統計諮詢 - 作業 7

國立成功大學統計學系暨數據科學研究所

廖傑恩（RE6094028）

2021-06-04

1 Exercise 15.4

1.1 問題敘述

研究者想了解若母親為有色人種，其新生兒體重與死亡率是否有所關聯。

1.2 資料介紹

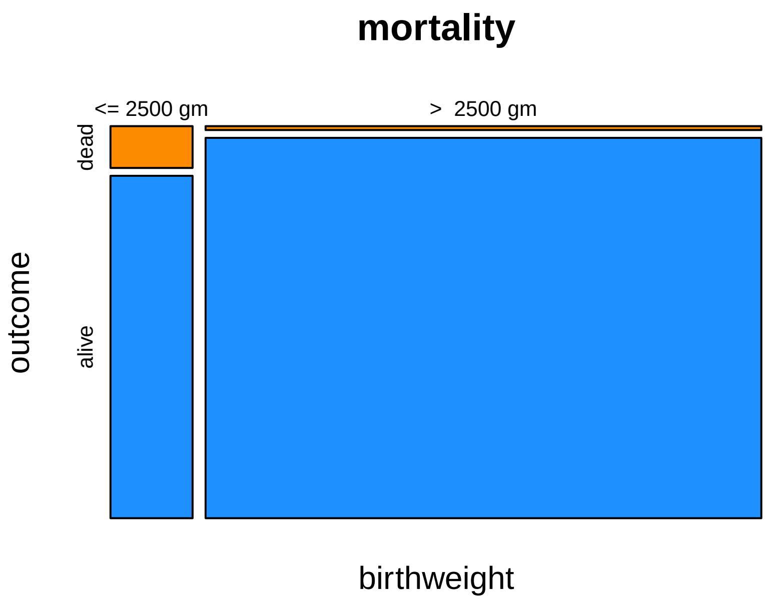
Fleiss（1981）收集了資料集 `mortality`，裡頭有 1974 年紐約市 37,840 名非白人母親分娩之新生兒存活概況，並以出生後一年以及出生重量 2500 公克作為分類基準。資料列聯表如下表所示，由此列聯表可以得到在存活超過一年的嬰兒中，大於 2500 克的嬰兒佔了 88.26% 而存活不到一年的嬰兒中，小於 2500 克的嬰兒佔了 38.59%。

出生體重	死亡數	存活數	總和
小於等於 2500 克	$a = 530$	$b = 4340$	$a + b = 4870$
大於 2500 克	$c = 333$	$d = 32637$	$c + d = 32970$
總和	$a + c = 863$	$b + d = 36977$	$a + b + c + d = 37840$

- a: 體重小於 (含)2500 公克且於出生一年內死亡的新生兒個數
- b: 體重小於 (含)2500 公克且於出生一年後存活的新生兒個數
- c: 體重大於 2500 公克且於出生一年內死亡的新生兒個數
- d: 體重大於 2500 公克且於出生一年後存活的新生兒個數

1.3 資料探索

下圖是將列聯表視覺化後的鑲嵌圖（mosaic plot），不論是由列聯表還是此圖都可以明顯看出趨勢：出生體重小於等於 2500 克的嬰兒中，死亡比率較出生體重大於 2500 克的嬰兒高；死亡的嬰兒中，出生體重小於等於 2500 克的嬰兒所佔比率比出生體重大於 2500 克的嬰兒高。



1.4 資料分析

1.4.1 卡方檢定

我們先以獨立性卡方檢定檢驗嬰兒出生體重高低與死亡之間的關聯，令顯著水準為 0.05，檢定假設如下：

- H_0: 嬰兒出生體重高低與其一年內死亡獨立
- H_1: 嬰兒出生體重高低與其一年內死亡有關

檢定統計量公式如下，因為在此資料中，列聯表為 2×2，因此需要進行 Yates 連續校正。此檢定統計量服從自由度（degree of freedom, df）為 1 的卡方分配。

$$\mathcal{X}^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}} \sim \mathcal{X}^2_{df=(2-1)(2-1)=1}$$

