**Logistic迴歸模型**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 社會科學充斥大量類別資料，例如性別可分為男性、女性；政治偏好可分為激進、中立、保守；宗教信仰可分為佛教、道教、基督教、回教等。  類別資料為自變數時，可以利用卡方分析或平均數檢定來進行分析，但如果要測量收入水準對政治偏好的影響，或身高體重與性別的關係時，  常受限於應變數為類別資料而缺乏適當分析工具。Logistic迴歸模型正是用以處理應變數屬於類別資料的情況，可說是社會科學最重要的統計模型。  Logistic迴歸則以Binomial分配為起點，假設應變數為二項式分配，線性模型為：  [https://sites.google.com/site/rlearningsite/_/rsrc/1445333866692/catagory/logit/logit%20function.png?height=68&width=320](https://sites.google.com/site/rlearningsite/catagory/logit/logit%20function.png?attredirects=0)  指數轉換後可改寫為：  [https://sites.google.com/site/rlearningsite/_/rsrc/1445333866692/catagory/logit/logit%20exp%20function.png](https://sites.google.com/site/rlearningsite/catagory/logit/logit%20exp%20function.png?attredirects=0)   1. **基本函數**  |  |  | | --- | --- | | **glm()** | 一般線性模型 | | **predict.glm()** | 一般線性估計 | | **exp()** | 計算自然指數 | | **contrasts()** | 回傳矩陣的對照值，搭配contr.treatment()可設定虛變數 | | **contr.treatment()** | 設定虛變數的參考 |   **Logistic迴歸模型** 鐵達尼號(RMS Titanic)於1912年4月10日載滿一千多名乘客，由英國南安普敦(Southampton)出發，途中經法國瑟堡(Cherbourg)及  愛爾蘭皇后鎮(Queenstown)，預定前往美國紐約。4月14日晚間11時不幸於航行途中撞上冰山，並於15日凌晨二時沉沒於大西洋外海，  是近代最嚴重的海難之一。  業者透過差別定價獲取最大利潤，搭飛機的時候是如此，像鐵達尼號這樣的豪華郵輪也不例外。根據歷史資料，鐵達尼號當年的票價也採取了  差別訂價的策略，換句話說傑克的三等艙票價比較便宜，蘿絲的頭等艙票價比較貴，令人驚訝的是，票價竟然也與生死有關。  由於鐵達尼號確切的搭乘名單及生還統計數字不一，直接採用 <http://www.encyclopedia-titanica.org/> 的歷史資料，並將所有乘客姓名、年齡、船票  等資訊整理成[titanic.csv](https://sites.google.com/site/rlearningsite/data/titanic.csv?attredirects=0)檔，檔案可從[範例檔](https://sites.google.com/site/rlearningsite/data)專區直接下載(註1)。   |  | | --- | | > titanic<-read.csv("c:/downloads/titanic.csv", header=T, sep=",") #讀取titanic.csv |   titanic.csv包含鐵達尼號乘客及所有機組人員清單，只需分析票價與生還間的關係，因此先將機組人員排除在外(註2)。   |  | | --- | | > titanic\_passenger<-titanic[1:1317,] #只讀取1317名乘客清單 > attach(titanic\_passenger) #將資料存入記憶體 > titanic\_passenger[fare==9999,"fare"]<-NA #設定票價的遺漏值為9999 > titanic\_passenger[age==9999,"age"]<-NA #設定年齡的遺漏值為9999 |   遺漏值設定完畢後，以survival(1=獲救生還)當作應變數，fare(英鎊)當作自變數進行Logistic迴歸分析，  注意要加上 na.action=na.exclude 排除遺漏值(註3)。   |  | | --- | | > model1<-glm(formula=survival~fare, data=titanic\_passenger, family=binomial(link="logit"), na.action=na.exclude) #將Logistic模型分析的結果指定給model1 > summary(model1) # 輸出分析結果   Call: glm(formula = survival ~ fare, family = binomial(link = "logit"),         data = titanic\_passenger, na.action = na.exclude)  Deviance Residuals:       Min       1Q        Median       3Q      Max   -2.2790  -0.8817  -0.8486   1.3470   1.5703    Coefficients:                      Estimate   Std. Error   z value   Pr(>|z|)     (Intercept) -0.927778   0.076671 -12.101  < 2e-16 \*\*\* fare               0.013108   0.001646    7.961   1.7e-15 \*\*\* --- Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1   (Dispersion parameter for binomial family taken to be 1)  Null deviance: 1710.1  on 1290  degrees of freedom Residual deviance: 1617.2  on 1289  degrees of freedom   (26 observations deleted due to missingness) AIC: 1621.2  Number of Fisher Scoring iterations: 4 |  1. **解釋Logistic迴歸模型** 鐵達尼號乘客獲救與票價之間的模型如下：   [https://sites.google.com/site/rlearningsite/_/rsrc/1445333866692/catagory/logit/logit%20exp%20result.png](https://sites.google.com/site/rlearningsite/catagory/logit/logit%20exp%20result.png?attredirects=0)  0.013為正數，代表票價與生還具有正相關。從csv檔中知道票價最便宜的是3英鎊，最貴的是512英鎊。  根據模型，付出3英鎊的乘客生還機率是：  [https://sites.google.com/site/rlearningsite/_/rsrc/1445333866692/catagory/logit/fare3result.png](https://sites.google.com/site/rlearningsite/catagory/logit/fare3result.png?attredirects=0)  付出512英鎊乘客的生還機率是：  [https://sites.google.com/site/rlearningsite/_/rsrc/1445333866692/catagory/logit/fare512result.png](https://sites.google.com/site/rlearningsite/catagory/logit/fare512result.png?attredirects=0)  透過predict.glm()函數快速獲得上面數學式的計算結果：   |  | | --- | | > predict.glm(model1, type="response", newdata=data.frame(fare=3))         1  0.2914288  > predict.glm(model1, type="response", newdata=data.frame(fare=512))         1 0.9969312 |   由此可見票價越高，生還的機率也越高。