類別資料分析

期末報告

組員:

楊善評　107354002

林志軒　107354025

李世摞　107354031

一、 研究動機與目的

探討鐵達尼號資料中，存活與否和客艙等級、性別、年齡、兄弟姐妹數/配偶數、父母數/子女數、船票價格、登船港口之間關連性，並評估在怎樣的情況下生存率會是最高的，故此為探索性研究，目的在於利用變數間關係配適出最佳的模型。

二、 資料來源及介紹

(1) 資料來源:kaggle(網址: https://www.kaggle.com/c/titanic/data)

(2) 資料介紹:此資料共 12 個變數，891 個樣本，以下列出前 10 筆:

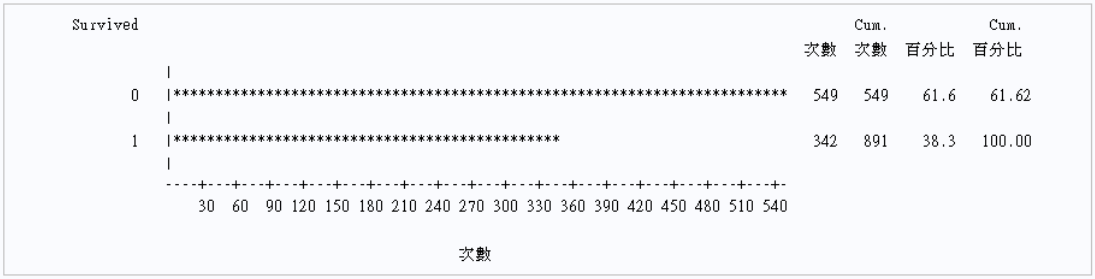
Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings,	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen	female	26	0	0	STON/O2	7.925		S
4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
6	0	3	Moran, M	male		0	0	330877	8.4583		Q
7	0	1	McCarthy	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, M	male	2	3	1	349909	21.075		S
9	1	3	Johnson, M	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C

(3) 各變數介紹:

變數	意義
PassengerId	每個乘客的編號
Survived	乘客是否倖存——倖存(1)、死亡(0)。我們將預測這一列。
Pclass	乘客所屬船艙等級——第一級(1),第二級(2),第三級(3)。
Name	乘客姓名。
Sex	乘客性別——男 male、女 female
Age	乘客年齡。
SibSp	船上兄弟姐妹和配偶的數量。
Parch	船上父母家長和孩子的數量。
Ticket	船票號碼。
Fare	票價。
Cabin	乘客所在船艙編號。
Embarked	乘客從哪個港口登船。C = Cherbourg 瑟堡-位於法國, Q = Queenstown 皇后鎮-位於紐西蘭, S = Southampton 南安普敦-位於英格蘭

取是否倖存當應變數，客艙等級、性別、年齡、船上兄弟姐妹數和配偶數、船上父母數和子女數、船票價格、登船港口當自變數，而剩下的乘客編號、姓名、船票號碼、船艙編號較沒討論的意義，故將其省略，以下對這一個應變數及七個自變數進行一些基本的介紹

1.Survived(是否倖存):



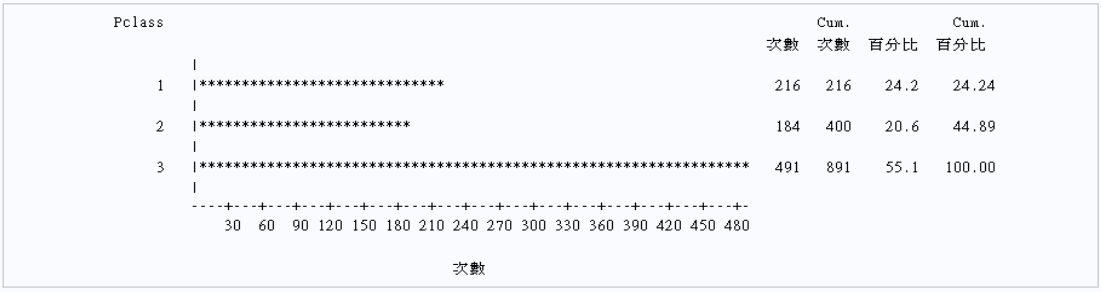
Survived	次數	Cum. 次數	百分比	Cum. 百分比
0	549	549	61.6	61.62
1	342	891	38.3	100.00

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540

次數

樣本的存活比率為 38.3%

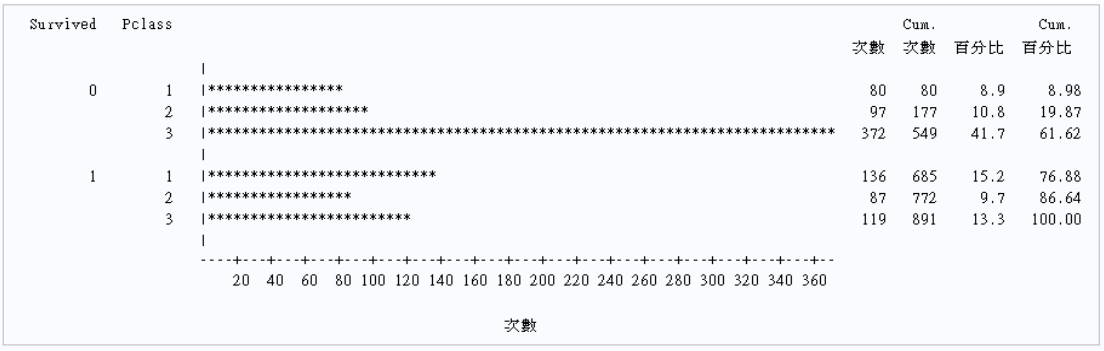
2. Pclass(客艙等級):



Pclass	次數	Cum. 次數	百分比	Cum. 百分比
1	216	216	24.2	24.24
2	184	400	20.6	44.89
3	491	891	55.1	100.00

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480

次數



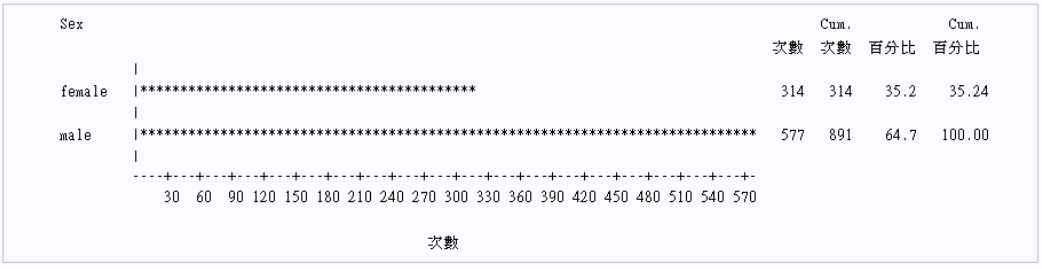
Survived	Pclass	次數	Cum. 次數	百分比	Cum. 百分比
0	1	80	80	8.9	8.98
	2	97	177	10.8	19.87
	3	372	549	41.7	61.62
1	1	136	685	15.2	76.88
	2	87	772	9.7	86.64
	3	119	891	13.3	100.00

