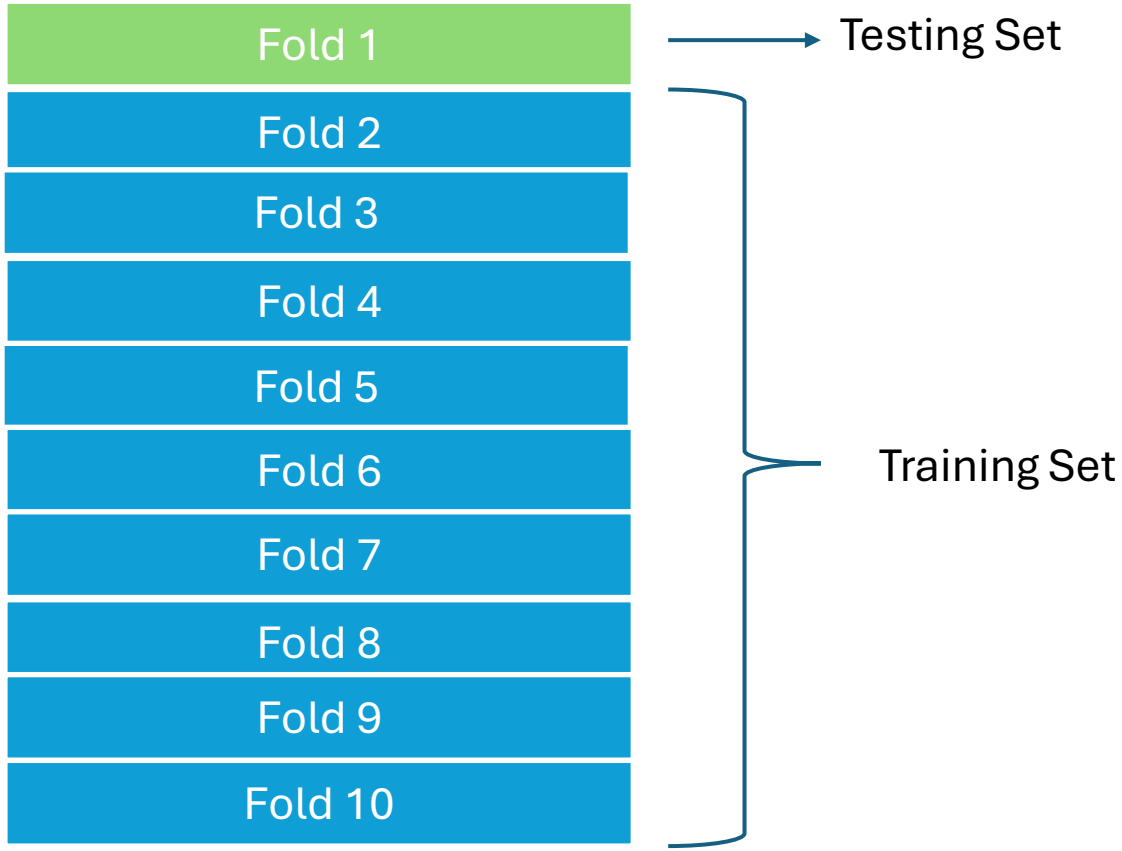
Out-of-sample RMSE

- Out-of-sample MSE = bias^2 + Variance



Validation Set Approach

10-fold cross validation



Testing	Training	Testing MSE
1	2,3,4,5,6,7,8,9,10	MSE1
2	1,3,4,5,6,7,8,9,10	MSE2
3	1,2,4,5,6,7,8,9,10	MSE3
4	1,2,3,5,6,7,8,9,10	MSE4
5	1,2,3,4,6,7,8,9,10	MSE5
6	1,2,3,4,5,7,8,9,10	MSE6
7	1,2,3,4,5,6,8,9,10	MSE7
8	1,2,3,4,5,6,7,9,10	MSE8
9	1,2,3,4,5,6,7,8,10	MSE9
10	1,2,3,4,5,6,7,8,9	MSE10

$$Ave.Testing\ MSE = \sum_{j}^{10} Testing\ MSE_j/10$$

แบบจำลองที่ใช้พยากรณ์

- เมื่อเลือกแบบจำลองที่ดีที่สุดแล้ว
- ใช้ข้อมูลทั้งหมด (full sample) ประมาณค่า แบบจำลองเพื่อใช้ในการพยากรณ์

