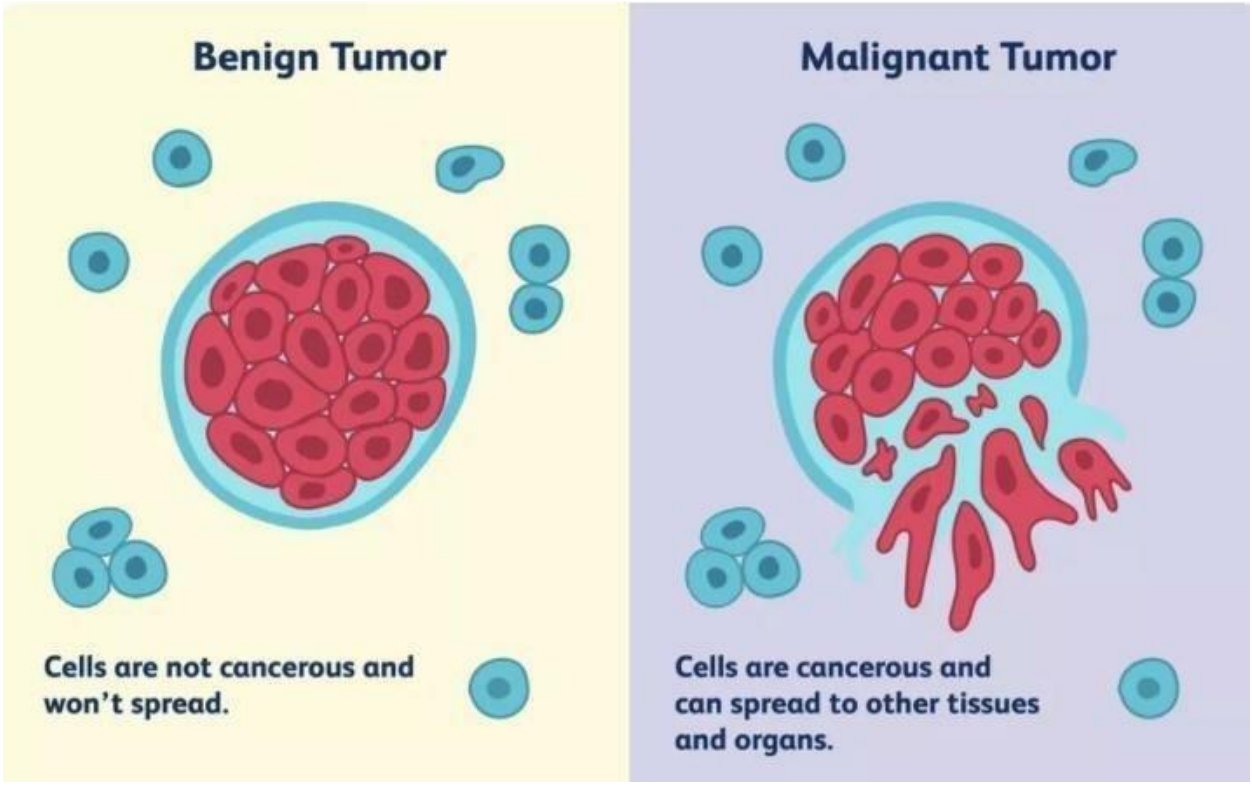

威斯康辛州乳癌（診斷）

劉羿暄

目錄

第壹章　緒論

第一節 前言

乳癌是女性中最常見的癌症之一，且每年新增的乳癌病例數量不斷增加，發病年齡也有下降的趨勢，其複雜多變的病理特徵和多樣化的亞型進一步增加了診斷和治療的難度，使乳癌成為高發病率和死亡率的癌症。因此早期檢測和預測具有重要意義，傳統的診斷方法主要包括臨床檢查、影像學檢查（如:乳腺 X 光、超聲波和核磁共振）以及病理學檢查。隨著基因分析技術的應用的進步，乳癌的診斷和分類得到了顯著改進。這些技術不僅可以精確區分乳癌的不同亞型，還可以預測患者對某些治療的反應，從而實現個體化治療。

在眾多診斷方法中，乳房攝影檢查是最基本且廣泛應用的檢查工具，能夠及早發現乳腺中的微小病變，以提高早期乳癌的檢出率，從而顯著降低死亡率。在台灣，各大醫療機構和公共衛生組織均建議，特別是 45 歲以上的女性，應每兩年進行一次乳房攝影檢查。影像學檢查在乳癌診斷中的優勢在於其早期發現能力、無創性、高準確性和廣泛的應用範圍，這些特點使其成為乳癌篩查和診斷的首選工具。故本研究將探索威斯康辛州的乳癌數據集，聚焦於影像學檢查的腫瘤特徵，建立一個邏輯迴歸模型，並進行預測。

第二節 動機與目的

乳癌的早期檢測可以極大地提高治療成功率和生存率，若能透過模型輔助醫療診斷，能使醫療人員在早期檢測和治療做出更準確的決策和制定更為精確的治療計劃。本研究旨在找出腫瘤外部特徵與癌症的關係，並嘗試利用邏輯迴歸建立一個有效的分類器希望能提供醫學界更好的乳癌診斷工具，從而改善患者的預後能力。