20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360

次數

此處第一級表示最高等級，第三級表示最低等級。第一級與第二級人數差不多，而第三級人數為第一、二級的兩倍以上，若加入 Survived 來分類，則可明顯看出，第三級也就是最低等級，死亡比率來的比其他兩級高很多

3.Sex(性別):



Sex	次數	Cum. 次數	百分比	Cum. 百分比
female	314	314	35.2	35.24
male	577	891	64.7	100.00

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570

次數

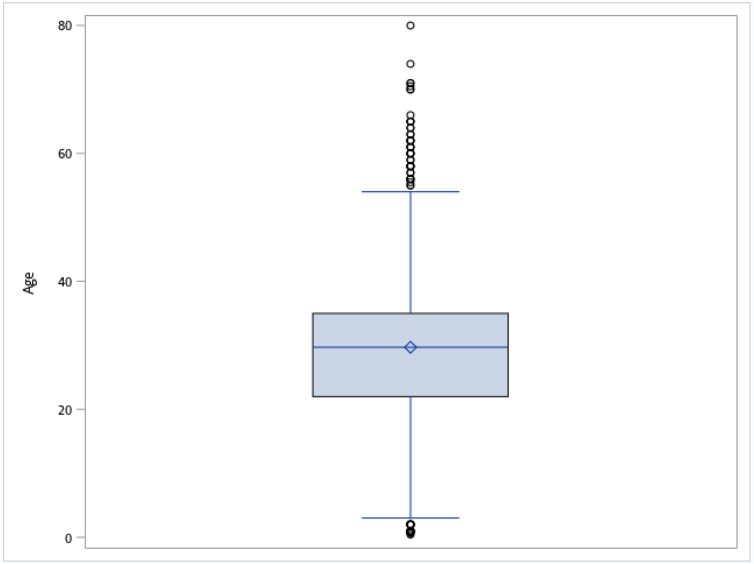
Survived	Sex		次數	Cum. 次數	百分比	Cum. 百分比
0	female	**********	81	81	9.0	9.09
	male	**********	468	549	52.5	61.62
1	female	**********	233	782	26.1	87.77
	male	**********	109	891	12.2	100.00

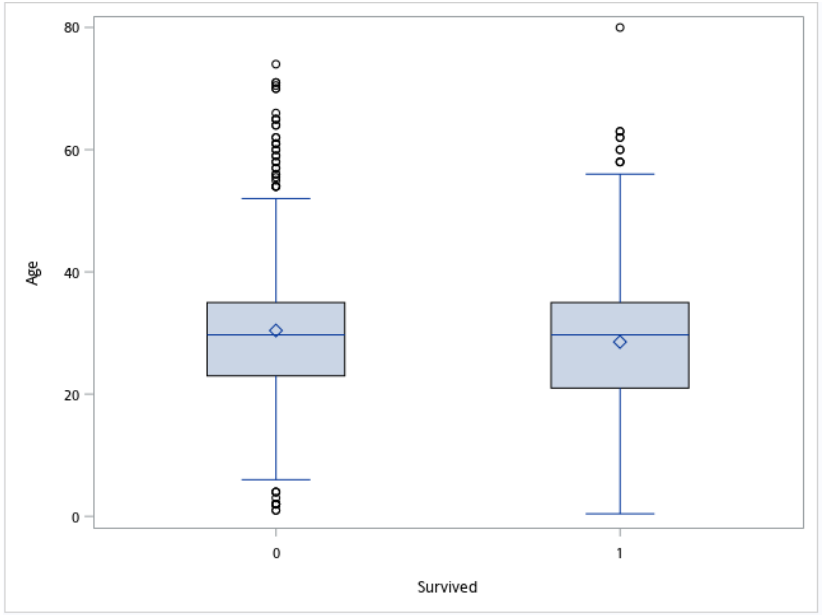
30 60 90 120 150 180 210 240 270 300 330 360 390 420 450

次數

男生人數大約為女生的兩倍，若加入 Survived 來分類，則可明顯看出，男生死亡比率比女生高了很多，且男生存活率也比女生低許多

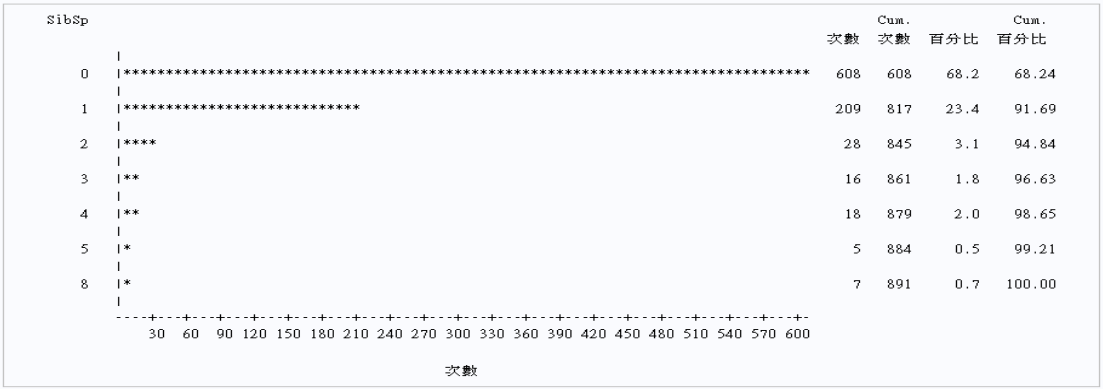
4.Age(年齡):