透過指數換算，鐵達尼號乘客每多付出1英磅，就增加1%的生存機會。   |  | | --- | | > exp(0.013) [1] 1.013085 |  1. **多元Logistic迴歸模型與虛變數** 常見的迴歸模型通常包含多項自變數，Logistic迴歸也不例外。在自變數中更常必須以虛變數來進行分析， 2. 這裡仍然以鐵達尼號為例，分析生還者與艙等、性別間的關係。  [titanic.csv](https://sites.google.com/site/rlearningsite/data/titanic.csv?attredirects=0)檔中乘客艙等共分為頭等艙、二等艙與三等艙，分別以1、2、3來表示；性別則有男性與女性，分別以0、1來表示 3. 。假設乘客所搭乘的艙等與性別為自變數，以多元Logistic迴歸分析性別是否影響獲救機會。艙等與性別是類別變數，先設定為虛變數。  |  | | --- | | > classf<-factor(class) #設定classf為類別變數 > genderf<-factor(gender) #設定genderf為類別變數 > is.factor(classf) #確定設定結果 > is.factor(genderf) #確定設定結果 |  1. 完成類別變數設定後，接來以contrasts()配合contr.treatment()，設定虛變數參考點，以三等艙及男性為參考點。  |  | | --- | | > contrasts(classf)<-contr.treatment(3, base=3)   #以contrast()設定虛變數classf，contr.treatment()設定三個艙等中，以編碼3當作參考點。 |  1. 由於genderf的編碼為0、1，因此無須額外設定虛變數。虛變數設定完成後利用glm()函數進行Logistic迴歸分析：  |  | | --- | | > model2<-glm(formula=survival~classf+genderf, data=titanic\_passenger, family=binomial(link="logit"), na.action=na.exclude) > summary(model2)  Call: glm(formula = survival ~ classf + genderf, family = binomial(link = "logit"),      data = titanic\_passenger, na.action = na.exclude)  Deviance Residuals:      Min       1Q         Median       3Q      Max   -2.0728  -0.6706  -0.4861   0.8247   2.0949    Coefficients:                       Estimate   Std. Error   z value    Pr(>|z|)     (Intercept)     -2.0761     0.1240      -16.75    < 2e-16 \*\*\* classf1           1.8188     0.1677       10.84     < 2e-16 \*\*\* classf2           0.6984     0.1724         4.05   5.12e-05 \*\*\* genderf1        2.2815     0.1433       15.92     < 2e-16 \*\*\* --- Signif. codes:  0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1   (Dispersion parameter for binomial family taken to be 1)  Null deviance: 1736.4  on 1316  degrees of freedom Residual deviance: 1301.3  on 1313  degrees of freedom AIC: 1309.3  Number of Fisher Scoring iterations: 4 |  1. summary()輸出分析結果後可以發現，艙等與性別都達到顯著水準。  |  | | --- | | > exp(1.8188)#頭等艙乘客與下等艙乘客相較的生還機率 [1] 6.164457  > exp(0.6984) #中等艙乘客與下等艙乘客相較的生還機率 [1] 2.010533  > exp(2.2815) #女性乘客與男性乘客相較的生還機率  [1] 9.791356 |  1. 經過指數換算後可以發現，頭等艙乘客的生還機率遠高於三等艙乘客六倍，二等艙乘客則高出兩倍，女性乘客獲救機率也高出男性九倍。  事實上，船難發生時，船長[Edward Smith](http://en.wikipedia.org/wiki/Edward_Smith_(sea_captain))下令讓婦女優先登上救生艇(women and children first)，由於登船順序以頭等艙乘客優先， 2. 因此下等艙乘客多數被困在船艙內無法登上救生艇(註4)。這樣的結果從交叉表來看更為明顯：  |  | | --- | | > classtab<-table(class, survival) > prop.table(classtab, 1)                   survival class                 0               1   1           0.3734568   0.6265432   2           0.6070175   0.3929825   3           0.7556497   0.2443503  > gendertab<-table(gender, survival) > prop.table(gendertab, 1)                    survival gender         0                   1  0           0.7995338    0.2004662  1            0.3115468     0.6884532 |  1. **備註**  * 註1：鐵達尼號旅客清單取材至  <http://www.encyclopedia-titanica.org/>。 * 註2：包含設計師[Thomas Andrews](http://en.wikipedia.org/wiki/Thomas_Andrews_(shipbuilder))、8名船上樂隊成員及部分船員因為有正式客房，因此排除在機組人員外。 * 註3：因1912年英國幣制仍使用先令(Shillin)，20先令等於1英鎊，為計算方便，但船票票價係直接取英鎊整數。先令於1971年停止流通。 * 註4：Smith船難發生後，下令women and children first，卻引發兩位大副[Charles Lightoller](http://en.wikipedia.org/wiki/Charles_Lightoller)與[William Murdoch](http://en.wikipedia.org/wiki/William_McMaster_Murdoch)不同解讀。 * 左舷的Lightoller解讀為「women and children only」，右舷的Murdoch解讀為「women and children first, then others」。 * 因此Murdoch在婦幼登艇後，開放讓男性登艇，Lightoller則嚴格遵守只讓婦幼登艇。這個決定性的差異，也導致右舷獲救的乘客較多。 |