第三節 研究資料

本研究使用 UCI 資料庫 1993 年威斯康辛州的乳癌數據集，從中保留診斷及 6 數值型特徵作為研究變數，包含：半徑、紋理、周長、面積、平滑度、凹點。利用 R 程式隨機抽取 140 筆作為訓練集，再從剩餘資料中抽取 60 筆做為測試集，共 200 筆資料以 7:3 分為兩份做分析及預測。

第四節 研究方法

首先，由訓練資料建模，分析其基本資料，診斷多元共線性，並刪除不合適的變數。隨後，檢測每個自變數對乳癌發生的影響，並透過選取法篩選變數，以提高模型簡潔性，得到最佳模型。最後，將此模型套入測試資料中，檢驗其預測能力。

第貳章　基本資料分析

第一節 基本敘述統計量

表 2-1 簡單統計值

變數	標籤	N	平均值	標準差	最小值	最大值
x1	半徑	140	14.2705214	3.8335434	7.691	27.22
x2	紋理	140	19.6798571	4.310632	10.38	32.47
x3	周長	140	93.1491429	26.645315	47.92	182.1
x4	面積	140	672.75	383.8285826	170.4	2250
x5	平滑度	140	0.0990366	0.0163256	0.05263	0.1634
x6	凹點	140	0.0548212	0.0445963	0	0.2012

註：所有數據皆以每個細胞核計算十個實值特徵

一、診斷(Y)

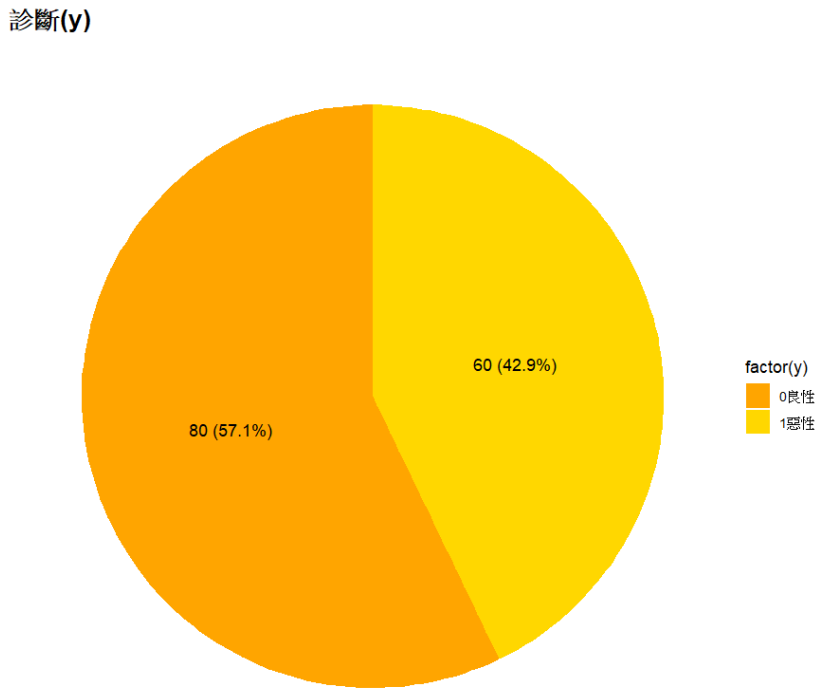

圖 2-1　診斷結果

由圖 2-1 可以了解到 140 個乳腺腫瘤樣本的診斷情形，其中良性有 80 個樣本，占總樣本數的 57.1%；惡性有 60 個樣本，占總樣本數的 42.9%。

二、半徑(X_1)

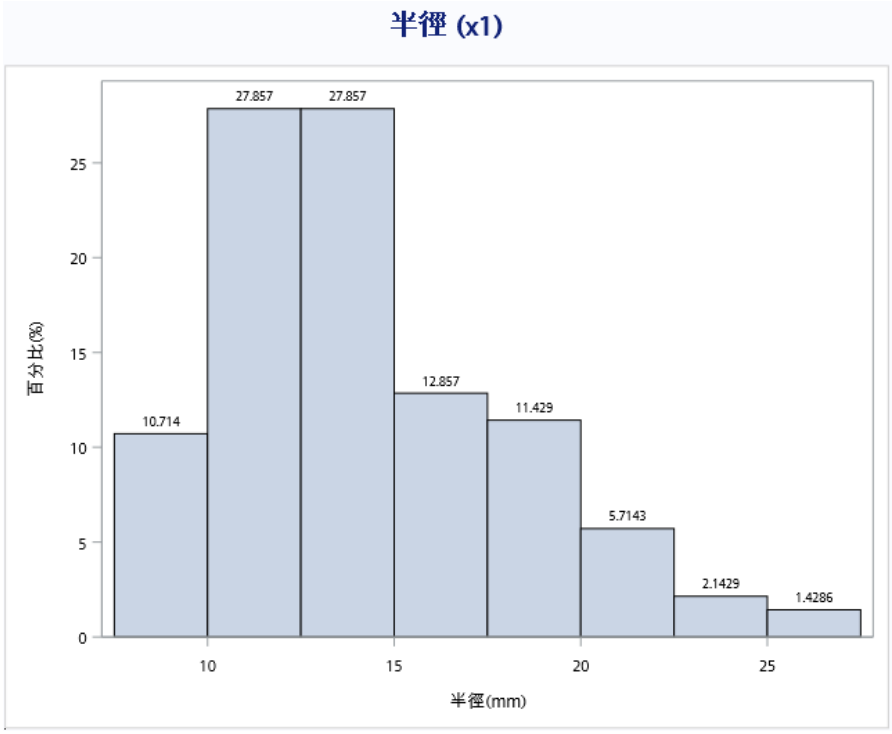

圖 2-2 半徑相對直方圖

半徑為腫瘤中心到週邊點距離的平均值，由表 2-1 可知最大值為 27.22，最小值為 7.691。由圖 2-2 可見數據呈現右偏分布，大部分樣本的半徑集中在 10 到 15 毫米之間，占總樣本的 55.741%。