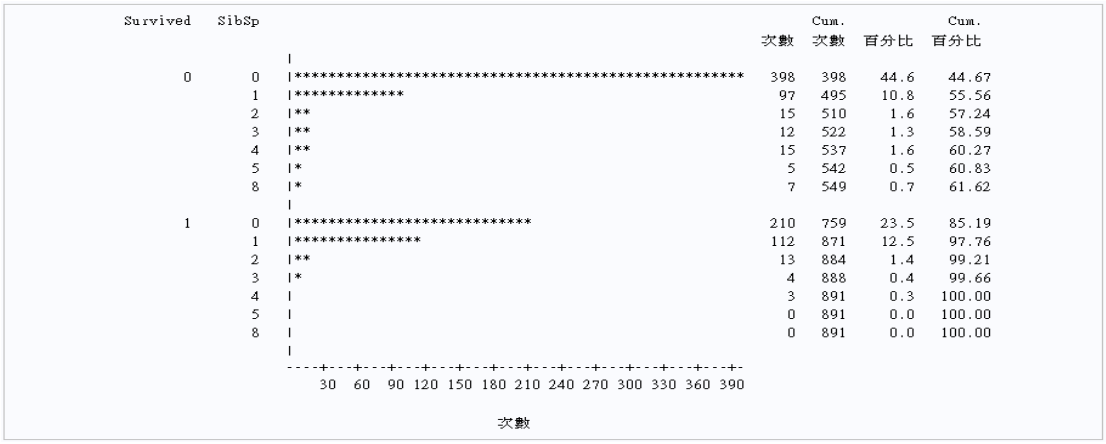




由於在原始資料中存在遺失值，故採用平均數做差補。年齡可能受平均數差補的影響，分配不管有沒有用 Survived 分類都大致呈現對稱，但仍有幾點較偏離大部份的資料

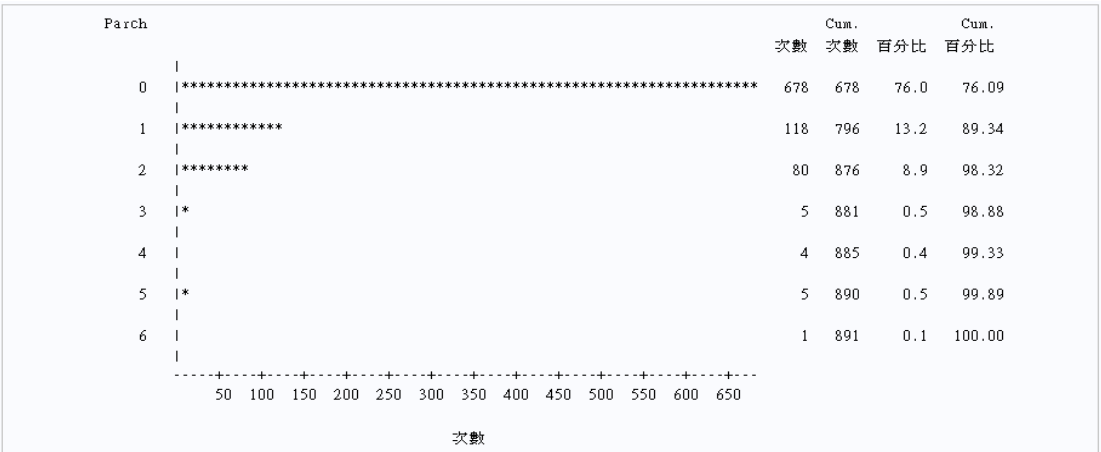
5.SibSp(船上兄弟姐妹數和配偶數):





船上兄弟姐妹數和配偶數越多的人數越少，若加入 Survived 來分類，船上兄弟姐妹數和配偶數越多，則死亡比率及生存比率都是下降，會造成這樣的狀況，也因為各類人數差異大

6.Parch(船上父母數和子女數):



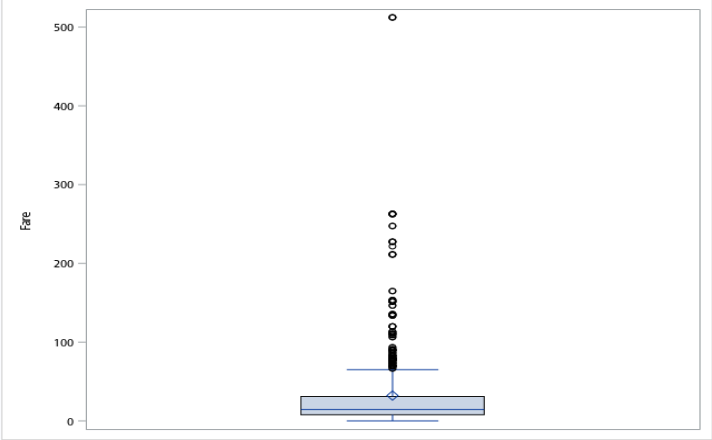
Survived	Parch		次數	Cum. 次數	百分比	Cum. 百分比
0	0	**********	445	445	49.9	49.94
	1	*******	53	498	5.9	55.89
	2	*****	40	538	4.4	60.38
	3		2	540	0.2	60.61
	4	*	4	544	0.4	61.05
	5	*	4	548	0.4	61.50
	6		1	549	0.1	61.62
1	0	********	233	782	26.1	87.77
	1	*********	65	847	7.3	95.06
	2	*****	40	887	4.4	99.55
	3		3	890	0.3	99.89
	4		0	890	0.0	99.89
	5		1	891	0.1	100.00
	6		0	891	0.0	100.00

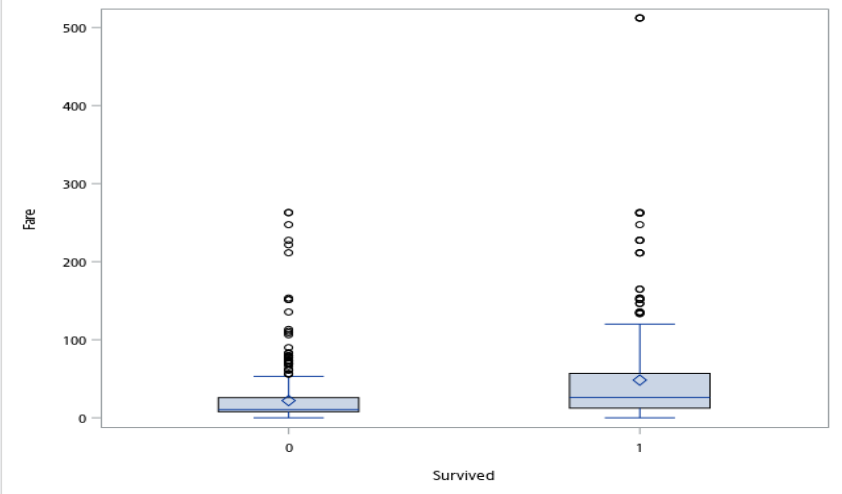
30 60 90 120 150 180 210 240 270 300 330 360 390 420

次數

船上父母數和子女數越多的人數越少，若加入 Survived 來分類，船上父母數和子女數越多，則死亡比率及生存比率都是下降，會造成這樣的狀況，也因為各類人數差異大

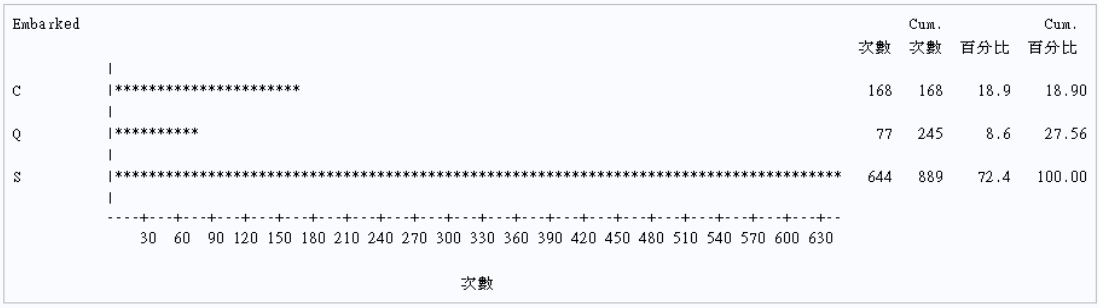
7.Fare(船票價格):





