



Fair's Affairs

107354014 粘明揚

目錄

CONTENT

01

資料來源

02

資料解讀

03

資料分析

04

結論

第一部分

資料來源

資料來源

<https://www.kaggle.com/clarkchong/fairs-affairs-dataset#Affairs.csv>

The screenshot shows the Kaggle dataset page for 'Fair's Affairs' dataset. The header includes the Kaggle logo, a search bar, and navigation links: Competitions, Datasets, Kernels, Discussion, and Learn. The dataset title is 'Fair's "Affairs" dataset' with the subtitle 'Are certain groups more likely to have extramarital affairs?'. It is by Fan Fei Chong, updated a year ago (Version 1). There are 2 voters and a share button. The dataset is 5 KB and has a 'New Kernel' button. The 'Data Sources' section shows 'Affairs.csv' (601 x 10). The 'About this file' section states 'No description yet'. The 'Columns' section lists: #, # affairs, A gender, # age, # yearsmarried, A children, # religiousness, # education, and # occupation.

Dataset

Fair's "Affairs" dataset
Are certain groups more likely to have extramarital affairs?

Fan Fei Chong • updated a year ago (Version 1)

2 voters

share

Data Overview Kernels (1) Discussion Activity

Download (5 KB) New Kernel

Data (5 KB)

API kaggle datasets download -d clarkchong/fairs-aff... ? Download All

Data Sources

File	Size
Affairs.csv	601 x 10

About this file

No description yet

Columns

- #
- # affairs
- A gender
- # age
- # yearsmarried
- A children
- # religiousness
- # education
- # occupation



第二部分

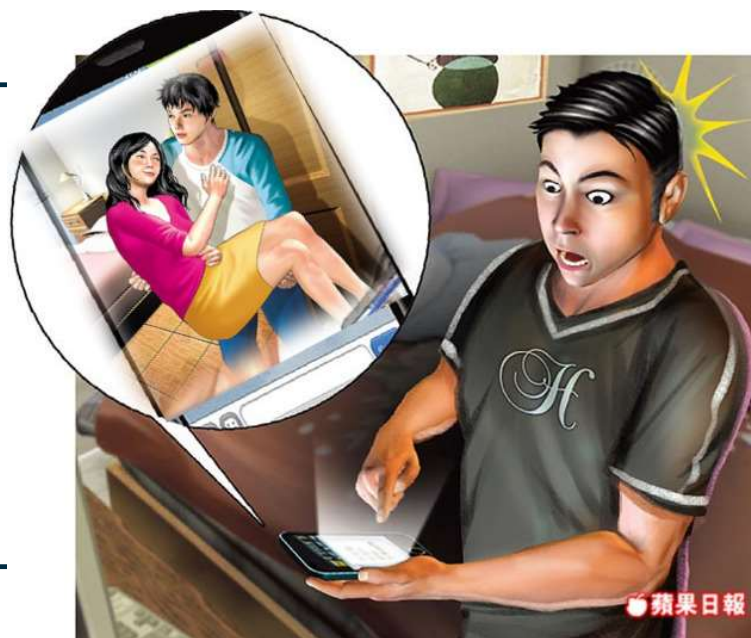
資料解讀

資料解讀

定義描述

出軌

法律上，有性關係為出軌必要因素(雙方自願)
僅有思想或行為不貞，為精神出軌



資料解讀

數據描述

取自於1969年(Psychology Today)所做的一個調查



affairs
過去一年出軌
次數



gender
性別(男=0女=1)



age
年齡



yearsmarried
結婚年數



children
是否有小孩(有=1無=0)



education
教育程度(20分滿分)



religiousness
宗教信仰程度(5分滿分)



occupation
職業種類



rating
婚姻滿意度(5分滿分)