三、紋理(X_2)

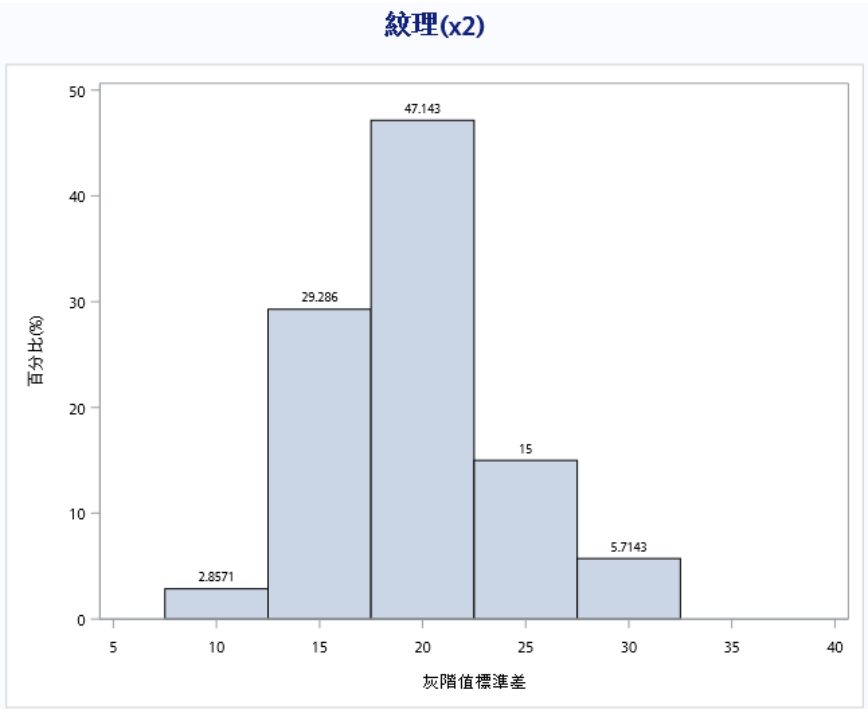

圖 2-3 紋理相對直方圖

灰階值的標準差(Standard Deviation of Gray Levels 本研究以紋理代稱)，是一種衡量圖像中灰階值分佈的離散程度的指標，反映了圖像的對比度和紋理複雜度，常用於腫瘤影像的特徵提取和分類，有助於描述圖像的紋理特性及區分不同的組織類型和病變狀況。標準差越大，表示圖像中灰階值的變化越大，紋理越複雜，可能是惡性腫瘤的特徵之一；標準差越小，表示圖像中灰階值的變化越小，紋理越均勻。計算公式如下：

$$\text{灰階值標準差} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - \mu)^2}$$

其中，μ 是灰階值的平均數，N 是像素數量，p_i是第 i 像素的灰階值。

灰階值：

數字越小，亮度越低：0 通常表示黑色。

數字越大，亮度越高：最大值（例如，8 位圖像中的 255）通常表示白色。

由表 2-1 可知最大值為 32.47，最小值為 10.38。由圖 2-3 可見，紋理呈現右偏分布，大部分樣本集中在 20 左右，占總樣的 47.143%。

四、周長(X_3)