分配不管有沒有用 Survived 分類都大致呈現右偏，但仍有幾點較偏離大部份的資料

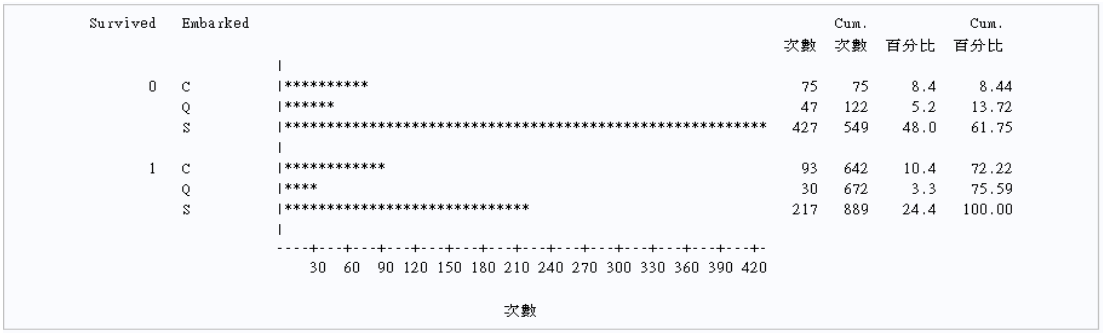
8.Embarked(登船港口):



Embarked		次數	Cum. 次數	百分比	Cum. 百分比
C	**********	168	168	18.9	18.90
Q	**********	77	245	8.6	27.56
S	**********	644	889	72.4	100.00

30 60 90 120 150 180 210 240 270 300 330 360 390 420 450 480 510 540 570 600 630

次數



Survived	Embarked	次數	Cum. 次數	百分比	Cum. 百分比
0	C	75	75	8.4	8.44
	Q	47	122	5.2	13.72
	S	427	549	48.0	61.75
1	C	93	642	10.4	72.22
	Q	30	672	3.3	75.59
	S	217	889	24.4	100.00

30 60 90 120 150 180 210 240 270 300 330 360 390 420

次數

從 S 港口登船人數比其他兩個港口登船人數多很多，也因此加入 Survived 分類後，從 S 港口登船的死亡比率及存活比率都比另外兩類高

以下為連續型變數的敘述統計:

Variable	Label	N	Mean	Median	Std Dev	Minimum	Maximum	Sum
Survived	Survived	891	0.3838384	0	0.4865925	0	1.0000000	342.0000000
Pclass	Pclass	891	2.3086420	3.0000000	0.8360712	1.0000000	3.0000000	2057.00
Age	Age	891	29.6992929	29.7000000	13.0020152	0.4200000	80.0000000	26462.07
SibSp	SibSp	891	0.5230079	0	1.1027434	0	8.0000000	466.0000000
Parch	Parch	891	0.3815937	0	0.8060572	0	6.0000000	340.0000000
Fare	Fare	891	32.2042080	14.4542000	49.6934286	0	512.3292000	28693.95

三、 原始模式

由於應變數為二元變數，所以建立羅吉斯迴歸模型，首先考慮七個自變數且有兩兩交互作用的模型

Warning: The maximum likelihood estimate may not exist.

Warning: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.

配適結果會有 mle 不存在的問題，且考慮到模型過於複雜，因此接下來都利用沒有交互作用的模型進行討論

再來建立有七個自變數的羅吉斯迴歸模型，其中 Pclass、Sex、Embarked 等三個變數為類別變數，所以分別設 2、1、2 個虛擬變數，但由於 Pclass 為次序尺度的類別變數，因此欲先檢定 Pclass 是否能直接當成連續型變數來放入模型中

模型配適統計值		
準則	僅截距	截距和共變量
AIC	1184.818	803.735
SC	1189.608	851.636
-2 對數 L	1182.818	783.735

模型配適統計值		
準則	僅截距	截距和共變量
AIC	1184.818	802.188
SC	1189.608	845.299
-2 對數 L	1182.818	784.188

H_0:logit(π) $= \alpha +$

$\beta_1$pclass+ $\beta_2$sex(male)+ $\beta_3$age+ $\beta_4$sibsp+ $\beta_5$parch+ $\beta_6$fare+ $\beta_7$embarked(C)+ β_8 embarked(Q)

vs

H_1:logit(π) $= \alpha +$

$\beta_1$pclass(1)+ $\beta_2$pclass(2)+ $\beta_3$sex(male)+ $\beta_4$age+ $\beta_5$sibsp+ $\beta_6$parch+ $\beta_7$fare+ β_8 embarked(C)+ β_9 embarked(Q)

TS=784.188-783.735=0.453，df=1

0.453<$X^2_{0.05}(1)$=3.841

所以不拒絕 H_0，在 α =0.05 下，我們沒有證據說 pclass 是類別變數的模型是對的

因此最後決定配適 Pclass 為連續型變數的羅吉斯迴歸模型，模型如下

最大概度估計值分析						
參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	4.8664	0.5378	81.8902	<.0001
Pclass		1	-1.1001	0.1435	58.7418	<.0001
Sex	male	1	-2.7187	0.2008	183.3429	<.0001
Age		1	-0.0399	0.00785	25.8063	<.0001
SibSp		1	-0.3258	0.1094	8.8699	0.0029
Parch		1	-0.0926	0.1187	0.6086	0.4353
Fare		1	0.00192	0.00238	0.6519	0.4194
Embarked	C	1	0.4188	0.2368	3.1283	0.0769
Embarked	Q	1	0.3848	0.3293	1.3655	0.2426

$\text{logit}(\hat{\pi}) =$ 4.8664-1.1001pclass-2.7187sex(male)-0.0399age-0.3258SibSp-0.0926Parch+0.00192Fare+0.4188Embarked(C)+0.3848Embarked(Q)

四、 最佳模型的選擇

(1) 利用 Wald 及 LR 檢定 full model 下個別參數是否顯著:

首先檢定整體模型參數是否顯著

檢定全域虛無假設: BETA=0			
檢定	卡方	DF	Pr > ChiSq
概度比	398.6294	8	<.0001
評分	352.3106	8	<.0001
Wald	240.3117	8	<.0001

H_0:$\beta_j = 0$，j=1,..,9

H_1:至少一$\beta_j \neq 0$

α=0.05

因為 p-value<0.0001<α，所以拒絕 H_0，我們有充分證據說至少一$\beta_j \neq 0$

接著檢定個別參數是否顯著

Wald Test:

