機器學習第四章：完整筆記與模擬出題（未壓縮）

本文件包含多元線性回歸的名詞解釋、名詞比較與模擬出題（中英文對照），所有內容逐字保留未壓縮

# 📘 第四章：多元線性回歸（Multiple Linear Regression）

## 🧠 名詞解釋（Term Explanation）

### 多元線性回歸（Multiple Linear Regression）

中文：多元線性回歸包含多個自變數，形式為 Y = β₀ + β₁X₁ + ... + βₚXₚ + ε。簡單線性回歸僅有一個自變數。

英文：Multiple linear regression includes multiple predictors and is written as Y = β₀ + β₁X₁ + ... + βₚXₚ + ε. Simple linear regression uses only one predictor.

### 共線性（Multicollinearity）

中文：共線性指自變數之間高度相關，會導致模型係數估計不穩定、標準誤增加，甚至影響變數解釋力。

英文：Multicollinearity occurs when predictors are highly correlated, causing unstable coefficient estimates, inflated standard errors, and reduced interpretability.

### 虛擬變數（Dummy Variable）

中文：類別變數以文字表示類別，無法直接用於回歸；虛擬變數為將其轉換成 0/1 數值後可供模型使用。

英文：Categorical variables are labels representing categories, not usable directly in regression. Dummy variables convert them into binary numerical form (0/1) for modeling.

### 變數交互作用（Interaction Effect）

中文：交互作用表示兩變數聯合對 Y 有影響，常見表示法為 X₁×X₂。

英文：Interaction effects indicate that two variables jointly affect Y, and are represented as X₁×X₂ in regression models.

### 變數選擇（Variable Selection）

中文：可考慮刪除其中一個變數、合併變數、或使用主成分分析（PCA）降低維度。

英文：You may drop one variable, combine them, or use Principal Component Analysis (PCA) to reduce dimensionality.

### 嵌入式方法（Embedded Method）

中文：將特徵選擇融入模型訓練過程，如 LASSO 或 Decision Tree 內建的重要性排名。

英文：Feature selection methods embedded in the training process of the model, such as LASSO or tree-based importance scoring.

### 非線性效應與多項式回歸（Polynomial Regression）

中文：如果變數與反應變數之間為曲線關係，可透過加入高次項（如 X², X³）模擬非線性。

英文：When the relationship is nonlinear, polynomial terms like X², X³ are added to the model to approximate the curve.

### 多變量多元回歸（Multivariate Multiple Regression）

中文：同時預測多個目標變數，並考慮其間的相關性。

英文：A regression setting where multiple dependent variables are predicted simultaneously, accounting for correlation between them.

### 過度擬合（Overfitting）

中文：過度擬合會導致模型在訓練資料表現很好，但在新資料表現不佳。可用交叉驗證、簡化模型、正則化等方式避免。

英文：Overfitting leads to poor generalization. Prevent it using cross-validation, model simplification, or regularization techniques.

### 欠擬合（Underfitting）

中文：模型太簡單，無法充分捕捉資料的結構與模式，導致訓練誤差與測試誤差都很高。

英文：When a model is too simple to capture the structure of the data, resulting in high error on both training and test sets.

## 📚 名詞比較（Term Comparison）

Dummy Variable vs Categorical Variable

|  |  |  |
| --- | --- | --- |
| 項目 | 類別變數（Categorical Variable） | 虛擬變數（Dummy Variable） |
| 定義 | 用來表示分類的文字或編號 | 將類別變數轉換成 0/1 的數值變數 |
| 在回歸模型中使用 | 無法直接使用 | 可作為模型中的數值特徵 |
| 範例 | “男”、“女” | 男 = 1, 女 = 0 或用多個變數表示多類別 |

Forward vs Backward Selection

|  |  |  |
| --- | --- | --- |
| 項目 | Forward Selection | Backward Elimination |
| 起始模型 | 從空模型開始 | 從包含所有變數的完整模型開始 |
| 操作方式 | 每次加入對模型最有貢獻的變數 | 每次移除對模型貢獻最小的變數 |
| 優點 | 避免加入無用變數 | 初始擁有完整資訊，有助於避免遺漏重要變數 |
| 適用情況 | 當變數很多時效果好 | 適合資料量大、變數相對較少的情況 |

# 📝 第四章模擬出題（Chapter 4 Mock Questions）

## 考題 1

中文問題：何謂多元線性回歸？其與簡單線性回歸有何差異？

英文問題：What is multiple linear regression? How does it differ from simple linear regression?

