機器學習第三章：完整筆記與模擬出題（未壓縮）

本文件包含線性回歸與模型準確度評估的名詞解釋、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

# 📘 第三章：線性回歸與模型準確度評估（Linear Regression and Accuracy Assessment）

## 🧠 名詞解釋（Term Explanation）

### 線性回歸（Linear Regression）

中文：線性回歸是一種監督式學習方法，用於預測連續變數。其假設自變數 X 與因變數 Y 之間存在線性關係。

英文：Linear regression is a supervised learning method used to predict a continuous variable, assuming a linear relationship between the input X and output Y.

### 最小平方法（Least Squares Method）

中文：一種估計線性模型參數的方法，透過最小化預測值與實際值之間平方差來找到最佳線。

英文：A method for estimating the parameters of a linear model by minimizing the sum of squared differences between the observed and predicted values.

### 係數估計值（Coefficient Estimates）

中文：線性回歸中，每個 βj 表示對應變數的邊際影響，即當 xj 增加一單位時 y 的變化。

英文：In linear regression, each coefficient βj represents the marginal effect of variable xj on y, i.e., the expected change in y when xj increases by one unit.

### 殘差（Residual）

中文：實際值與預測值之間的差距，即 Residual = yi - ŷi。

英文：The difference between the observed value and the predicted value, defined as yi - ŷi.

### 標準誤（Standard Error, SE）

中文：表示係數估計的不確定性，越小代表估計越穩定。

英文：Represents the uncertainty in coefficient estimates. A smaller SE indicates more precise estimation.

### t-統計量（t-Statistic）

中文：用於假設檢定的統計量，計算方式為：t = β / SE(β)。

英文：A test statistic used in hypothesis testing, computed as t = β / SE(β).

### p-值（p-value）

中文：在虛無假設（如 β = 0）下，觀察到比目前更極端的統計量的機率，常用於決定變數是否顯著。

英文：The probability of obtaining a test statistic at least as extreme as the one observed under the null hypothesis; used to assess variable significance.

### R平方（R-squared, R²）

中文：衡量模型對資料的擬合程度，範圍為 0 到 1，值越高代表模型解釋越多變異。

英文：Measures the proportion of variance in the dependent variable that is explained by the independent variables. Ranges from 0 to 1.

### 調整後R平方（Adjusted R-squared）

中文：在考慮變數數量後對 R² 進行修正，用於多變量模型的比較。

英文：Adjusted version of R² that penalizes for adding irrelevant predictors; useful for comparing models with different numbers of variables.

### F-統計量（F-Statistic）

中文：檢驗整體模型是否具有預測能力的統計量，用來比較模型是否比僅使用截距更好。

英文：A statistic that tests whether the overall regression model provides a better fit than a model with no predictors (intercept-only model).

### 信賴區間（Confidence Interval）

中文：某一參數的估計值落在特定區間內的機率，例如「95%信賴區間」表示有95%的信心該參數落在此範圍內。

英文：A range of values within which the true parameter is expected to lie with a specified level of confidence, such as 95%.

### 假設檢定（Hypothesis Testing）

中文：檢驗某參數是否為特定值（通常是 0）的方法，結合 t 統計與 p 值進行決策。

英文：A method to determine whether a model parameter is equal to a hypothesized value (often 0), using the t-statistic and p-value.

## 📚 名詞比較（Term Comparison）

t-test vs. F-test

|  |  |  |
| --- | --- | --- |
| 項目 | t-統計檢定（t-test） | F-統計檢定（F-test） |
| 測試對象 | 單一係數是否為零 | 所有變數是否同時為零 |
| 適用情境 | 變數單獨的重要性 | 整體模型是否有效 |
| 統計量 | t = β / SE(β) | 比較 SSE 之間的比例 |
| 解釋力 | 單一變數是否顯著 | 整體模型是否顯著 |

R-squared vs. Adjusted R-squared

|  |  |  |
| --- | --- | --- |
| 項目 | R-squared | 調整後 R-squared |
| 定義 | 模型解釋的變異比例 | 修正後的 R²，考慮變數數量 |
| 是否變數越多越大 | 是（即使無效變數） | 不一定，可能會下降 |
| 評估功能 | 衡量擬合程度 | 比較不同模型複雜度 |

# 📝 第三章模擬出題（Chapter 3 Mock Questions）

## 考題 1：名詞解釋 Term Explanation

中： 何謂線性回歸？其基本假設為何？

EN: What is linear regression? What are its basic assumptions?