其中 O_{ij} 為列聯表中第 i 列第 j 行的樣本觀察值個數（observations）；E_{ij} 為若 H_0 成立，列聯表中第 i 列第 j 行的期望個數（expectations）。

檢定結果：檢定統計量 $\mathcal{X}^2 = 1851.45$，其 p 值趨近於 0，遠小於顯著水準，因此我們拒絕 H_0，顯示嬰兒出生體重高低與死亡有顯著關聯。其關聯的方向以及程度需要進行勝算比分析來探討。

1.4.2 勝算比分析

除了檢驗出生體重高低與死亡之關聯之外，本研究也將建構樣本勝算比（odds ratio）及解釋其代表意義，並建構母體勝算比之 95% 信賴區間。首先定義名詞如下：

- 勝算（odds）：發生某事件比率與未發生該事件比率之比值。以此資料集為例，小於等於 2500 克的嬰兒發生死亡之勝算（odds）為 0.1221。

- 勝算比（odds ratio）：兩個勝算之比值即為勝算比。以此資料集為例，小於等於 2500 克的嬰兒死亡之勝算，對上大於 2500 克的嬰兒死亡勝算之比值即為勝算比，若此比值顯著大於 1，則顯示小於等於 2500 克的嬰兒死亡勝算明顯高於大於 2500 克的嬰兒。樣本勝算比公式如下：

$$\hat{\Psi} = \frac{odds_1}{odds_2}$$

其中 Ψ 為母體勝算比，$odds_1$ 與 $odds_1$ 分別為兩組樣本勝算。
在此資料中，勝算與勝算比計算結果：

- 出生體重小於等於 2500 克的嬰兒發生死亡之勝算：$odds_{\leq 2500} = a/b = 530/4340 \approx 0.1221$

- 出生體重大於 2500 克的嬰兒發生死亡之勝算：$odds_{>2500} = c/d = 333/32637 \approx 0.0102$

- 樣本勝算比：$\hat{\Psi} = \frac{odds_{\leq 2500}}{odds_{>2500}} = \frac{a/b}{c/d} = \frac{ad}{bc} = \frac{0.1221}{0.0102} \approx 11.97$

1.4.2.1 勝算比雙尾檢定與信賴區間之計算

接著我們對勝算比（OR）建立信賴區間並進行假說檢定。我們先進行雙尾檢定來看兩組勝算是否有顯著差異。檢定假說為：$H_0 : \Psi = 1 \ v.s. \ H_1 : \Psi \neq 1$

樣本勝算比會服從常態分配：

$$\ln(\hat{\Psi}) \sim N(ln(\Psi), \sigma^2_{\ln(\hat{\Psi})})$$

其中 $\sigma_{\ln(\hat{\Psi})} = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$
而母體勝算比之 95% 信賴區間計算過程如下：

1. 計算取自然對數的樣本勝算比，也就是檢定統計量：

$$\ln(\hat{\Psi}) = \ln(OR) = \ln(\frac{odds_{\leq 2500}}{odds_{>2500}}) = \ln(\frac{0.1221}{0.0102}) \approx 2.48$$

2. 計算取自然對數的母體勝算比之 95% 信賴區間：

$$CI[ln(\Psi)]_{95\%,\ two.sided} = \ln(\hat{\Psi}) \pm Z_{\frac{1-.95}{2}} \sigma_{\ln(\hat{\Psi})} \approx [2.34, 2.62]$$

3. 透過 exp 轉換成母體勝算比之 95% 信賴區間：

$$CI[\Psi]_{95\%,\ two.sided}[e^{2.34}, e^{2.62}] \approx [10.40, 13.78]$$

由於 $0 \notin CI[ln(\Psi)]_{95\%,\ two.sided} \approx [2.34, 2.62]$（也就是 $1 \notin CI[\Psi]_{95\%,\ two.sided} \approx [10.40, 13.78]$），檢定統計量 $\ln(\hat{\Psi})$ 之 p 值小於顯著水準 $1 - 95\% = 0.05$，我們拒絕 H_0，表示母體勝算比顯著不為 1，也就是說，出生體重小於等於 2500 克以及大於 2500 克這兩組嬰兒的死亡勝算有顯著差異。