中文答案：多元線性回歸包含多個自變數，形式為 Y = β₀ + β₁X₁ + ... + βₚXₚ + ε。簡單線性回歸僅有一個自變數。

英文答案：Multiple linear regression includes multiple predictors and is written as Y = β₀ + β₁X₁ + ... + βₚXₚ + ε. Simple linear regression uses only one predictor.

## 考題 2

中文問題：什麼是共線性？對模型有什麼影響？

英文問題：What is multicollinearity? How does it affect the model?

中文答案：共線性指自變數之間高度相關，會導致模型係數估計不穩定、標準誤增加，甚至影響變數解釋力。

英文答案：Multicollinearity occurs when predictors are highly correlated, causing unstable coefficient estimates, inflated standard errors, and reduced interpretability.

## 考題 3

中文問題：請比較 Dummy Variable 與 Categorical Variable 的差異。

英文問題：Compare dummy variables and categorical variables.

中文答案：類別變數以文字表示類別，無法直接用於回歸；虛擬變數為將其轉換成 0/1 數值後可供模型使用。

英文答案：Categorical variables are labels representing categories, not usable directly in regression. Dummy variables convert them into binary numerical form (0/1) for modeling.

## 考題 4

中文問題：若 X₁ 和 X₂ 高度相關，應如何處理？

英文問題：If X₁ and X₂ are highly correlated, what can you do?

中文答案：可考慮刪除其中一個變數、合併變數、或使用主成分分析（PCA）降低維度。

英文答案：You may drop one variable, combine them, or use Principal Component Analysis (PCA) to reduce dimensionality.

## 考題 5

中文問題：請說明變數交互作用及其在回歸中的表達方式。

英文問題：Explain interaction effects and how they are expressed in regression.

中文答案：交互作用表示兩變數聯合對 Y 有影響，常見表示法為 X₁×X₂。

英文答案：Interaction effects indicate that two variables jointly affect Y, and are represented as X₁×X₂ in regression models.

## 考題 6

中文問題：比較 Forward Selection 與 Backward Elimination。

英文問題：Compare forward selection and backward elimination.

中文答案：Forward 由空模型逐步加入變數；Backward 由完整模型逐步移除不重要變數。前者較適合高維度資料，後者適合中低維資料。

英文答案：Forward starts with no variables and adds the most significant ones; Backward starts with all and removes the least. Forward suits high-dimensional data, Backward suits simpler data.

## 考題 7

中文問題：什麼是多變量多元回歸？

英文問題：What is multivariate multiple regression?

中文答案：同時預測多個目標變數，並考慮其間的相關性。

英文答案：A regression setting where multiple dependent variables are predicted simultaneously, accounting for correlation between them.

## 考題 8

中文問題：何時需要使用 Dummy Variable？如何建立？

英文問題：When do you need dummy variables? How are they created?

中文答案：當變數為類別型時需使用 Dummy Variable，可用 0 和 1 表示每一類的存在與否。

英文答案：When a categorical variable is used, dummy variables are created to represent categories using 0 and 1 values.

## 考題 9

中文問題：為什麼過度擬合不好？要如何避免？

英文問題：Why is overfitting bad? How can it be prevented?

中文答案：過度擬合會導致模型在訓練資料表現很好，但在新資料表現不佳。可用交叉驗證、簡化模型、正則化等方式避免。

英文答案：Overfitting leads to poor generalization. Prevent it using cross-validation, model simplification, or regularization techniques.

## 考題 10

中文問題：若你發現模型的 Adjusted R² 比 R² 還低，如何解釋？

英文問題：If the Adjusted R² is lower than R², how would you interpret this?

中文答案：這表示加入的變數並未改善模型解釋能力，甚至可能是無效或雜訊變數。

英文答案：This suggests that added variables did not improve model performance and may be irrelevant or noisy.