效果的第三型分析			
效果	DF	Wald 卡方	Pr > ChiSq
Pclass	1	58.7418	<.0001
Sex	1	183.3429	<.0001
Age	1	25.8063	<.0001
SibSp	1	8.8699	0.0029
Parch	1	0.6086	0.4353
Fare	1	0.6519	0.4194
Embarked	2	4.0583	0.1314

$H_0:\beta_1 = 0$ $H_1:\beta_1 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Pclass 是一個顯著的解釋變數。

$H_0:\beta_2 = 0$ $H_1:\beta_2 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Sex 是一個顯著的解釋變數。

$H_0:\beta_3 = 0$ $H_1:\beta_3 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Age 是一個顯著的解釋變數。

$H_0:\beta_4 = 0$ $H_1:\beta_4 \neq 0$

p-value=0.0029<α=0.05 故拒絕 H_0 。即有充分證據顯示 SibSp 是一個顯著的解釋變數。

$H_0:\beta_5 = 0$ $H_1:\beta_5 \neq 0$

p-value=0.4353>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Parch 是一個顯著的解釋變數。

$H_0:\beta_6 = 0$ $H_1:\beta_6 \neq 0$

p-value=0.4194>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Fare 是一個顯著的解釋變數。

$H_0:\beta_7 = 0$ $H_1:\beta_7 \neq 0$

p-value=0.1314>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Embarked 是一個顯著的解釋變數。

LR Test:

類型 3 分析的 LR 統計值			
來源	DF	卡方	Pr > ChiSq
Pclass	1	59.39	<.0001
Sex	1	230.59	<.0001
Age	1	27.86	<.0001
SibSp	1	10.40	0.0013
Parch	1	0.62	0.4322
Fare	1	0.69	0.4051
Embarked	2	4.04	0.1329

$H_0:\beta_1 = 0$ $H_1:\beta_1 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Pclass 是一個顯著的解釋變數。

$H_0:\beta_2 = 0$ $H_1:\beta_2 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Sex 是一個顯著的解釋變數。

$H_0:\beta_3 = 0$ $H_1:\beta_3 \neq 0$

p-value<0.0001 故拒絕 H_0 。即有充分證據顯示 Age 是一個顯著的解釋變數。

$H_0:\beta_4 = 0$ $H_1:\beta_4 \neq 0$

p-value=0.0013<α=0.05 故拒絕 H_0 。即有充分證據顯示 SibSp 是一個顯著的解釋變數。

$H_0:\beta_5 = 0$ $H_1:\beta_5 \neq 0$

p-value=0.4322>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Parch 是一個顯著的解釋變數。

$H_0:\beta_6 = 0$ $H_1:\beta_6 \neq 0$

p-value=0.4051>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Fare 是一個顯著的解釋變數。

$H_0:\beta_7 = 0$ $H_1:\beta_7 \neq 0$

p-value=0.1329>α=0.05 故不拒絕 H_0 。即沒有充分證據顯示 Embarked 是一個顯著的解釋變數。

為確保沒有共線性導致整體及個別參數是否顯著的檢定結果不一致的問題，因此看變數間的 VIF

參數估計值							
變數	標籤	DF	參數估計值	標準誤差	t 值	Pr > |t|	變異數膨脹
Intercept	Intercept	1	1.06826	0.08089	13.21	<.0001	0
Pclass	Pclass	1	-0.20687	0.02299	-9.00	<.0001	1.65341
Age	Age	1	-0.00764	0.00126	-6.08	<.0001	1.19500
SibSp	SibSp	1	-0.04279	0.01530	-2.80	0.0053	1.27426
Parch	Parch	1	0.04499	0.02086	2.16	0.0313	1.26539
Fare	Fare	1	0.00078381	0.00037892	2.07	0.0389	1.58654

由此可見 VIF 都不大，因此確定無共線性的問題

(2) 向前選取法

前進選擇的摘要						
步階	輸入的效果	DF	數目於	計分卡方	Pr > ChiSq	變數標籤
1	Sex	1	1	260.7563	<.0001	Sex
2	Pclass	1	2	90.2592	<.0001	Pclass
3	Age	1	3	21.5589	<.0001	Age
4	SibSp	1	4	12.7128	0.0004	SibSp

效果的第三型分析			
效果	DF	Wald 卡方	Pr > ChiSq
Pclass	1	95.3656	<.0001
Sex	1	197.7849	<.0001
Age	1	26.5064	<.0001
SibSp	1	11.7842	0.0006

(3) 向後消去法

向後消去的摘要						
步階	移除的效果	DF	數目於	Wald 卡方	Pr > ChiSq	變數標籤
1	Parch	1	6	0.6086	0.4353	Parch
2	Fare	1	5	0.4469	0.5038	Fare
3	Embarked	2	4	5.0471	0.0802	Embarked

效果的第三型分析			
效果	DF	Wald 卡方	Pr > ChiSq
Pclass	1	95.3656	<.0001
Sex	1	197.7849	<.0001
Age	1	26.5064	<.0001
SibSp	1	11.7842	0.0006

(4) 逐步迴歸選取法

逐步選擇的摘要								
步階	效果		DF	數目於	計分卡方	Wald卡方	Pr > ChiSq	變數標籤
	輸入	移除						
1	Sex		1	1	260.7563		<.0001	Sex
2	Pclass		1	2	90.2592		<.0001	Pclass
3	Age		1	3	21.5589		<.0001	Age
4	SibSp		1	4	12.7128		0.0004	SibSp

效果的第三型分析			
效果	DF	Wald卡方	Pr > ChiSq
Pclass	1	95.3656	<.0001
Sex	1	197.7849	<.0001
Age	1	26.5064	<.0001
SibSp	1	11.7842	0.0006

(5) LASSO

LASSO 選擇摘要				
步階	輸入的效果	移除的效果	效果數目	SBC
0	Intercept		1	-1276.1706
1	Sex_male		2	-1467.8220
2	Pclass		3	-1643.6151
3	Age		4	-1647.9202
4	Embarked_C		5	-1650.0780
5	SibSp		6	-1674.7006*
* 準則的最佳值				