1.4.2.2 勝算比單尾檢定與信賴區間之計算

我們接著進行單尾檢定來確認方向性，檢定假設為：$H_0 : \Psi \leq 1 \ v.s. \ H_1 : \Psi > 1$。

1. 計算取自然對數的樣本勝算比，也就是檢定統計量：

$$\ln(\hat{\Psi}) = \ln(OR) = \ln(\frac{odds_{\leq 2500}}{odds_{>2500}}) = \ln(\frac{0.1221}{0.0102}) \approx 2.48$$

2. 計算取自然對數的母體勝算比之 95% 信賴區間：

$$CI[ln(\Psi)]_{95\%,\ one.sided} = [\ln(\hat{\Psi}) - Z_{1-.95}\sigma^2_{\ln(\hat{\Psi})}, \infty) \approx [2.36, \infty)$$

3. 透過 exp 轉換成母體勝算比之 95% 信賴區間：

$$CI[\Psi]_{95\%,\ one.sided}[e^{2.36}, e^{\infty}) \approx [10.59, \infty)$$

由於 $0 \notin CI[ln(\Psi)]_{95\%,\ one.sided} \approx [2.36, \infty)$ （也就是 $1 \notin CI[\Psi]_{95\%,\ one.sided} \approx [10.59, e^{\infty})$），檢定統計量 $\ln(\hat{\Psi})$ 之 p 值小於顯著水準 $1 - 95\% = 0.05$，我們拒絕 H_0，表示母體勝算比顯著大於 1，也就是說，出生體重小於等於 2500 克的嬰兒死亡勝算顯著大於 2500 克的嬰兒。

1.5 結論

根據獨立性卡方檢定與勝算比分析，我們得知在紐約市非白人母親分娩之新生兒中，嬰兒出生體重高低與其一年內死亡與否有顯著關聯：出生體重不超過 2500 公克的嬰兒之死亡勝算約為出生體重超過 2500 公克嬰兒的 11.97 倍（$\hat{\Psi} = \frac{odds_{\leq 2500}}{odds_{>2500}} = \frac{0.1221}{0.0102} \approx 11.97$ ）。由此研究可提出相關政策建議：呼籲孕婦於產期注意胎兒發展，且若新生兒出生時體重過輕，可能需要額外營養補充。

1.6 建議

資料僅提供各嬰兒出生時體重是否大於 2500 公克，若能取得嬰兒出生時實際體重，則出生體重為連續變項，會帶有更細緻的資訊，可以進行更多分析，例如建立邏輯式迴歸模型，以嬰兒出生體重（連續變項）來預測其是否於一年內死亡（二分變項），並可藉由迴歸係數得知當嬰兒出生體重每下降（或上升）1 公克，其一年內死亡機率上升（或下降）多少，這樣的發現會更細緻。此外，嬰兒出生一年內死亡可能與出生體重之外的因子有關，例如母親社經地位、社會支持程度，如果也收集這些資料並納入分析，可以得到更多發現，也能更加確認出生體重與一年內死亡的關聯。

2 Exercise 17.4

2.1 問題敘述

研究者發現有些年輕男性病患於加護病房（Intensive Care Unit, ICU）死亡，然而卻沒有女性病患死亡案例，因此想探討性別與年齡是否為病患於 ICU 死亡的重要因素。

2.2 資料介紹

Hosmer 與 Lemeshow (2000) 收集了 2000 年的 ICU 200 筆病患資料，除了以下列出的變項之外，資料集也包含了病患在 ICU 接受的處置與生理狀況的變項，例如於是否有感染可能、血壓與心率等等，不過在此分析中，我們僅考慮諮詢者關注的年齡與性別這兩個因子。