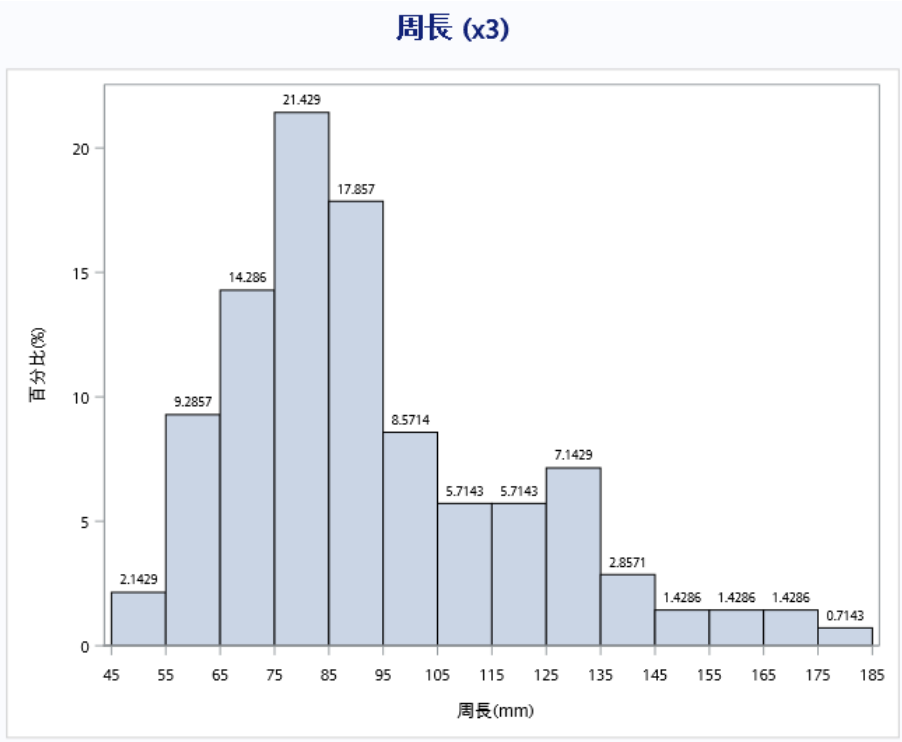

圖 2-4 周長的相對直方圖

周長為核心腫瘤的平均大小，由表 2-1 可知最大值為 182.1，最小值為 47.92。由圖 2-4 可見，周長呈現右偏分布，大部分樣本集中在 75 到 85 毫米，占總樣本的 21.429%。

五、面積(X_4)

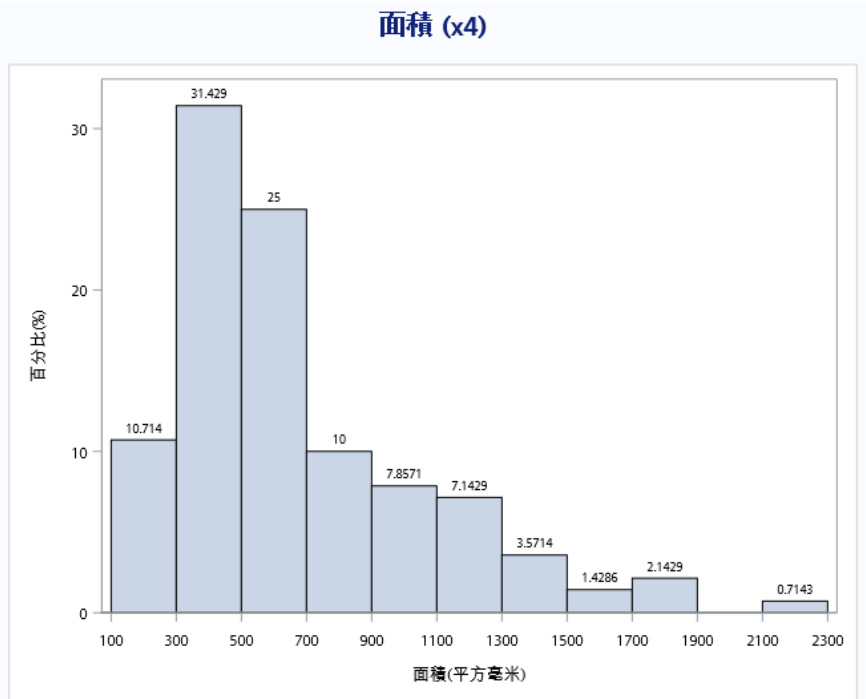

圖 2-5 面積平均值

面積為核心腫瘤的平均面積，由表 2-1 可知最大值為 2250，最小值為 170.4。由圖 2-5 可見，面積呈現右偏分布，大部分樣本集中在 300 到 500 平方毫米，占總樣本的 31.429%；其次為 500 到 700 平方毫米區間，占總樣本的 25%。

六、平滑度($\mathbf{X_5}$)