選擇已在 SBC 準則的本機最小值處停止。

停止詳細資料				
候選項目	效果	候選項目 SBC		比較 SBC
項目	Fare	-1673.5925	>	-1674.7006

(6) 各方法所選變數之比較

- WALD 及 LR TEST

  Pclass、Sex、Age、SibSp
- 向前選取法

  Pclass、Sex、Age、SibSp
- 向後消去法

  Pclass、Sex、Age、SibSp
- 逐步迴歸選取法

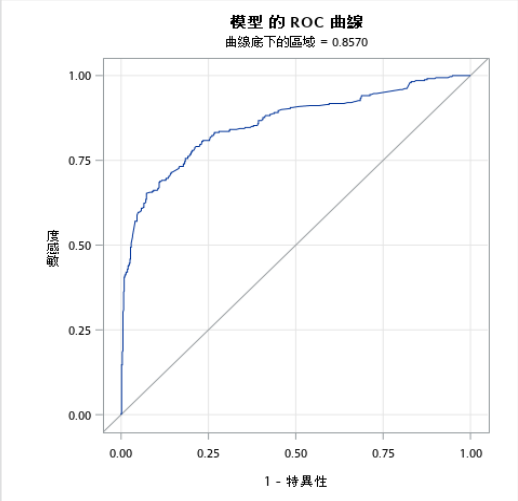
  Pclass、Sex、Age、SibSp
- Lasso

  Pclass、Sex、Age、SibSp、Embarked

(7) 利用 AIC 及 ROC CURVE 來比較(6)中兩模型

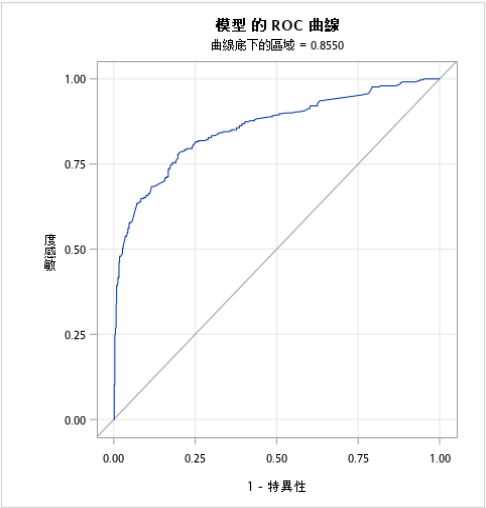
Lasso 所選的模型:AIC=799.272、AUC=0.8570

模型配適統計值		
準則	僅截距	截距和共變量
AIC	1184.818	799.272
SC	1189.608	832.803
-2 對數 L	1182.818	785.272



其餘方法選的模型: AIC=800.845、AUC=0.8550

模型配適統計值		
準則	僅截距	截距和共變量
AIC	1188.655	800.845
SC	1193.447	824.806
-2 對數 L	1186.655	790.845



由於兩者差異不大，所以最終選擇較好解釋的模型

$logit(\pi) = \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp

$logit(\hat{\pi}) = 5.1919 - 1.1724$pclass-2.7398sex(male)-0.0398age-0.3578SibSp

五、 模型診斷及校正

(1)利用 Hosmer and Lemeshow test 來檢定最終選擇的模型是否配適的好

Hosmer 和 Lemeshow 配適度檢定		
卡方	DF	Pr > ChiSq
17.0499	8	0.0296

H_0:$logit(\pi) = \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp

H_1:$logit(\pi) \neq \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp

p-value=0.0296<0.05 故拒絕 H_0 ，在$\alpha = 0.05$ 下，有充分證據顯示模型配適的不好

(2)偵測影響點，以下列出前 10 筆

迴歸診斷																
案例編號	共變量				Pearson 殘差	偏差殘差	Hat 矩陣對角	Intercept DfBeta	Pclass DfBeta	Sexmale DfBeta	Age DfBeta	SibSp DfBeta	信賴區間置換 C	信賴區間置換 CBar	Delta 偏差	Delta 卡方
	Pclass	Sex male	Age	SibSp												
1	3.0000	1.0000	22.0000	1.0000	-0.3169	-0.4375	0.00239	0.00637	-0.00863	-0.0101	-0.00008	-0.00591	0.000241	0.000241	0.1917	0.1007
2	1.0000	0	38.0000	1.0000	0.3414	0.4695	0.00440	0.0130	-0.0153	-0.0142	0.000665	0.00345	0.000517	0.000515	0.2210	0.1170
3	3.0000	0	26.0000	0	0.7261	0.9203	0.00656	0.0103	0.0168	-0.0367	-0.00441	-0.0272	0.00351	0.00348	0.8503	0.5307
4	1.0000	0	35.0000	1.0000	0.3216	0.4436	0.00402	0.0133	-0.0146	-0.0130	-0.00159	0.00257	0.000419	0.000417	0.1972	0.1038
5	3.0000	1.0000	35.0000	0	-0.2926	-0.4054	0.00229	0.00898	-0.0102	-0.00756	-0.00608	0.00122	0.000197	0.000197	0.1645	0.0858
6	3.0000	1.0000	29.7000	0	-0.3252	-0.4483	0.00233	0.00790	-0.0108	-0.00876	-0.00343	0.00241	0.000247	0.000247	0.2013	0.1060
7	1.0000	1.0000	54.0000	0	-0.6476	-0.8369	0.00878	0.00837	0.0147	-0.0138	-0.0374	-0.00140	0.00375	0.00372	0.7041	0.4231
8	3.0000	1.0000	2.0000	3.0000	-0.3299	-0.4546	0.00951	0.00167	-0.00430	-0.0143	0.0104	-0.0234	0.00105	0.00104	0.2077	0.1099
9	3.0000	0	27.0000	0	0.7407	0.9353	0.00661	0.00849	0.0184	-0.0374	-0.00176	-0.0273	0.00367	0.00365	0.8783	0.5522
10	2.0000	0	14.0000	1.0000	0.3805	0.5201	0.00446	0.0211	-0.0131	-0.0156	-0.0166	-0.00051	0.000651	0.000648	0.2711	0.1455

利用 Hat 矩陣對角、信賴區間置換 C、Intercept DfBeta 的絕對值、Pclass DfBeta 的絕對值、Sexmale DfBeta 的絕對值、Age DfBeta 的絕對值、SibSp DfBeta 的絕對值、Delta 偏差、Delta 卡方大於 3 則認為是影響點的準則，挑出了以下共 62 筆資料