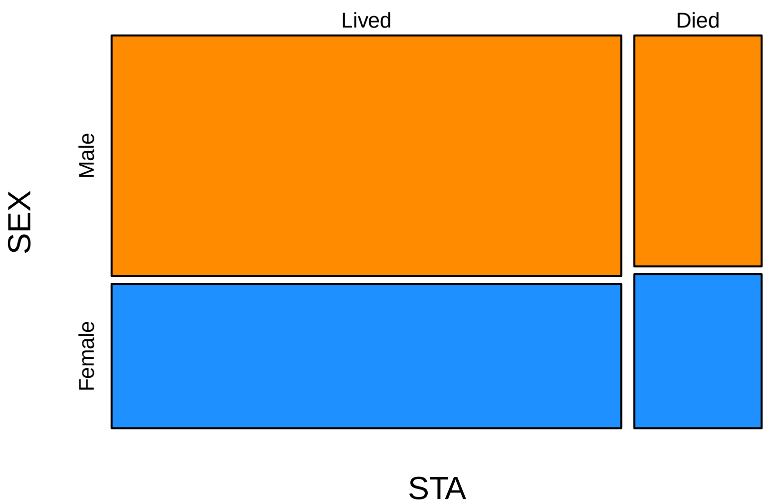
- STA：生存狀態，0= 生存；1= 死亡，依變項
- SEX：性別，0= 男性；1= 女性
- AGE：年齡

2.3 資料探索

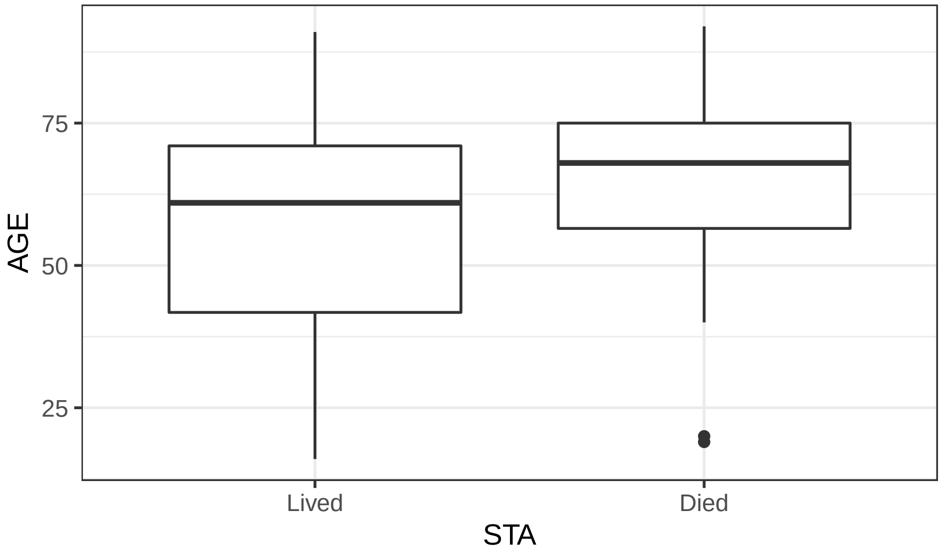
不同性別存活與死亡個數如下列聯表：

性別＼生存狀態	存活數	死亡數	總和
男性	100	24	124
女性	60	16	76
總和	160	40	200

下圖是將列聯表視覺化後的鑲嵌圖（mosaic plot），由圖與表皆可觀察到：不同性別死亡比率接近。



由下面盒鬚圖（box plot）可以發現，與存活病患相比，死亡病患的年齡中位數較大，變異則較小。



由下面以性別分層的盒鬚圖可以發現，不論性別為何，與存活病患相比，死亡病患的年齡中位數都較大，變異則較小。

2.4 資料分析

2.4.1 模型

我們將建立邏輯式迴歸模型，以 ICU 病患性別及年齡預測其生存狀態。由於在資料探索的圖片中沒有觀察到交互作用存在的可能，因此模型不考慮交互作用項。模型定義如下：

$$\log[\frac{p_i}{1-p_i}] = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2},\ \forall i = 1, 2, ..., 200$$

其中 p_i 為病患 i 死亡的機率；X_{i1} 為病患 i 性別，男性記為 0，女性記為 1；X_{i2} 為病患 i 年齡；β_0、β_1 與 β_2 為迴歸係數。根據模型，病患 i 死亡的機率可以表示成：

$$p_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2})}$$

2.4.2 資料切割

為了驗證模型的預測能力，避免過度配適（overfitting），我們依據生存狀態的比率對病患進行分層隨機抽樣，將原始資料的 20 作為測試資料集（testing data），剩下的 80 為訓練資料集（training data），用來配適模型。