答：  
線性回歸是預測連續目標變數的一種方法，假設因變數與自變數間為線性關係。基本假設包含：線性、誤差常態分布、同方差性與獨立性。  
Linear regression is a method for predicting a continuous target variable, assuming a linear relationship between dependent and independent variables. Basic assumptions include linearity, normally distributed errors, homoscedasticity, and independence.

## 考題 2：名詞解釋 Term Explanation

中： 何謂最小平方法？其目標為何？

EN: What is the least squares method? What is its goal?

答：  
最小平方法透過最小化預測值與實際值的平方誤差總和來估計模型參數，目標是找到最佳擬合線。  
The least squares method estimates model parameters by minimizing the sum of squared differences between predicted and actual values, aiming to find the best fit line.

## 考題 3：名詞比較 Term Comparison

中： 比較 t 檢定與 F 檢定在回歸分析中的差異。

EN: Compare t-test and F-test in the context of regression analysis.

答：  
t 檢定檢視單一變數是否顯著，F 檢定則評估整體模型是否有效。前者用於變數選擇，後者用於模型整體評估。  
t-tests assess the significance of individual variables, while F-tests evaluate whether the regression model as a whole is meaningful.

## 考題 4：名詞解釋 Term Explanation

中： 什麼是 R-squared？調整後 R-squared 有什麼不同？

EN: What is R-squared? How does Adjusted R-squared differ?

答：  
R-squared 測量模型解釋資料變異的能力；調整後 R-squared 修正變數數量對結果的影響，適合比較不同複雜度的模型。  
R-squared measures how much variance is explained by the model; Adjusted R-squared accounts for the number of variables, making it better for comparing models.

## 考題 5：應用題 Application

中： 當模型中的某個變數 t 值小、p 值大時，該如何解釋？

EN: What does it imply if a variable in the model has a small t-value and a large p-value?

答：  
表示該變數在模型中可能不顯著，無法拒絕其係數為零的虛無假設，可能考慮剃除該變數。  
It suggests the variable is not statistically significant, failing to reject the null hypothesis that the coefficient is zero, and might be excluded from the model.

## 考題 6：應用題 Application

中： 若一個模型的 R-squared 很高但 p 值皆不顯著，可能原因為何？

EN: If a model has a high R-squared but non-significant p-values, what could be the cause?

答：  
可能存在共線性，導致模型整體看似良好但個別變數不顯著。  
Multicollinearity may exist, making the overall model seem good while individual predictors appear insignificant.

## 考題 7：計算題 Calculation

中： 若估計係數為 2，標準誤為 0.5，請計算其 t 值。

EN: If the estimated coefficient is 2 and standard error is 0.5, calculate the t-value.

答：  
t = 2 / 0.5 = 4  
 t = 2 / 0.5 = 4

## 考題 8：名詞解釋 Term Explanation

中： 何謂信賴區間？其與標準誤有何關係？

EN: What is a confidence interval? How is it related to standard error?

答：  
信賴區間表示參數的可能落點範圍，通常為 β̂ ± 1.96 × SE（95% CI）。  
A confidence interval gives the likely range of a parameter, typically calculated as β̂ ± 1.96 × SE for a 95% confidence level.

## 考題 9：名詞解釋 Term Explanation

中： 殘差是什麼？如何評估模型是否有異常殘差？

EN: What is a residual? How can you assess if the model has unusual residuals?

答：  
殘差是實際值與預測值之間的差異；可利用殘差圖檢查是否有系統性趨勢或極端值。  
Residuals are the differences between observed and predicted values; residual plots help identify patterns or outliers.

## 考題 10：綜合題 Integration

中： 若 F 統計顯著但 Adjusted R-squared 僅 0.1，該模型是否可靠？

EN: If the F-statistic is significant but the Adjusted R-squared is only 0.1, is the model reliable?

答：  
雖有統計顯著性，但解釋力有限，模型預測效果差，實務上不一定可接受。  
Despite statistical significance, the low Adjusted R-squared implies weak explanatory power, which may not be acceptable in practice.