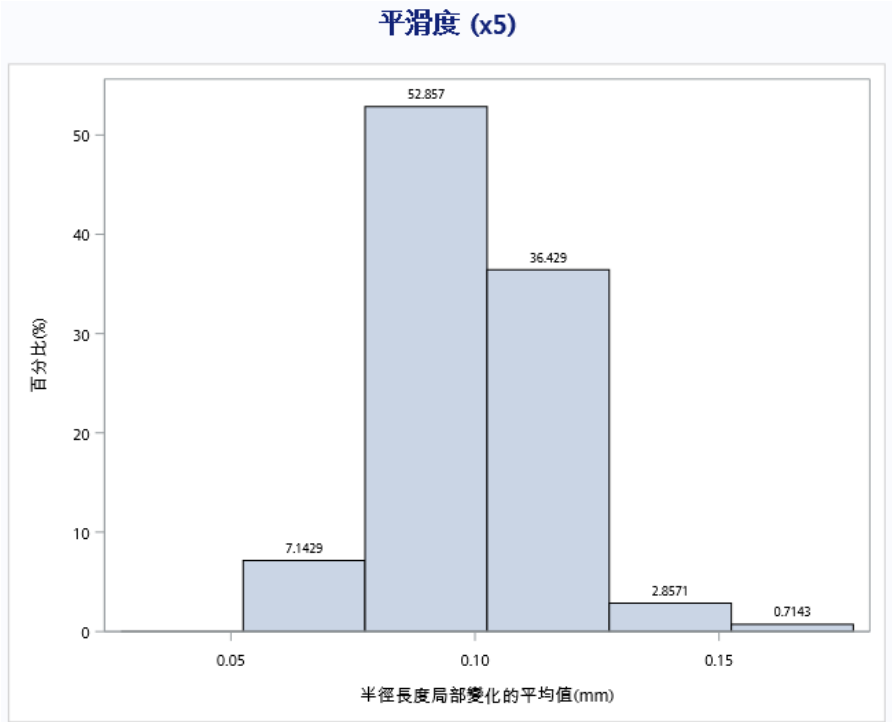

圖 2-6 平滑度相對直方圖

平滑度通常定義為腫瘤邊界上半徑長度的局部變化，邊界越平滑，變化越小。

具體的計算方式如下：

$$\text{平滑度} = \frac{1}{N}\sum_{i=1}^{N}|R_i - R_{i+1}|$$

每個邊界點的半徑R_i：從腫瘤的質心（中心點）到每個邊界點的距離

N 是邊界上的點數

由表 2-1 可知最大值為 0.1634，最小值為 0.05263。由圖 2-6 可見，大部分樣本集中 0.075 到 0.125 毫米。

七、凹點(X_6)

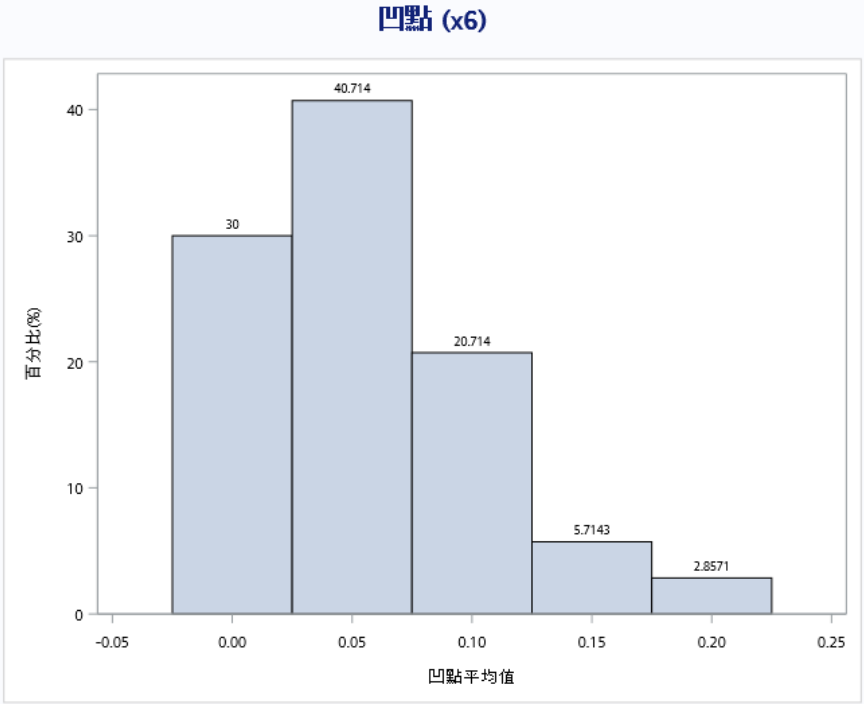

圖 2-7 凹點相對直方圖

輪廓凹部數量的平均值(Concave points mean 本研究簡稱凹點）是描述腫瘤邊緣凹點數量的平均值，用於量化腫瘤形狀的複雜度。此指標可以深入了解腫瘤的形狀不規則性，較高的數值通常表示惡性腫瘤，在腫瘤的形態學分析中具有重要意義。

由表 2-1 可知最大值為 0.2012，最小值為 0。由圖 2-7 可見，凹點呈現右偏分布，大部分樣本集中 0.05 左右，占總樣本的 40.714%。

第二節 Variance Inflation Factor(VIF)

變異數膨脹因子(VIF)為診斷多元共線性嚴重程度的指標。

表 2-2 各變異數的 VIF

變數	標籤	變異數膨脹
Intercept	Intercept	0
X1	半徑	1267.632
X2	紋理	1.842
X3	周長	939.271
X4	面積	105.579
X5	平滑度	4.685
X6	凹點	4.391

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2} \ , i = 1,2, \ldots ,6$$

觀察自變數 X_i的模型時，複判定係數R_i^2反映了X_i與其他自變數的相依程度。R_i^2愈大，VIF 值愈大，可能導致參數估計不穩定，增加係數估計的變異性。通常，VIF 值大於 10 被視為存在共線性的警告信號。

根據表 2-2，有三個變數的 VIF 大於 10，其中以X_1最大，故刪除變數後再次檢驗。

表 2-3 各變異數的 VIF

變數	標籤	變異數膨脹
Intercept	Intercept	0
X2	紋理	1.941
X3	周長	67.036
X4	面積	67.597
X5	平滑度	5.616
X6	凹點	3.680

根據表 2-3，有兩個變數的 VIF 大於 10，其中以X_4最大，故刪除變數後再次檢驗。

表 2-4 各變異數的 VIF

變數	標籤	變異數膨脹
Intercept	Intercept	0
X2	紋理	1.726
X3	周長	2.908
X5	平滑度	5.530
X6	凹點	3.343

根據表 2-4，所有變數的 VIF 皆小於 10，保留$\mathrm{X}_2, \mathrm{X}_3, \mathrm{X}_5, \mathrm{X}_6$ 四個變數建立模型。