2.4.3 模型配適與係數檢定

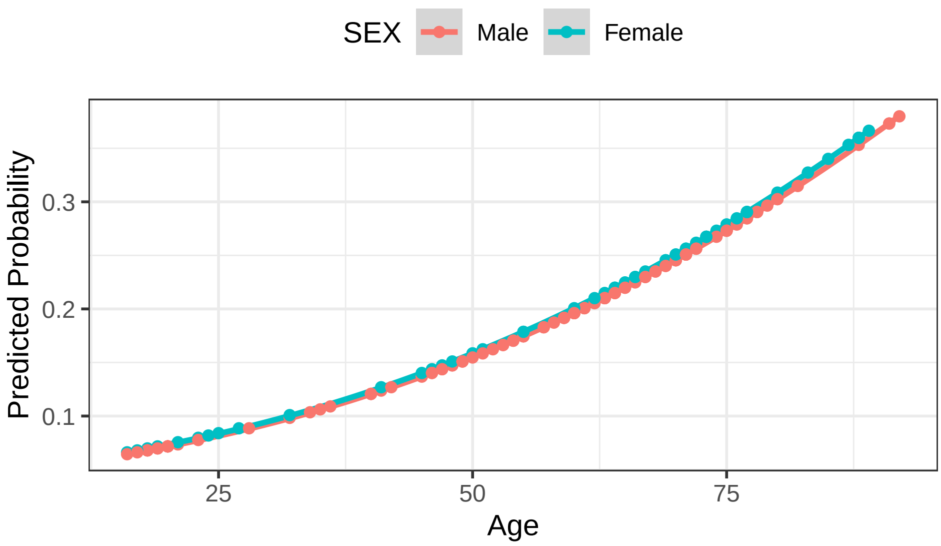
我們對截距以外之迴歸係數進行 Z 檢定，檢定假說為 $H_0 : \beta_j = 0 \ v.s. \ H_1 : \beta_j \neq 0,\ j = 1, 2$。顯著水準設定為 0.05。

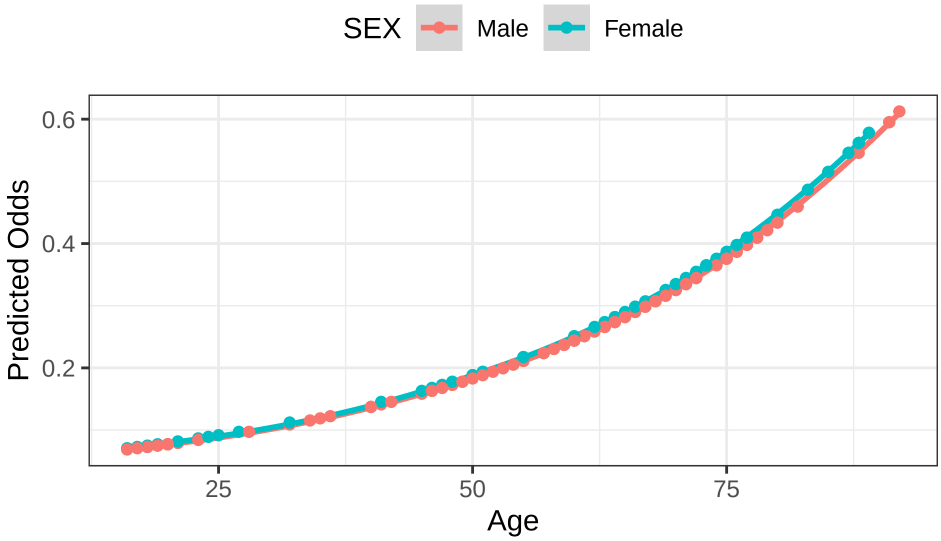
	係數估計值	標準誤	Z 檢定統計量	p 值
beta_0	-3.1365	0.7668	-4.0902	0.0000
beta_1	0.0287	0.4241	0.0678	0.9460
beta_2	0.0288	0.0117	2.4534	0.0142

