資料處理作業第十一&十二週

統計二 108304010 趙啟宏

**EX\_A:**

H0:mu甲=mu乙=mu丙

H1:not H0

ANOVA Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Sum of squares | df | Mean square | F |
| Between group | 173.17 | 2 | 86.57 | 13.32 |
| Within group | 58.5 | 9 | 6.5 |  |
| Total | 231.67 | 11 |  |  |

**由F-table可得知，拒絕域為F>4.26，13.32>4.26，所以我們拒絕H0。因此，在顯示水準等於0.05的情況下，有足夠的證據顯示三個班級的每周平均看電視時數有差異。**

EX\_B:

##讀入資料##

tscs951 <- read\_sav('tscs951.sav')

tscs001 <- read\_sav('tscs001.sav')

tscs051 <- read\_sav('tscs051.sav')

tscs101 <- read\_sav('tscs101.sav')

##調整歸因分數變項##

tscs951$v54x <- recode(tscs951$v54,'1=4;2=3;3=2;4=1;else=NA')

tscs001$v60x <- recode(tscs001$v60,'1=4;2=3;3=2;4=1;else=NA')

tscs051$v39ex <- recode(tscs051$v39e,'1=4;2=3;3=2;4=1;else=NA')

tscs101$v39ex <- recode(tscs101$v39e,'1=4;2=3;3=2;4=1;else=NA')

##調整教育年度##

tscs951$edu <- recode(tscs951$v7a,'1:5=1;6:13=2;14:19=3;else=NA')

tscs001$edu <- recode(tscs001$v7a,'1:5=1;6:13=2;14:19=3;else=NA')

tscs051$edu <- recode(tscs051$v7,'1:5=1;6:13=2;14:21=3;else=NA')

tscs101$edu <- recode(tscs101$v7a,'1:5=1;6:13=2;14:21=3;else=NA')

tscs951$edu <- factor(tscs951$edu,levels = c(1,2,3),labels = c('國中以下','高中職','大專以上'))

tscs001$edu <- factor(tscs001$edu,levels = c(1,2,3),labels = c('國中以下','高中職','大專以上'))

tscs051$edu <- factor(tscs051$edu,levels = c(1,2,3),labels = c('國中以下','高中職','大專以上'))

tscs101$edu <- factor(tscs101$edu,levels = c(1,2,3),labels = c('國中以下','高中職','大專以上'))

##擷取成年人資料##

tscs951\_cut <- tscs951[which(tscs951$age>19 & tscs951$age<70),]

tscs001\_cut <- tscs001[which(tscs001$age>19 & tscs001$age<70),]

tscs051\_cut <- tscs051[which(tscs051$age>19 & tscs051$age<70),]

tscs101\_cut <- tscs101[which(tscs101$age>19 & tscs101$age<70),]

##擷取歸因分數+年度##

site1 <- tscs951\_cut$v54x

site2 <- tscs001\_cut$v60x

site3 <- tscs051\_cut$v39ex

site4 <- tscs101\_cut$v39ex

ex\_b1 <- data.frame(Y=c(site1,site2,site3,site4),Site=factor(rep(c('1995','2000','2005','2010'),times=c(length(site1),length(site2),length(site3),length(site4)))))

b1 <- aov(Y~Site,data=ex\_b1)

anova(b1)

##擷取歸因分數+教育分組##

new1 <- tscs951\_cut[which(tscs951\_cut$edu=="國中以下"),]

new2 <- tscs951\_cut[which(tscs951\_cut$edu=="高中職"),]

new3 <- tscs951\_cut[which(tscs951\_cut$edu=="大專以上"),]

new4 <- tscs001\_cut[which(tscs001\_cut$edu=='國中以下'),]

new5 <- tscs001\_cut[which(tscs001\_cut$edu=="高中職"),]

new6 <- tscs001\_cut[which(tscs001\_cut$edu=="大專以上"),]

new7 <- tscs051\_cut[which(tscs051\_cut$edu=='國中以下'),]

new8 <- tscs051\_cut[which(tscs051\_cut$edu=="高中職"),]

new9 <- tscs051\_cut[which(tscs051\_cut$edu=="大專以上"),]

new10 <- tscs101\_cut[which(tscs101\_cut$edu=='國中以下'),]

new11 <- tscs101\_cut[which(tscs101\_cut$edu=="高中職"),]

new12 <- tscs101\_cut[which(tscs101\_cut$edu=="大專以上"),]

site5 <- c(new1$v54x,new4$v60x,new7$v39ex,new10$v39ex)

site6 <- c(new2$v54x,new5$v60x,new8$v39ex,new11$v39ex)

site7 <- c(new3$v54x,new6$v60x,new9$v39ex,new12$v39ex)

ex\_b2 <- data.frame(Y=c(site5,site6,site7),Site=factor(rep(c('國中以下','高中職','大專以上'),times=c(length(site5),length(site6),length(site7)))))

b2 <- aov(Y~Site,data = ex\_b2)

anova(b2)