資料解讀

將類別變數轉成數值型變數

```
15 table(Affairs$affairs)
16 Affairs$affairs[Affairs$affairs >= 1] <- 1 將出軌次數>0者令為
17 a <- sub("female",1,Affairs$gender)
18 b <- sub("male",0,a) 1(有出軌)
19 Affairs$gender <- a
20 Affairs$gender <-b 女生令為0，男生令為1
21 as.numeric(Affairs$gender)
22
23 table(Affairs$children)
24 c <- sub("yes",1,Affairs$children)
25 d<- sub("no",0,c) 有小孩令為1，無小孩令為0
26 Affairs$children <- c
27 Affairs$children<-d
23 Affairs$occupation <-as.factor(Affairs$occupation)
```

令occupation中的數值為
類別型(dummy variable)



第三部分

資料分析

資料分析

羅吉斯回歸

觀察模型及變數是否顯著

篩選重要變數

子集法: Stepwise Regression



配適模型好壞

ROC Curve

AUC

資料分析

羅吉斯回歸

羅吉斯回歸(尚未篩選變數)

有些變數對模型的貢獻並不顯著，於是想篩選變數，能否用簡單模型就可達到與原模型的表現與效果

```
Call:
glm(formula = affairs ~ ., data = train[, 2:10])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7588 -0.2870 -0.1517  0.4072  1.0468

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.726838    0.205164   3.543 0.000436 ***
gender1        0.014745    0.048659   0.303 0.762007
age          -0.006731    0.003435  -1.959 0.050667 .
yearsmarried   0.012804    0.006375   2.008 0.045167 *
children1      0.109777    0.054529   2.013 0.044670 *
religiousness -0.059472    0.016500  -3.604 0.000347 ***
education      0.005934    0.010297   0.576 0.564706
occupation2    0.167727    0.153641   1.092 0.275539
occupation3    0.085403    0.085153   1.003 0.316413
occupation4    0.144831    0.076974   1.882 0.060517 .
occupation5    0.003673    0.060927   0.060 0.951950
occupation6    0.072683    0.076470   0.950 0.342367
occupation7    0.174566    0.145908   1.196 0.232142
rating        -0.097584    0.018680  -5.224 2.64e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1755021)

    Null deviance: 94.865  on 480  degrees of freedom
Residual deviance: 81.959  on 467  degrees of freedom
AIC: 543.82

Number of Fisher Scoring iterations: 2
```

資料分析

篩選重要變數

Stepwise Regression

採用Stepwise Regression中的Backward Stepwise
在完整回歸中，逐一移除變數，直到移除任一變數，
模型會損失過多解釋力時，則停止。

在Backward Stepwise中挑選的變數，為被留下的重
要變數

```
70 #Backward Stepwise regression
71 # 1. 先建立一個完整的線性迴歸
72 full = glm(affairs ~ ., data = train[,2:10])
73
74 # 2. 用'step()'，一個一個把變數移除，看移除哪個變數後 AIC 下降最多
75 backward.glm = step(full,
76                      scope = list(upper=full),
77                      direction="backward")
78 summary(backward.glm)
```

Call:
glm(formula = affairs ~ age + yearsmarried + children + religiousness +
rating, family = binomial(), data = train)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7237	-0.7951	-0.5714	1.0059	2.4027

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	1.92183	0.69562	2.763	0.005732	**
age	-0.03089	0.01889	-1.636	0.101899	
yearsmarried	0.06081	0.03464	1.755	0.079180	.
children1	0.73318	0.32237	2.274	0.022946	*
religiousness	-0.34394	0.09461	-3.635	0.000278	***
rating	-0.50726	0.10077	-5.034	4.81e-07	***