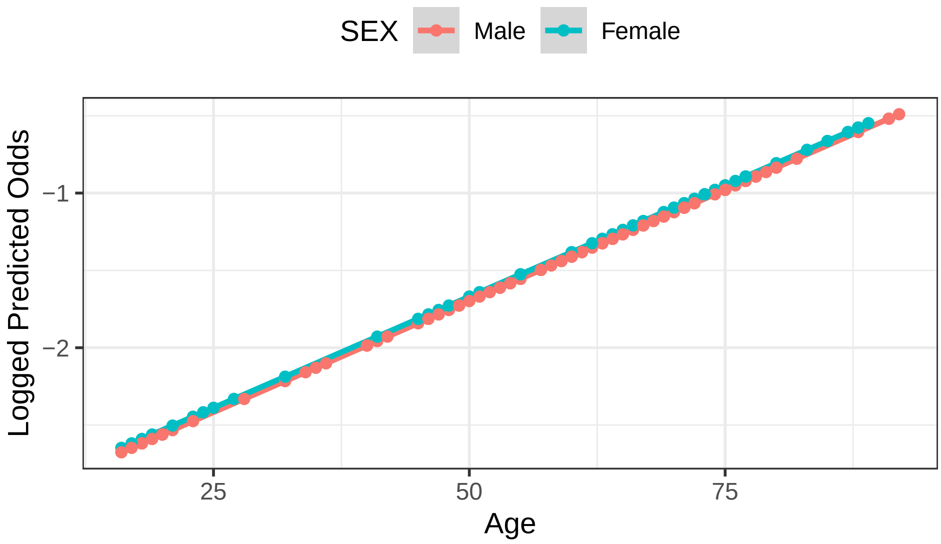
上表為邏輯式迴歸模型配適結果。針對 β_1 的檢定的 p 值小於顯著水準，因此我們拒絕 H_0。而針對 β_2 的檢定的 p 值大於顯著水準，因此我們不拒絕 H_0。結果顯示 ICU 病患年齡對於其死亡機率預測力顯著，而性別則無。

2.4.4 模型視覺化

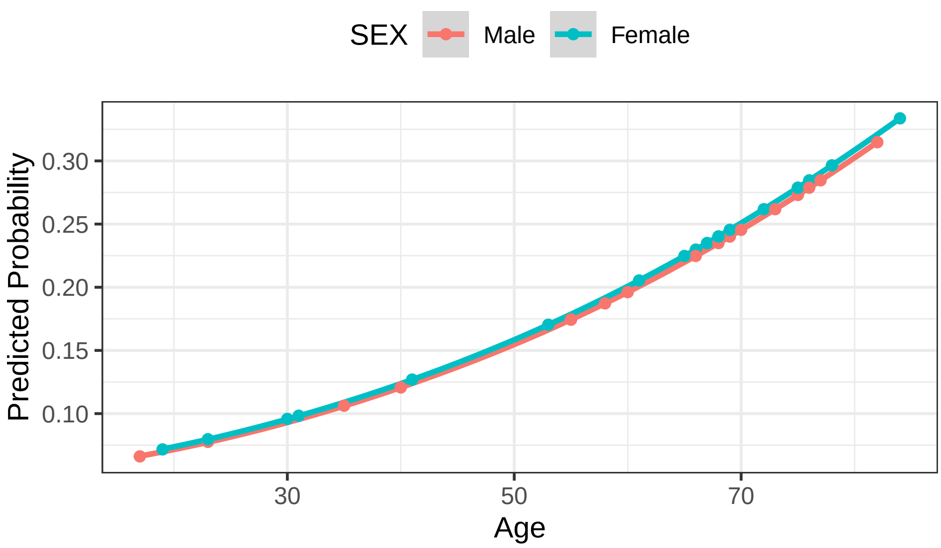
我們將模型在訓練資料集上的表現視覺化成下三圖：最上方的圖縱軸為預測死亡機率 $\hat{p}_i$，呈現其與年齡間的關係，由圖可見年齡越大預測死亡機率也越大，且會隨著年齡增加上升速度增加；中間的圖縱軸為 $\frac{\hat{p}_i}{1-\hat{p}_i}$，顯示隨著年齡的增加，死亡勝算增加，且上升幅度越來越大，也就是年齡越大死亡機率越高；最下方的圖縱軸為 $\log[\frac{\hat{p}_i}{1-\hat{p}_i}]$，即為邏輯式迴歸模型最原始的預測目標。三張圖都以性別分層，不同性別用不同顏色標注。