Analysis of Variance Table

Response: Y

Df Sum Sq Mean Sq F value Pr(>F)

Site 3 16.69 5.5623 16 2.306e-10 \*\*\*

Residuals 6928 2408.52 0.3477

Analysis of Variance Table

Response: Y

Df Sum Sq Mean Sq F value Pr(>F)

Site 2 17.98 8.9918 25.878 6.358e-12 \*\*\*

Residuals 6924 2405.93 0.3475

##執行事後檢定##

TukeyHSD(b1,p.adjust.methods='bonf')

TukeyHSD(b2,p.adjust.methods='bonf')

diff lwr upr p adj

2000-1995 0.11960933 0.06812124 0.17109741 0.0000000

2005-1995 0.10954063 0.05913462 0.15994664 0.0000001

2010-1995 0.10403858 0.05251810 0.15555906 0.0000013

2005-2000 -0.01006869 -0.06158408 0.04144670 0.9585614

2010-2000 -0.01557074 -0.06817710 0.03703562 0.8721391

2010-2005 -0.00550205 -0.05704981 0.04604571 0.9927939

diff lwr upr p adj

高中職-大專以上 -0.02465754 -0.06927438 0.01995929 0.3977187

國中以下-大專以上 -0.12157655 -0.16783923 -0.07531387 0.0000000

國中以下-高中職 -0.09691901 -0.13429812 -0.05953990 0.0000000

1995 2000 2005 2010 國中以下 高中職 大專以上

(如果有三條畫這條就好了)

**結論:(1)根據本次分析，我們算出來的p-value遠小於0.05，所以拒絕H0。所以，是有足夠證據顯示歸因分數存在年度別上的差異。而透過事後分析我們可以發現，造成整體上有差異的原因是<1>2000年和2005年 <2>2000年和2010年 <3>2005年和2010年，這三組間的平均數在alpha=0.05下有著顯著差異。**

1. **根據本次分析，我們算出來的p-value遠小於0.05，所以拒絕H0。所以，是有足夠證據顯示歸因分數存在教育程度別上的差異。而透過事後分析我們可以發現，造成整體上有差異的原因是<1>高中職和大專以上這一組的平均數在alpha=0.05的情況下有著顯著差異。**

**EX\_C(.1):**

##讀入資料##

tscs131 <- read\_sav('tscs131.sav')

##調整年齡變項##

tscs131$age <- 102-as.numeric(tscs131$v2y)+1

##調整籍貫&社會地位##

tscs131$home <- recode(tscs131$v5,'1=1;2=2;3=3;4=4;else=NA')

tscs131$score <- recode(tscs131$v84,'1=1;2=2;3=3;4=4;5=5;6=6;7=7;8=8;9=9;10=10;else=NA')

##擷取成年人資料##

tscs131\_cut <- tscs131[which(tscs131$age>19 & tscs131$age<66),]

##擷取性別+籍貫+社會地位##

mydata <- data.frame(sex=tscs131\_cut$v1,home=tscs131\_cut$home,score=tscs131\_cut$score)

##刪除含有遺漏值的樣本##

mydata <- mydata[complete.cases(mydata),]

##製造Dummy variables(sex\_2為女性，home\_1為閩南人)##

install.packages('fastDummies')

library(fastDummies)

mydata <- dummy\_cols(mydata,select\_columns = c('sex','home'))

##建立男性vs社會地位模型##

library(lm.beta)

ex\_c1 <- lm(mydata$score~mydata$sex\_1)

summary(lm.beta(ex\_c1))

##建立籍貫vs社會地位模型(閩南人為參考組)##

ex\_c2 <- lm(mydata$score~mydata$home\_2+mydata$home\_3+mydata$home\_4)

summary(lm.beta(ex\_c2))

##建立籍貫+男性vs社會地位模型(女性閩南人為參考組，無互動項)##

ex\_c3 <- lm(mydata$score~mydata$sex\_1+mydata$home\_2+mydata$home\_3+mydata$home\_4)

summary(lm.beta(ex\_c3))

##建立籍貫+男性vs社會地位模型(女性閩南人為參考組，有互動項)##

mydata$int1 <- mydata$sex\_1\*mydata$home\_1

mydata$int2 <- mydata$sex\_1\*mydata$home\_2

mydata$int3 <- mydata$sex\_1\*mydata$home\_3

mydata$int4 <- mydata$sex\_1\*mydata$home\_4

ex\_c4 <- lm(mydata$score~mydata$sex\_1+mydata$home\_2+mydata$home\_3+mydata$home\_4+mydata$int1+mydata$int2+mydata$int3)

summary(lm.beta(ex\_c4))

─────────────────────────────────────────────────────────────────────────────────