第參章　原始模型檢定

第一節 建立邏輯迴歸模型

為了解影響乳癌的因素，我們令X_{2i}為紋理、X_{3i}為周長、X_{5i}為平滑度、X_{6i}為凹點建立原始模型：

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \beta_6 X_{6i}$$

$$i = 1,2,\ldots,140$$

其中p_i是第i個觀測值為惡性的機率，Y_i=0 表示良性，Y_i=1 表示惡性。

用 R 程式建立邏輯迴歸模型：

表 3-1 參數估計

	Estimate	Std.Error	Z	Pr(>|Z|)
(Intercept)	-28.67824	9.83959	-2.915	0.003562
X2	0.40803q	0.12179	3.35	0.000807
X3	0.12722	0.05068	2.51	0.012073
X5	-66.84705	53.69137	1.245	0.213123
X6	43.86848	33.08651	1.426	0.183882

由表 3-1，可得：

$\hat{\beta}_0$=－28.67824　$\hat{\beta}_2$=0.4083　$\hat{\beta}_3$=0.12722　$\hat{\beta}_5$=－66.84705　$\hat{\beta}_6$= 43.8648

原始模型：

$$\text{logit}[P(Y=1)] = -28.67824 + 0.40803X_2 + 0.12722X_3 - 66.84705X_5 + 43.8648X_6$$

第二節 單一參數 Wald test

得知β值之後，接著判斷各解釋變 X_2(紋理)、X_3(周長)、X_5(平滑度)、X_6(凹點)與 Y(診斷)是否存在顯著關係。

一、β_2之 Wald test

欲了解 X_2(紋理)與 Y(診斷)是否存在顯著關係，假定其他變數為固定的情況下假設：

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0$$

在顯著水準 α=0.05 下，檢定結果如下：

因為 p-value = 0.000807 < α=0.05，因此拒絕 H_0 的假設，表示有充分證據顯示 $\beta_2 \neq 0$，即 X_2(紋理) 與 Y(診斷) 有顯著相關。

二、β_3之 Wald test

欲了解 X_3(周長)與 Y(診斷)是否存在顯著關係，假定其他變數為固定的情況下假設：

$$H_0 : \beta_3 = 0$$

$$H_1 : \beta_3 \neq 0$$

在顯著水準 α=0.05 下，檢定結果如下：

因為 p-value = 0.012073 < α=0.05，因此拒絕 H_0 的假設，表示有充分證據顯示 $\beta_3 \neq 0$，即 X_3 (紋理) 與 Y (診斷) 有顯著相關。

三、β_5 之 Wald test

欲了解 X_5 (平滑度)與 Y (診斷)是否存在顯著關係，假定其他變數為固定的情況下假設：

$$H_0 : \beta_5 = 0$$

$$H_1 : \beta_5 \neq 0$$

在顯著水準 α=0.05 下，檢定結果如下：

因為 p-value = 0.213123 > α=0.05，因此不拒絕虛無假設，表示沒有充分證據顯示 $\beta_5 \neq 0$，即 X_5 (平滑度) 與 Y (診斷) 沒有顯著相關。

四、β_6 之 Wald test

欲了解 X_6 (凹點)與 Y (診斷)是否存在顯著關係，假定其他變數為固定的情況下假設：

$$H_0 : \beta_6 = 0$$

$$H_1 : \beta_6 \neq 0$$

在顯著水準 α=0.05 下，檢定結果如下：

因為 p-value = 0.184882 > α=0.05，因此拒絕虛無假設，表示沒有充分證據顯示 $\beta_6 \neq 0$，即 X_6 (凹點) 與 Y (診斷) 沒有顯著相關。

第肆章 模型的選取方法

第一節 向前選取法(Forward)

逐個將每個變數添加到邏輯回歸模型中，並分別估計迴歸係數。我們優先選擇能夠最大程度提升模型性能的特徵，即具有最小 p-value 的特徵先進入模型，這有助於提高模型的簡潔性並確保模型的解釋性。

R 程式結果如下:

```
Start:  AIC=60.84
y ~ x2 + x3 + x5 + x6
Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -28.67824     9.83959   -2.915  0.003562  **
x2            0.40803     0.12179    3.350  0.000807  ***
x3            0.12722     0.05068    2.510  0.012073  *
x5           66.84705    53.69137    1.245  0.213123
x6           43.86848    33.08651    1.326  0.184882
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

由向前選取法可得其邏輯迴歸模型為:

$$\text{logit}[P(Y=1)] = -28.67824 + 0.40803\,X_2 + 0.12722\,X_3 + 66.84705\,X_5 + 43.86848\,X_6$$

第二節 後退刪去法(Backward)

與向前選取法相反，後退刪去法是從包含所有特徵的模型開始，然後逐步刪除對模型性能影響最小的特徵。能夠快速減少特徵數量，提高模型的簡潔性和解釋性。

R 程式結果如下:

```
Start:  AIC=60.84
y ~ x2 + x3 + x5 + x6

       Df Deviance    AIC
- x5    1   52.340 60.340
<none>      50.842 60.842
- x6    1   52.852 60.852
- x3    1   59.478 67.478
- x2    1   69.092 77.092
```

Step1: AIC=60.34

y ~ x2 + x3 + x6

```
       Df Deviance    AIC
<none>      52.340 60.340
- x3    1   61.789 67.789
- x2    1   69.180 75.180
- x6    1   81.405 87.405
> summary(step_backward)