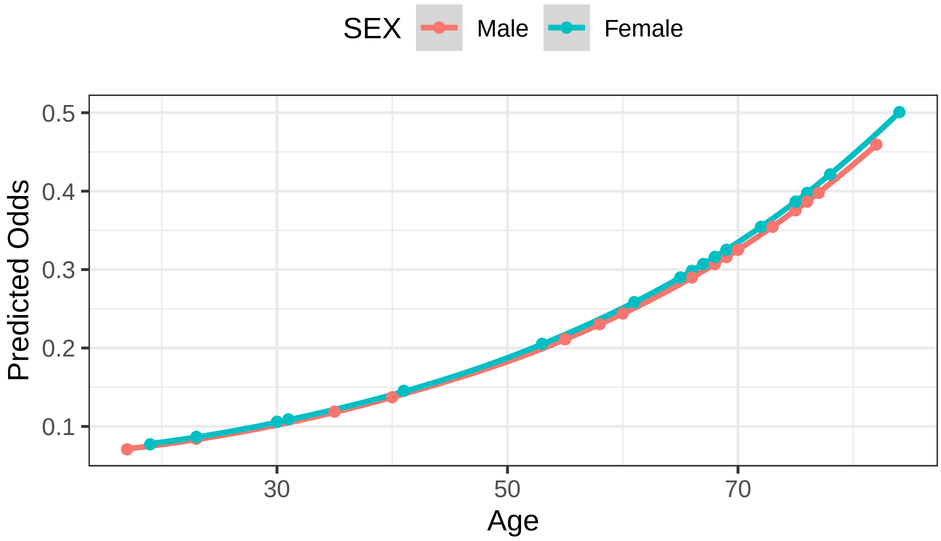


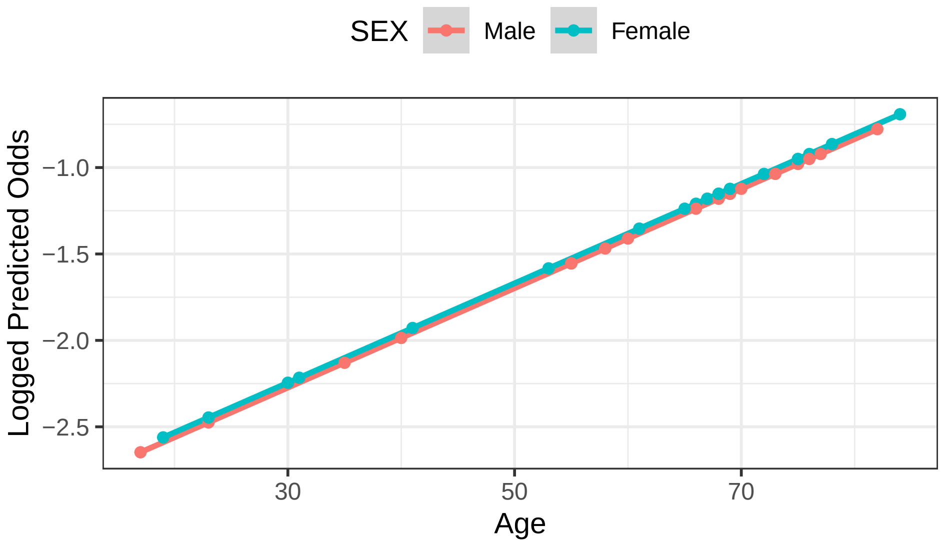




以下三圖則是模型在測試資料集的結果，與其在訓練資料集上的表現（上三圖）一致。







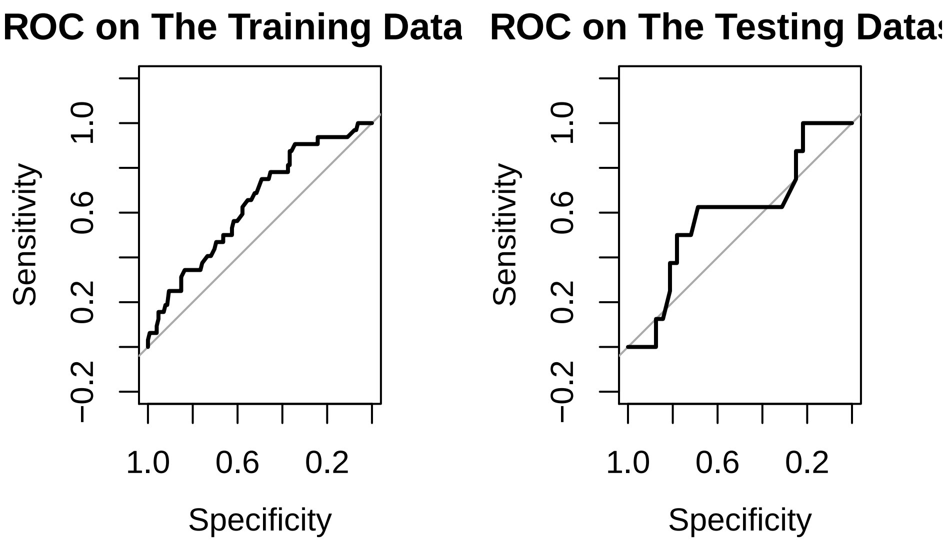
這 6 張圖中，代表不同性別的兩條線非常接近、幾乎重疊，表示其實不同性別之死亡率差異不大，不過因為諮詢者希望模型納入性別，因此我們保留這個變數。

2.4.5 模型預測表現評估

若將機率切點定為 0.5，模型在訓練與測試資料集的預測正確率分別為 0.8 與 0.8，看似很高，但由下面此模型在訓練與測試資料集的預測結果混淆矩陣（confusion matrix）可知，其實模型都只是隨意將所有病患都預測為存活，高正確率是由存活病患的高比率造成的，而非模型的預測能力。

模型預測＼實際情形	病患實際存活	病患實際死亡
訓練集中，模型預測病患為存活	128	32
訓練集中，模型預測病患為死亡	0	0
測試集中，模型預測病患為存活	32	8
測試集中，模型預測病患為死亡	0	0

我們接著繪製模型在訓練與測試資料集上的接收者操作特徵曲線（receiver operating characteristic curve，ROC 曲線），其橫軸為特異度（i.e., 1 - 偽陽性率），縱軸為敏感度（i.e., 真陽性率）。我們也計算相對應的 ROC 曲線下面積（area under ROC curve, AUROC）。