# 📝 第四章：模擬考題與答案（題目與答案寫在一起）

1. Q1. (a)(b) 題組：多元線性回歸與共線性

(a) What is multiple linear regression? How does it differ from simple linear regression?  
Answer: Multiple linear regression includes multiple predictors and is written as Y = β₀ + β₁X₁ + ... + βₚXₚ + ε. Simple linear regression uses only one predictor.  
（a）何謂多元線性回歸？其與簡單線性回歸有何差異？  
答：多元線性回歸包含多個自變數，形式為 Y = β₀ + β₁X₁ + ... + βₚXₚ + ε。簡單線性回歸僅有一個自變數。  
  
(b) What is multicollinearity? How does it affect the model?  
Answer: Multicollinearity occurs when predictors are highly correlated, causing unstable coefficient estimates, inflated standard errors, and reduced interpretability.  
（b）什麼是共線性？對模型有什麼影響？  
答：共線性指自變數之間高度相關，會導致模型係數估計不穩定、標準誤增加，甚至影響變數解釋力。

1. Q2. (a)(b) 題組：類別變數與交互作用

(a) When do you need dummy variables? How are they created?  
Answer: When a categorical variable is used, dummy variables are created to represent categories using 0 and 1 values.  
（a）何時需要使用 Dummy Variable？如何建立？  
答：當變數為類別型時需使用 Dummy Variable，可用 0 和 1 表示每一類的存在與否。  
  
(b) Explain interaction effects and how they are expressed in regression.  
Answer: Interaction effects indicate that two variables jointly affect Y, and are represented as X₁×X₂ in regression models.  
（b）請說明變數交互作用及其在回歸中的表達方式。  
答：交互作用表示兩變數聯合對 Y 有影響，常見表示法為 X₁×X₂。

1. Q3. (a)(b) 題組：變數選擇方法與嵌入式特徵選擇

(a) Compare forward selection and backward elimination.  
Answer: Forward starts with no variables and adds the most significant ones; Backward starts with all and removes the least. Forward suits high-dimensional data, Backward suits simpler data.  
（a）比較 Forward Selection 與 Backward Elimination。  
答：Forward 由空模型逐步加入變數；Backward 由完整模型逐步移除不重要變數。前者較適合高維度資料，後者適合中低維資料。  
  
(b) What are embedded methods? Give one example.  
Answer: Embedded methods perform feature selection during model training. An example is LASSO regression, which shrinks some coefficients to zero.  
（b）什麼是嵌入式方法？請舉一例。  
答：嵌入式方法在模型訓練時進行特徵選擇，例如 LASSO 會將不重要的係數縮為零。

1. Q4. (a)(b) 題組：過擬合與多變量多元回歸

(a) Why is overfitting bad? How can it be prevented?  
Answer: Overfitting leads to poor generalization. Prevent it using cross-validation, model simplification, or regularization techniques.  
（a）為什麼過度擬合不好？要如何避免？  
答：過度擬合會導致模型在訓練資料表現很好，但在新資料表現不佳。可用交叉驗證、簡化模型、正則化等方式避免。  
  
(b) What is multivariate multiple regression?  
Answer: It predicts multiple dependent variables simultaneously, accounting for correlation among them.  
（b）什麼是多變量多元回歸？  
答：同時預測多個目標變數，並考慮其間的相關性。

1. Q5. (a)(b) 題組：變數交互作用視覺化解釋（圖：interaction\_effect\_plot.png）

(a) Based on the interaction plot, how does x2 affect the influence of x1 on y?  
Answer: When x2 is higher, the slope of x1 on y becomes steeper, indicating a stronger positive interaction effect. This means x1 has a greater effect on y when x2 is high.  
（a）根據交互作用圖，x2 如何影響 x1 對 y 的影響？  
答：當 x2 較高時，x1 對 y 的影響更大，斜率更陡，代表存在正向交互作用，x2 增強了 x1 的影響力。  
  
(b) In practical terms, what could x1 and x2 represent in a marketing scenario?  
Answer: x1 could represent social media ad spending, and x2 could represent TV ad spending. The interaction suggests their combined use may lead to a greater increase in sales than using either alone.  
（b）在行銷場景中，x1 和 x2 可以代表什麼？  
答：x1 可代表社群廣告費用，x2 代表電視廣告支出。交互作用顯示兩者一起使用可比單獨使用任一種更有效提升銷售。

圖 Q5：Interaction Effect between x1 and x2