Call:
glm(formula = y ~ x2 + x3 + x6, family = binomial, data = train1)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -18.55594    3.90677  -4.750 2.04e-06 ***
x2            0.34918    0.10174   3.432 0.000599 ***
x3            0.08270    0.02984   2.771 0.005585 **
x6           79.76518   20.08365   3.972 7.14e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

首先評估了包含所有變數的起始模型的 AIC 值，然後逐步地刪除一個變數，直到無法再進一步改善模型的性能。第一步刪除了變數 X_5，結果發現刪除 X_5 後的模型的 AIC 從 60.84 降至 60.34，表示刪除 X_5 後的模型在 AIC 準則下更好，因此刪除 X_5。

由後退刪去法選取的解釋變數建立的邏輯迴歸模型如下：

$$\text{logit}[P(Y=1)] = -18.55594 + 0.34918\ X_2 + 0.0827\ X_3 + 79.76518\ X_6$$

第三節 逐步迴歸法

結合向前選取法和後退刪去法的特徵選取方法，逐步地選擇和排除特徵，以優化模型的性能。

R 程式結果如下:

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-18.55594	3.90677	-4.750	2.04e-06	***
x2	0.34918	0.10174	3.432	0.000599	***
x3	0.08270	0.02984	2.771	0.005585	**
x6	79.76518	20.08365	3.972	7.14e-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

由逐步迴歸法選取的解釋變數建立的邏輯迴歸模型如下：

$$\text{logit}[P(Y=1)] = -18.55594 + 0.34918\ X_2 + 0.0827\ X_3 + 79.76518\ X_6$$

第四節 結論

表 4-1 模型選取方法與結論

方法	選擇之變數
向前選取法	X_2、 X_3、X_5、X_6
後退刪去法	X_2、 X_3、X_6
逐步迴歸法	X_2、X_3、X_6

綜合以上各種檢定方法，可知我們應剔除X_5來進行分析，以求得最佳模型：

$$\text{logit}[P(Y=1)] = -18.55594 + 0.34918\,X_2 + 0.0827\,X_3 + 79.76518\,X_6$$

第伍章 模型確認

模型預測能力

用最佳模型預測測試集資料，預測能力如下:

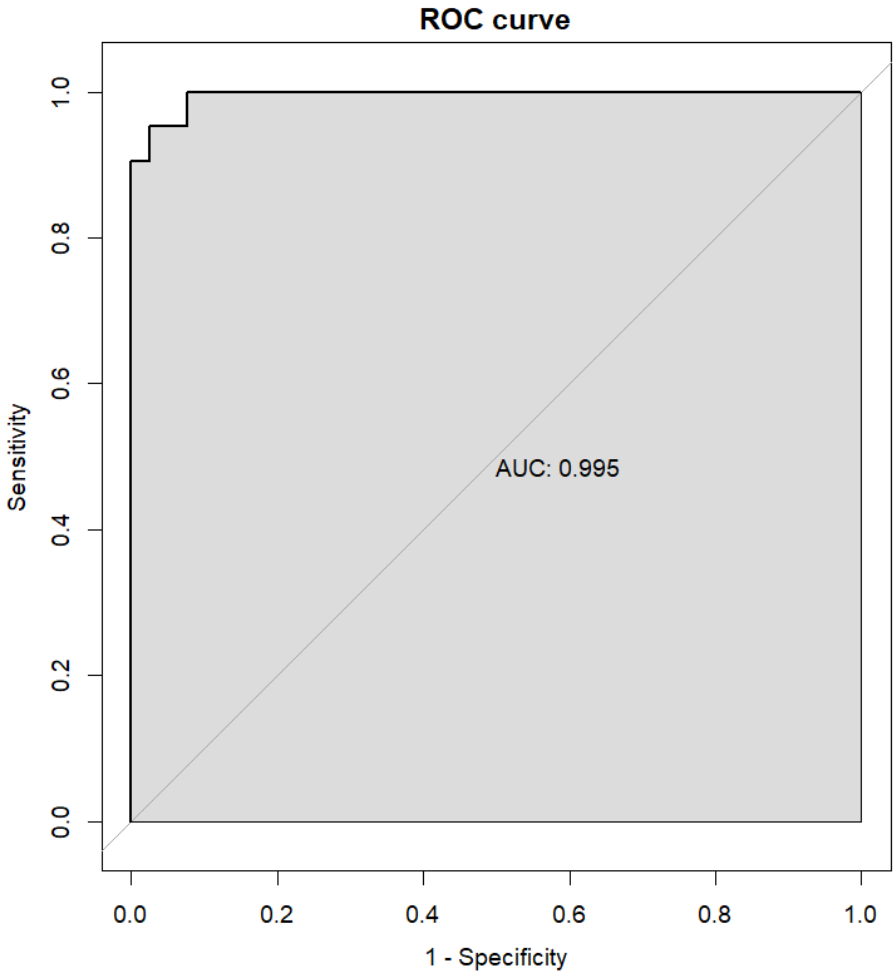

圖 5-1 ROC curve

ROC 曲線呈現出真陽性率與偽陽性率之間的關係，其中座標上(0,1)代表最好的分類情況，即完全預測正確的情況，因此越靠近 ROC 曲線的左上方，代表模型預測結果越傑出。