不論是在訓練還是測試資料集，ROC 曲線都十分貼近 45 度灰色斜直線，且 AUROC 在兩資料集都不高，分別為 0.6432 與 0.5938，顯示此模型預測能力不佳。

2.5 結論與建議

我們建立一個邏輯式迴歸模型，以 ICU 病患年齡與性別預測其生存狀態，模型如下：

$$\log[\frac{\hat{p}}{1-\hat{p}}] = -3.1365 + 0.0287x_1 + 0.0288x_2$$

也就是說，預測死亡機率為

$$\hat{p} = \frac{\exp(-3.1365 + 0.0287x_1 + 0.0288x_2)}{1 + \exp(-3.1365 + 0.0287x_1 + 0.0288x_2)}$$

從係數估計值可知，當性別固定時，女性病患死亡的勝算會是男性病患的 $e^{0.0287} = 1.0291$ 倍（此效果未達統計顯著）；而病患年齡每增加 1 歲，其死亡勝算會是原本的 $e^{0.0288} = 1.0291$ 倍（此效果達統計顯著）。

根據此模型，可以給出相關政策參考：ICU 患者死亡機率隨著其年齡上升而提高，醫護人員應加強照護高齡 ICU 患者。

關於分析的建議：由混淆矩陣、ROC 曲線與 AUROC 可知此模型預測能力不佳，建議將資料集中的其他變項納入模型，並進行變數挑選與模型比較。