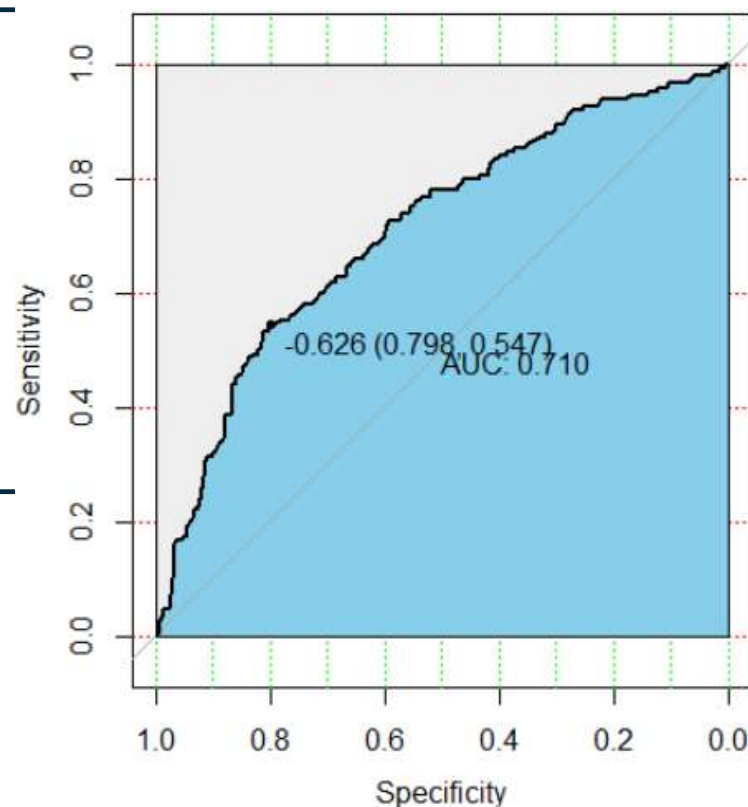
資料分析

配適模型好壞

```
94 expect <- fit.backward.a[  
95 #ROC Curve  
96 install.packages("pROC")  
97 library(pROC)  
98  
99 pre <- predict(fit.backward,Affairs)  
100 modelroc <- roc(Affairs$affairs,pre)  
101 plot(modelroc, print.auc=TRUE, auc.polygon=TRUE, grid=c(0.1, 0.2),  
102       grid.col=c("green", "red"), max.auc.polygon=TRUE,  
103       auc.polygon.col="skyblue", print.thres=TRUE)
```

ROC Curve AUC

可以看到AUC值為0.71，表現還算尚可



第四部分

結論

結論

模型解讀

配適模型解讀

```
> coef(fit.backward)
(Intercept)      age yearsmarried    children1 religiousness      rating
  1.92182841 -0.03089141  0.06080538  0.73318263  -0.34393790  -0.50725840
> #odds
> exp(coef(fit.backward))
(Intercept)      age yearsmarried    children1 religiousness      rating
  6.8334414  0.9695809  1.0626921  2.0816953   0.7089730   0.6021442
```

```
93 coef(fit.backward)
94 #odds
95 exp(coef(fit.backward))
```

在羅吉斯回歸中，迴歸係數的含義是當其他預測變量不變，一單位預測變量的變化可引起對數勝算比的變化，對其指數化即為勝算比的變化。例如：婚齡每增加一年，婚外情勝算比會乘以1.06

結論

模型預測

模型預測

X	affairs	gender	age	yearsmarried	children	religiousness	education	occupation	rating	prob
4	0	0	37.0	10.000	0	3	18	7	4	0.15789862
55	0	0	27.0	4.000	1	4	18	6	4	0.20740801
86	0	1	27.0	4.000	0	4	14	5	4	0.11166920
115	0	1	22.0	1.500	0	2	16	5	5	0.13115985
139	0	1	22.0	1.500	0	2	18	5	5	0.13115985
155	0	1	27.0	4.000	1	3	17	5	4	0.26959413
192	0	0	22.0	1.500	1	1	14	3	5	0.30711925
224	0	1	27.0	4.000	1	2	18	6	1	0.70454389
320	0	0	37.0	10.000	1	3	20	6	4	0.28074592
321	0	1	22.0	0.750	0	2	16	5	5	0.12604980
334	0	0	32.0	7.000	0	4	20	6	4	0.11447297
355	0	0	37.0	10.000	1	4	20	6	4	0.21675082
362	0	1	22.0	1.500	0	5	16	5	5	0.05104974

```
107 test$prob <- predict(fit.backward,  
108                       newdata=test,  
109                       type="response")  
110 test|
```


感謝觀看