- AUC ＝1 時，為最理想的情況，分類器做完美的預測。
- AUC ＞0.5 時， 分類器效果比隨機猜測的結果還理想，預測有效果。

而此模型之 AUC=0.995，預測能力良好，模型能夠有效地區分良性和惡性腫瘤。

第陸章 附錄

一、資料來源

https://www.kaggle.com/code/jagannathrk/predicting-breast-cancer-logistic-regression/input
https://www.researchgate.net/figure/Illustrations-of-Cancerous-Non-Cancerous-Cells-A-B-C_fig1_366929515

二、程式碼

(SAS)

```
proc means data=work.train(drop=y);
run;
proc sgplot data=work.train;
histogram x1/  datalabel;
title "半徑 (x1)";
xaxis label="半徑(mm)" ;
 yaxis label="百分比(%)" ;
run;
proc sgplot data=work.train
histogram x2/binwidth=5 datalabel;
title "紋理(x2)";
xaxis label="灰階值標準差" values=(5 to 40 by 5);
 yaxis label="百分比(%)" ;
run;
proc sgplot data=work.train;
histogram x3/binwidth=10   datalabel;
title "周長 (x3)";
xaxis label="周長(mm)" values=(45 to 185 by 10);
 yaxis label="百分比(%)" ;
run;
proc sgplot data=work.train;
histogram x4/binwidth=200    datalabel;
title "面積 (x4)";
```

```
xaxis label="面積(平方毫米)" values=(100 to 2300 by 200);
 yaxis label="百分比(%)" ;
run;
proc sgplot data=work.train;
histogram x5/binwidth=0.025  datalabel;
title "平滑度 (x5)";
xaxis label="半徑長度局部變化的平均值(mm)" ;
 yaxis label="百分比(%)" ;
run;
proc sgplot data=work.train;
histogram x6/binwidth=0.05 datalabel;
title "凹點 (x6)";
xaxis label= "凹點平均值" values=(-0.05 to 0.25 by 0.05);
 yaxis label="百分比(%)" ;
run;
```

(R)

```
pacman::p_load(ggplot2, dplyr,
tidyverse,readxl,MASS,stats,rJava,glmulti,glmnet,car,pROC,DescTools)
set.seed(666)

cost <- read_excel("C:\\Users\\niuno\\Desktop\\類別\\breast_cancer.xlsx")

train <- cost %>%
   sample_n(140, replace = FALSE)
# 未被抽取的數據
remaining_data <- anti_join(cost, train ,by='id')
# 再從未被抽取的數據中抽取 60 筆
test<- remaining_data %>% sample_n(60, replace = FALSE)
train1 <- train[, 2:8]
test1 <- test[, 2:8]
# 匯出抽取的 data 到 Excel
#library(openxlsx)
#write.xlsx(train1, file = "C:\\Users\\niuno\\Desktop\\類別\\train.xlsx", rowNames= F)
#write.xlsx(test1, file =   "C:\\Users\\niuno\\Desktop\\類別\\test.xlsx", rowNames= F)

freq_y <- train1 %>%
```

```
  count(y) %>%
  mutate(percentage = n / sum(n) * 100)  # 計算百分比

ggplot(freq_y, aes(x = "", y = n, fill = factor(y))) +
  geom_bar(width = 1, stat = "identity") +
  geom_text(aes(label = paste0(n, " (", sprintf("%.1f", percentage), "%)")),
            position = position_stack(vjust = 0.5), size = 4) +
  coord_polar(theta = "y") +
  labs(title = "診斷(y)", x = NULL, y = NULL) +
  scale_fill_manual(values = c("#ffa500", "#ffd700"),
                    labels = c("0 良性", "1 惡性")) +
  theme_void() +
  theme(plot.title = element_text(size =16, face = "bold"))


# Fit initial null model (intercept only)
null_model <- glm(y ~ 1, data = train1, family = binomial)

# Full model with all predictors
full_model <- glm(y ~ x1+x2+x3+x4+x5+x6, data = train1, family = binomial)
# 顯示模型摘要
summary(full_model)


# 計算變異數膨脹因子 VIF
vif_values <- vif(full_model)
print(vif_values)

# Full model with all predictors
train_model <- glm(y ~ x2+x3+x5+x6, data = train1, family = binomial)
# 顯示模型摘要
summary(train_model)

# Forward selection
step_forward <- step(train_model, scope = list(lower = train_model, upper = train_model), direction =
"forward")
summary(step_forward)
```

```
# Backward elimination
step_backward <- step(train_model, direction = "backward")
summary(step_backward)

# 逐步迴歸
stepwise_model <- stepAIC(train_model, direction = "both", trace = FALSE)
summary(stepwise_model)



# Fit the linear model
train.fit <-    glm(y ~x2+x3+x6, data = train1, family = binomial)
summary(train.fit)

# 繪製 ROC 曲線，計算 AUC 值
prop <- sum(test1$y) / 60
fit <- predict(train.fit, newdata = test1, type = "response")
predicted <- as.numeric(fit > prop) # predict y=1 when est. > prop
xtabs(~ test1$y + predicted)
rocplot <- roc(test1$y, fit)
plot.roc(rocplot, legacy.axes = TRUE, legacy.col = "black", print.auc = TRUE, auc.polygon = TRUE)

# 添加標題
title("ROC curve", line = 2.5)
```