(1) (2) (3) (4)

─────────────────────────────────────────────────────────────────

(Intercept) 4.690 \*\*\* 4.582 \*\*\* 4.623 \*\*\* 4.651 \*\*\*

(0.060) (0.048) (0.064) (0.068)

mydata$sex\_1 -0.074 -0.078 0.313

(0.084) (0.084) (0.251)

mydata$home\_2 0.179 0.180 0.166

(0.138) (0.138) (0.199)

mydata$home\_3 0.070 0.047 -0.040

(0.357) (0.358) (0.406)

mydata$home\_4 0.445 \*\*\* 0.446 \*\*\* 0.209

(0.134) (0.134) (0.195)

mydata$int1 -0.447

(0.268)

mydata$int2 -0.419

(0.360)

mydata$int3 -0.124

(0.893)

─────────────────────────────────────────────────────────────────

N 1648 1648 1648 1648

R2 0.000 0.007 0.008 0.009

logLik -3212.380 -3206.849 -3206.416 -3204.966

AIC 6430.760 6423.698 6424.831 6427.932

─────────────────────────────────────────────────────────────────────────────────

**結論:(1)從模型一中我們可以得知，以成年女性作為參考組，男性的社會地位平均分數比女性低了大約0.074分，女性的社會地位平均分數為4.69分。**

1. **從模型二中我們可以得知，以閩南人作為參考組，客家人的社會地位平均分數比閩南人高了大約0.18分，而原住民的社會地位平均分數比閩南人高了大約0.07分，最後，大陸其他省人的社會地位平均分數比閩南人高了大約0.45分。對照組的平均分數為4.58分。**

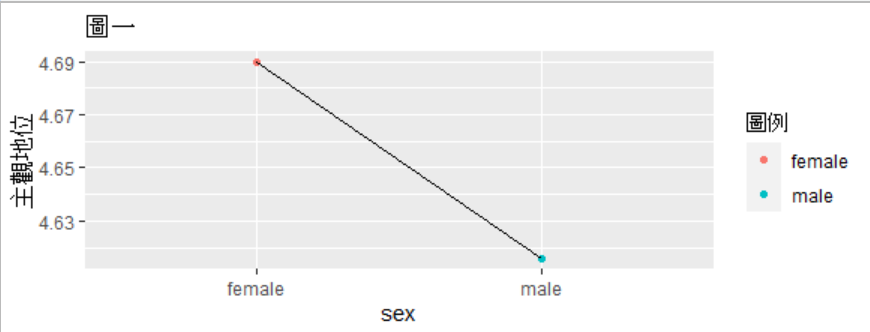
**(3)從模型三中我們可以得知，以女性閩南人作為參考組，且籍貫和性別沒有互動效果時，女性客家人的社會地位平均分數比對照組高了大約0.18分，女性原住民的社會地位平均分數比對照組高了大約0.047分，大陸其他女性省人的社會地位平均分數比對照組高了大約0.45分。**

**而若為男性，男性客家人的社會地位平均分數比對照組高了大約0.102分，女性原住民的社會地位平均分數比對照組低了大約0.031分，大陸其他女性省人的社會地位平均分數比對照組高了大約0.372分。對照組的平均分數為4.623分。**

**(4)最後，從模型四我們可以發現，以女性閩南人作為參考組，且籍貫和性別「有」互動效果時，(當控制其他變項)女性客家人的社會地位平均分數比對照組高了大約0.17分，女性原住民的社會地位平均分數比對照組低了大約0.04分，大陸其他女性省人的社會地位平均分數比對照組高了大約0.21分。只要多變項分析都要寫這句**

**而若為男性，男性客家人的社會地位平均分數比對照組高了大約0.47分，女性原住民的社會地位平均分數比對照組高了大約0.26分，大陸其他女性省人的社會地位平均分數比對照組高了大約0.51分。對照組的平均為4.65分。**

**\*C的前三題作法是把X軸視為類別變數，概念上只要兩個點就能畫出一條線，所以你可以透過方程式求得你需要的兩個點**

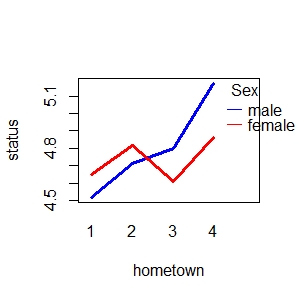


**這個圖用excel就能做出來了，但是我們禮拜四會統一講解怎麼用R做。**

**EX\_C(.2):**

mydata$sex1 <- factor(mydata$sex,levels = c('1','2'),labels = c('male','female'))

interaction.plot(mydata$home,mydata$sex1,mydata$score,type='l',xlab = 'hometown',ylab='status',col = c('blue','red'),lty = 1,lwd=3,trace.label = 'Sex')



108304010@nccu.edu.tw