Obs	Pclass	Sex	Age	SibSp	h	c	dfbeta0	dfbeta1	dfbeta2	dfbeta3	dfbeta4	difchisq	difdev
1	2	male	34.0	0	0.002524	0.00881	-0.00710	0.00501	0.04877	0.01536	-0.01928	3.4837	3.00577
2	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
3	2	female	27.0	1	0.003808	0.01580	-0.07269	0.04971	0.09065	0.02386	-0.01620	4.1322	3.28066
4	3	male	29.7	1	0.002083	0.02830	-0.10717	0.11098	0.10840	0.05657	0.07358	13.5544	5.38014
5	3	male	32.0	0	0.002288	0.02382	-0.08772	0.10946	0.08520	0.04904	-0.01913	10.3880	4.88472
6	3	male	29.0	0	0.002349	0.02171	-0.07082	0.09958	0.08212	0.02747	-0.02393	9.2197	4.66604
7	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
8	3	male	12.0	1	0.003927	0.02647	-0.00943	0.05439	0.08923	-0.06694	0.04027	6.7142	4.10564
9	3	male	24.0	0	0.002658	0.02014	-0.04332	0.08334	0.07690	-0.00743	-0.03162	7.5586	4.30926
10	3	male	27.0	0	0.002437	0.02080	-0.05971	0.09304	0.08004	0.01334	-0.02706	8.5151	4.52214
11	3	male	9.0	0	0.006445	0.02710	0.03182	0.03741	0.06084	-0.10116	-0.05148	4.1770	3.30495
12	1	female	50.0	0	0.005777	0.04450	-0.07629	0.10152	0.13423	-0.06631	0.04353	7.6577	4.35089
13	2	female	24.0	0	0.003863	0.02583	-0.12217	0.06778	0.12161	0.06771	0.07810	6.6594	4.09085
14	3	male	18.0	0	0.003535	0.02114	-0.01174	0.06436	0.07054	-0.04718	-0.04021	5.9584	3.89490
15	3	female	2.0	0	0.009590	0.04819	-0.14334	0.02384	0.10465	0.16688	0.10298	4.9771	3.60757
16	3	male	26.0	0	0.002497	0.02049	-0.05421	0.08979	0.07900	0.00635	-0.02860	8.1834	4.45078

17	3	male	16.0	0	0.003991	0.02206	-0.00162	0.05819	0.06840	-0.05981	-0.04290	5.5051	3.76038
18	3	male	3.0	4	0.010546	0.14729	-0.05190	0.05002	0.15956	-0.06781	0.32168	13.8198	5.51790
19	3	male	25.0	1	0.002210	0.02490	-0.08026	0.09560	0.10348	0.02224	0.06456	11.2437	5.03081
20	3	male	25.0	0	0.002571	0.02027	-0.04875	0.08656	0.07795	-0.00057	-0.03012	7.8647	4.37981
21	3	male	19.0	0	0.003341	0.02078	-0.01688	0.06747	0.07161	-0.04074	-0.03884	6.1991	3.96286
22	3	male	30.0	0	0.002320	0.02231	-0.07642	0.10286	0.08315	0.03461	-0.02235	9.5937	4.73856
23	2	male	42.0	0	0.002983	0.01434	-0.04656	0.02657	0.05721	0.06782	-0.00827	4.7918	3.52223
24	1	female	2.0	1	0.003199	0.11575	-0.32754	0.25505	0.16023	0.26631	0.03292	36.0678	7.33464
25	3	male	29.7	2	0.002792	0.05431	-0.14041	0.11965	0.13441	0.08146	0.17438	19.3986	6.07977
26	2	female	26.0	1	0.003795	0.01638	-0.07803	0.05269	0.09199	0.03035	-0.01482	4.2999	3.34552
27	3	male	45.0	0	0.002546	0.04448	-0.16343	0.15298	0.09811	0.14643	0.00296	17.4305	5.86756
28	3	male	3.0	1	0.007298	0.03462	0.03494	0.02777	0.07888	-0.12152	0.02445	4.7090	3.50638
29	2	female	38.0	0	0.004761	0.01827	-0.04593	0.02616	0.10208	-0.02462	0.05061	3.8184	3.15549
30	3	male	21.0	0	0.003013	0.02028	-0.02732	0.07377	0.07374	-0.02763	-0.03601	6.7104	4.10011
31	3	male	39.0	0	0.002365	0.03254	-0.12809	0.13280	0.09224	0.10082	-0.00745	13.7259	5.40726
32	3	male	44.0	0	0.002513	0.04219	-0.15750	0.14961	0.09714	0.13877	0.00120	16.7500	5.79011
33	3	female	10.0	0	0.008113	0.02957	-0.09754	0.00220	0.09329	0.10782	0.08504	3.6147	3.07508
34	3	male	32.0	0	0.002288	0.02382	-0.08772	0.10946	0.08520	0.04904	-0.01913	10.3880	4.88472
35	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
36	3	male	29.0	0	0.002349	0.02171	-0.07082	0.09958	0.08212	0.02747	-0.02393	9.2197	4.66604
37	3	male	9.0	1	0.004789	0.02870	0.00588	0.04529	0.08582	-0.08592	0.03488	5.9638	3.90180
38	1	female	25.0	1	0.003400	0.04928	-0.18799	0.17608	0.13695	0.08864	-0.00828	14.4452	5.51734
39	3	male	26.0	0	0.002497	0.02049	-0.05421	0.08979	0.07900	0.00635	-0.02860	8.1834	4.45078
40	3	male	29.0	0	0.002349	0.02171	-0.07082	0.09958	0.08212	0.02747	-0.02393	9.2197	4.66604
41	2	male	32.0	1	0.002802	0.01293	-0.02451	0.00541	0.07069	0.02391	0.06222	4.6024	3.45467
42	3	male	22.0	0	0.002878	0.02015	-0.03261	0.07694	0.07479	-0.02096	-0.03457	6.9819	4.16940
43	3	male	32.0	0	0.002288	0.02382	-0.08772	0.10946	0.08520	0.04904	-0.01913	10.3880	4.88472
44	2	male	62.0	0	0.005540	0.05932	-0.15961	0.08827	0.07791	0.21764	0.02483	10.6475	4.95902
45	3	male	32.0	0	0.002288	0.02382	-0.08772	0.10946	0.08520	0.04904	-0.01913	10.3880	4.88472
46	1	male	60.0	1	0.010689	0.04729	-0.07863	-0.01931	0.06229	0.16157	0.09840	4.3769	3.39352
47	3	male	20.0	1	0.002569	0.02374	-0.05229	0.07947	0.09811	-0.01320	0.05508	9.2184	4.66743
48	1	male	80.0	0	0.015059	0.10417	-0.15986	0.03001	0.05911	0.28898	0.04664	6.8130	4.18775
49	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
50	1	male	50.0	2	0.012114	0.05163	-0.05661	-0.03997	0.07568	0.11753	0.16759	4.2107	3.33276
51	3	male	20.0	1	0.002569	0.02374	-0.05229	0.07947	0.09811	-0.01320	0.05508	9.2184	4.66743
52	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
53	3	male	29.7	1	0.002083	0.02830	-0.10717	0.11098	0.10840	0.05657	0.07358	13.5544	5.38014
54	3	male	31.0	0	0.002300	0.02302	-0.08206	0.10616	0.08417	0.04180	-0.02074	9.9829	4.81146
55	3	male	6.0	0	0.007983	0.02988	0.04509	0.02895	0.05759	-0.11732	-0.05472	3.7127	3.11755
56	3	male	20.0	0	0.003167	0.02049	-0.02207	0.07061	0.07267	-0.03422	-0.03744	6.4496	4.03126
57	3	male	1.0	1	0.008407	0.03691	0.04410	0.02216	0.07654	-0.13261	0.02110	4.3536	3.37843
58	3	male	27.0	0	0.002437	0.02080	-0.05971	0.09304	0.08004	0.01334	-0.02706	8.5151	4.52214
59	3	male	27.0	0	0.002437	0.02080	-0.05971	0.09304	0.08004	0.01334	-0.02706	8.5151	4.52214
60	3	male	29.7	0	0.002328	0.02212	-0.07474	0.10188	0.08284	0.03246	-0.02283	9.4799	4.71677
61	3	male	32.0	0	0.002288	0.02382	-0.08772	0.10946	0.08520	0.04904	-0.01913	10.3880	4.88472
62	3	male	4.0	1	0.006798	0.03352	0.03026	0.03062	0.08005	-0.11582	0.02615	4.8977	3.57110