การประเมินแบบจำลอง Regression

- แบบจำลองเส้นตรง $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
- เราต้องการเลือกตัวแปรพยากรณ์

$$\{x_1, \ldots, x_d\} \in \{x_1, \ldots, x_p\}$$

เราประมาณสมการ Regression ด้วย Least Square Method

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_d X_d$$

เราสามารถคำนวณ In-sample Residual Sum of Square (RSS)

$$Training\ MSE = \frac{RSS}{n} = \sum_i \frac{(y_i - \hat{y}_i)^2}{n}$$

การใช้ In-sample MSE ในการเปรียบเทียบแบบจำลอง จะก่อให้เกิด Overfitting

การประเมินแบบจำลอง Regression

1) แบ่งข้อมูลออกเป็น training sample ($\boldsymbol{x}$) และ testing sample ($\boldsymbol{x}^*$)

- ประมาณค่า $\hat{\beta}_i$ ด้วยชุดข้อมูล training sample ($\boldsymbol{x}$)
- แทนค่า x จาก testing sample ($\boldsymbol{x}^*$)

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_p x_p^*$$

คำนวณ Testing MSE (Validation MSE)

$$\sum_j^m \frac{\left(y_j^* - \hat{y}_j^*\right)^2}{m}$$

เลือกแบบจำลองที่ Testing MSE มีค่าต่ำที่สุด

การประเมินแบบจำลอง Regression

2) K-fold cross-validation

- จากตัวอย่าง N ตัวอย่าง แบ่งเป็น k กลุ่ม (fold) แต่ละกลุ่มมีสมาชิก N/K
- แต่ละครั้งที่ $\boldsymbol{i}$ จะเลือก fold $\boldsymbol{i}$ เป็น Testing sample และ fold ที่เหลือเป็น Training sample

$$CV_k = \sum \frac{Validation\ MSE_i}{k}$$

เราจะเปรียบเทียบ CVk ของแต่ละ subset ของ $\{x_1, \ldots, x_p\}$

Best subset selection

1. M0 คือแบบจำลองที่มีแต่ละตัวแปร
2. สำหรับแต่ละ k =1,...,p
   - ประมาณค่าทุกกรณีของ k
   - เลือกแบบจำลองที่ดีที่สุดในแต่ละจำนวนตัวแปร – Mk (ใช้ RSS ต่ำที่สุด)
3. เลือกแบบจำลองที่ดีที่สุดใน M0, M1,..., Mp โดยใช้ Cross-validation หรือ Model Selection criteria (AIC, BIC)

จำนวนแบบจำลองที่ต้องพิจารณาสูง

Forward selection

1. M0 คือแบบจำลองที่มีแต่ละตัวแปร
2. สำหรับแต่ละ k =0,...,p-1
   - ประมาณค่าทุกแบบจำลอง p-k แบบจำลอง ซึ่งเพิ่มตัวแปร 1 ตัวแปรจากรอบที่ผ่านมา k-1
   - เลือกแบบจำลองที่ดีที่สุด เป็นตัวแทนของ M_(k+1)
   - M0 → M1 → M2 → ... → Mp
3. เลือกแบบจำลองที่ดีที่สุดใน M0, M1,..., Mp โดยใช้ Cross-validation หรือ Model Selection criteria (AIC, BIC)

Backward selection

- Mp คือแบบจำลองที่มีทุกตัวแปร
- สำหรับแต่ละ k = p-1, p-2,...1
  - ประมาณค่าทุกแบบจำลอง k แบบจำลอง ซึ่งลดตัวแปร 1 ตัวแปรจากรอบที่ผ่านมา k-1
  - เลือกแบบจำลองที่ดีที่สุด เป็นตัวแทนของ M_(k)
  - Mp → M_p-1 → M_p-2 → ... → M0
- เลือกแบบจำลองที่ดีที่สุดใน M0, M1,..., Mp โดยใช้ Cross-validation หรือ Model Selection criteria (AIC, BIC)

การเลือกแบบจำลอง

- การเพิ่มตัวแปรจะทำให้ RSS ลดลง และ Training Error มักจะมีค่าต่ำกว่า Testing Error
- เราสามารถแก้ไขปัญหาได้โดย
- 1) การประมาณค่า Testing Error จาก Validation Set หรือ Cross Validation
- 2) การประมาณค่า Testing Error ทางอ้อม โดยการปรับ Bias จาก Training Error

การปรับ Bias จาก Training Set

- Mellow's Cp เป็น unbiased estimator ของ Tesing MSE

$$Mellow's\ Cp = \frac{1}{n}(RSS + 2d\ \hat{\sigma}^2)$$

โดยที่ $\hat{\sigma}^2$ คือ estimate of variance of error และ d คือจำนวนตัวแปร

- Akaike Information Criteria (AIC)

$$AIC = \frac{1}{n\hat{\sigma}^2}(RSS + 2d\ \hat{\sigma}^2)$$

- Bayesian Information Criteria (BIC)

$$BIC = \frac{1}{n\hat{\sigma}^2}(RSS + \log(n)d\ \hat{\sigma}^2)$$