# 📝 第三章：模擬考題與答案（題目與答案寫在一起）

1. Q1. (a)(b) 題組：線性回歸與基本假設

(a) What is linear regression? What are its basic assumptions?  
Answer: Linear regression is a method for predicting a continuous target variable, assuming a linear relationship between dependent and independent variables. Basic assumptions include linearity, normally distributed errors, homoscedasticity, and independence.  
（a）何謂線性回歸？其基本假設為何？  
答：線性回歸是預測連續目標變數的一種方法，假設因變數與自變數間為線性關係。基本假設包含：線性、誤差常態分布、同方差性與獨立性。  
  
(b) What is the least squares method? What is its goal?  
Answer: The least squares method estimates model parameters by minimizing the sum of squared differences between predicted and actual values, aiming to find the best fit line.  
（b）何謂最小平方法？其目標為何？  
答：最小平方法透過最小化預測值與實際值的平方誤差總和來估計模型參數，目標是找到最佳擬合線。

1. Q2. (a)(b) 題組：模型評估與顯著性檢定

(a) Compare t-test and F-test in the context of regression analysis.  
Answer: t-tests assess the significance of individual variables, while F-tests evaluate whether the regression model as a whole is meaningful.  
（a）比較 t 檢定與 F 檢定在回歸分析中的差異。  
答：t 檢定檢視單一變數是否顯著，F 檢定則評估整體模型是否有效。前者用於變數選擇，後者用於模型整體評估。  
  
(b) What is R-squared? How does Adjusted R-squared differ?  
Answer: R-squared measures how much variance is explained by the model; Adjusted R-squared accounts for the number of variables, making it better for comparing models.  
（b）什麼是 R-squared？調整後 R-squared 有什麼不同？  
答：R-squared 衡量模型解釋資料變異的能力；調整後 R-squared 修正變數數量對結果的影響，適合比較不同複雜度的模型。

1. Q3. (a)(b) 題組：標準誤與信賴區間

(a) What is standard error? How is it interpreted?  
Answer: Standard error measures the uncertainty in the coefficient estimates. Smaller SE indicates more stable and reliable estimates.  
（a）什麼是標準誤？其意義為何？  
答：標準誤衡量係數估計的不確定性，數值越小代表估計越穩定與可靠。  
  
(b) What is a confidence interval and how is it calculated?  
Answer: A confidence interval gives the likely range of a parameter. For 95% CI, it is calculated as β̂ ± 1.96 × SE.  
（b）何謂信賴區間？如何計算？  
答：信賴區間表示參數的可能落點範圍，95% 信賴區間可由 β̂ ± 1.96 × SE 得出。

1. Q4. (a)(b) 題組：模型整體評估與殘差解釋

(a) What is the meaning of a high R-squared with all p-values insignificant?  
Answer: This may indicate multicollinearity—variables are correlated with each other, making individual effects indistinguishable.  
（a）若 R-squared 很高但 p 值皆不顯著，可能原因為何？  
答：可能存在共線性，導致模型整體表現良好但各變數個別影響難以區分。  
  
(b) What is a residual? How do we identify unusual residuals?  
Answer: A residual is the difference between the actual and predicted value. Residual plots help identify patterns or outliers.  
（b）什麼是殘差？如何判斷是否異常？  
答：殘差是實際值與預測值的差距，可利用殘差圖檢查系統性趨勢或極端值。

1. Q5. (a)(b) 題組：殘差圖判讀與模型診斷（圖：residual\_plot.png）

(a) Based on the residual plot, how can you tell if a linear model is appropriate for the data?  
Answer: If the residuals are randomly scattered around zero with no clear pattern, the linear model is likely appropriate. Systematic patterns suggest model misspecification.  
（a）根據殘差圖，如何判斷線性模型是否適用？  
答：若殘差在0附近隨機分布、無明顯圖樣，代表線性模型適用；若殘差呈現系統性趨勢，可能表示模型設計不良。  
  
(b) What potential issues can you identify from this plot?  
Answer: If the residuals display a fan shape or curve, it could imply heteroscedasticity or a nonlinear relationship.  
（b）從圖中可以觀察出哪些潛在問題？  
答：若殘差呈現扇形或曲線分布，可能代表誤差變異不等（異質性）或資料存在非線性關係。

圖 Q5：Residual Plot