可以看出大部份資料都是在 Delta 偏差及 Delta 卡方大於 3

去除這 62 筆資料後，以剩下的 829 筆資料，且同樣是最終選擇的模型，重新做一次 Hosmer and Lemeshow test

Hosmer 和 Lemeshow 配適度檢定		
卡方	DF	Pr > ChiSq
11.9143	8	0.1551

H_0:$logit(\pi) = \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp
H_1:$logit(\pi) \neq \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp
p-value=0.1551>0.05 故不拒絕 H_0 ，在$\alpha = 0.05$ 下，沒有充分證據顯示模型配適的不好，因此以此資料配出的模型為最後選定的模型

六、 結論

最大概度估計值分析						
參數		DF	估計值	標準誤差	Wald 卡方	Pr > ChiSq
Intercept		1	11.0675	1.0189	117.9970	<.0001
Pclass		1	-2.9042	0.2841	104.4977	<.0001
Sex	male	1	-5.7534	0.4890	138.4531	<.0001
Age		1	-0.0725	0.0115	39.8162	<.0001
SibSp		1	-0.5268	0.1446	13.2750	0.0003

最後選取模型:
$logit(\pi) = \alpha + \beta_1$pclass+$\beta_2$sex(male)+$\beta_3$age+$\beta_4$SibSp
預測模型:(以去除影響點後的 829 筆資料配適)
$logit(\hat{\pi}) = 11.0675$-2.9042pclass-5.7534sex(male)-0.0725age-0.5268SibSp

模型參數的解釋:
α之估計值=11.0675
β_1之估計值=-2.9042 代表當其他變數固定不變時，pclass 每增加一單位，存活的 odds 會是原先的 exp(-2.9042) = 0.0548 倍
β_2之估計值=-5.7534 代表當其他變數固定不變時，男生存活的 odds 是女生的 exp(-5.7534)= 0.0032 倍
β_3之估計值=-0.0725 代表當其他變數固定不變時，age 每增加一單位，存活的 odds 會是原先的 exp(-0.0725) = 0.9301 倍
β_4之估計值=-0.5268 代表當其他變數固定不變時，sibsp 每增加一單位，存活的 odds 會是原先的 exp(-0.5268) = 0.5905 倍

由上述可知船艙等級及性別影響較大，且等級越低的船艙及男生的存活率都較低，正好印證了一開始樣本資料所給的描述，而年齡越大確實會降低生存率，但並不會很明顯的降低，還有船上兄弟姊妹和配偶數量越多，也會造成生存率下降，這點比較無法從樣本資料中看出，總結，以此樣本推測，若想提高生存率要在等級較高的船艙，且為女性，越年輕及船上兄弟姊妹和配偶數量越少越好，這也相當符合現實狀況。

七、 SAS CODE

```
PROC IMPORT OUT= WORK.d1
            DATAFILE= "C:\Users\107354002\Desktop\train.xlsx"
            DBMS=EXCEL REPLACE;
     RANGE="train$";
     GETNAMES=YES;
     MIXED=NO;
     SCANTEXT=YES;
     USEDATE=YES;
     SCANTIME=YES;
RUN;
data d2;
set WORK.d1;
if age=. then age=29.7;/*平均差補*/
n=1;
/*敘述統計*/
proc chart data=d2;
hbar Survived Pclass Sex SibSp Parch Embarked/discrete;
proc chart data=d2;
hbar Pclass Sex SibSp Parch Embarked/discrete group=Survived;
proc sgplot data=d2;
vbox age;
proc sgplot data=d2;
vbox Fare;
proc sgplot data=d2;
vbox age/category=Survived;
proc sgplot data=d2;
vbox Fare/category=Survived;
proc means data=d2    n mean median std min max sum;
var Survived Pclass Age SibSp Parch Fare;
```

```
/*原始模型*/
  PROC LOGISTIC DATA=d2;
CLASS Pclass(REF='3') Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass|Sex|Age|SibSp|Parch|Fare|Embarked@2;
  PROC LOGISTIC DATA=d2;
CLASS Pclass(REF='3') Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Parch Fare Embarked;
  PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Parch Fare Embarked/LRCL
WALDCL;
  PROC GENMOD DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived/n=Pclass Sex Age SibSp Parch Fare Embarked/DIST=BIN LINK=LOGIT
TYPE3;
/*VIF*/
proc reg data=d2;
model Survived=Pclass Age SibSp Parch Fare/vif;
output out=outlier h=h;
/*選模*/
  PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Parch Fare
Embarked/SELECTION=BACKWARD;
  PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Parch Fare
Embarked/SELECTION=STEPWISE;
  PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Parch Fare
Embarked/SELECTION=FORWARD;
proc glmselect data=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived=Pclass Sex Age SibSp Parch Fare Embarked/selection=lasso;
/*ROC CURVE*/
PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female')/PARAM=REF;
```

```
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp/OUTROC=roc1;
PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female') Embarked(REF ='S')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp Embarked/OUTROC=roc2;
/*模型診斷及校正*/
PROC LOGISTIC DATA=d2;
CLASS Sex(REF ='female')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp/lackfit influence;
OUTPUT OUT=predict H=h C=c DFBETAS=dfbeta0 dfbeta1 dfbeta2 dfbeta3 dfbeta4
DIFCHISQ=difchisq DIFDEV=difdev;
DATA predict2;
SET predict;
IF difchisq>3 & difdev>3 then output;
PROC PRINT DATA=predict2;
VAR Pclass Sex Age SibSp h c dfbeta0 dfbeta1 dfbeta2 dfbeta3 dfbeta4 difchisq
difdev;
DATA predict3;
SET predict;
IF difchisq>3 & difdev>3 then delete;
PROC PRINT DATA=predict3;
VAR Pclass Sex Age SibSp h c dfbeta0 dfbeta1 dfbeta2 dfbeta3 dfbeta4 difchisq
difdev;
PROC LOGISTIC DATA=predict3;
CLASS Sex(REF ='female')/PARAM=REF;
MODEL Survived(EVENT='1')=Pclass Sex Age SibSp/lackfit;
```