#William Zhang

#ID:110768508

#Problem1

library(ISwR)

tooth <- data("ToothGrowth")

#Notice that the data is independent since each dose gets tested different

#Extract the tooth length with dosage 0.5

dose1 <- ToothGrowth$len[ToothGrowth$dose == 0.50]

#Extract the tooth length with dosag 1.0

dose2 <-ToothGrowth$len[ToothGrowth$dose == 1.0]

#Extract the tooth length with dosage 2.0

dose3 <- ToothGrowth$len[ToothGrowth$dose == 2.0]

data <- c(dose1,dose2,dose3)

group<-c(rep(0.5,20), rep(1.0,20), rep(2.0,20))

ydata <- data.frame(y= data, group = factor(group))

bartlett.test(data ~ group)

#Well we can certainly use AOV in the case.

#But since the P value is massive here we will run on equal variance assumption

fit1 <- aov(y ~ group,data = ydata)

summary(fit1)

#AOV tells us that there P valueis 9.53....e^-16. Whcih is far smaller than 0.01

#Which is way too small reject null hypothesis and conclude means are different

#Part 2

pairwise.t.test(data, group, p.adj ='bonferroni')

#As one can see the the lvl(1,2) and (1,3) are different. Lvl (2,3) are different as well.

#Lvl 1 has dosage 0.5 and lvl 2 has dosage 1.0 and Lvl 3 has dosage 2.0.

#When compared all the means are different

#Part 3

#Now we will use 2 sample testing because are are only 2 samples. They are indepdent however

doseVC <-ToothGrowth$len[ToothGrowth$supp == "VC"]

#Extract the tooth length with dosage 2.0

doseOJ <- ToothGrowth$len[ToothGrowth$supp == "OJ"]

shapiro.test(doseVC)

shapiro.test(doseOJ)

#They all follow a normal distribution because their values exceed 0.01'

var.test(doseVC,doseOJ)

#0.2331 SHOWS that we will fail to reject null hypothesis conclude they have equal variance

t.test(doseVC,doseOJ,var.equal = T)

#0.06 > 0.01. Fail to reject null hypothesis again and conclude their means are equal

#Part 4

anova2<-aov(ToothGrowth$len~

as.factor(ToothGrowth$dose)\*as.factor(ToothGrowth$supp),data = ToothGrowth)

summary(anova2)

#According to the P value Dosage is signficant. As well as well as the Type of vitamin C supplementation

#Both of them are vital signficant in contributing to toothgrowth.

#There interaction is significant.

#Problem 2

library(ISwR)

data(kfm)

#Sex has a problem its a categorical data with non integer. We need to perform a multiple linear regression

y <- kfm$dl.milk

x2 <- as.numeric(kfm$sex)-1

x3 <- kfm$weight

x4 <- kfm$ml.suppl

x5 <- kfm$mat.height

x6 <- kfm$no

#Now perform a linear model regression

linearmodel <- lm(y ~ x2 + x3 + x4 + x5 + x6,data =kfm)

summary(linearmodel)

#The general idea is that for any P value that is less than 0.05 that the variable is significant.

#It appears that weight,suppl,height are significant variables and the sex and $No is are not significant at all

#It appears that x3,x4,x5 are significant and x2,x6 are useless. However it is to be noted that x4(suppl)

#We can either choose to include or not include supplementation as part. If we really want the most minimum amount then use x3 and x5

library(leaps)

leaps1 <- regsubsets(y ~ x2+ x3 + x4 + x5+ x6, data = kfm, nbest = 10)

summary(leaps1)

plot(leaps1, scale = 'r2')

#As one can see the plot, That you dont see much improvement beyond 2 variables at x3,x5

#Rerun a linear model

linearmodel2 <- lm(y ~ x3 + x5 ,data =kfm)

summary(linearmodel2)

#Better fit is indeed achieved here

Problem 3

**data** problem3;

INPUT ID $ **1**-**3** GENDER $ **5**

@**7** SCORE **3.**

@**11** DATE MMDDYY10.

@**22** AMOUNT COMMA5.

@**28** PROFIT DOLLAR6.;

datalines;

001,M,98 ,09/11/1981,1,000,$5.81

002,F,100,02/12/1986,9,966,$17.81

003,M,89 ,04/11/1976,1,530,$11.42

004,M,77 ,02/14/1985,7,320,$0.26

;

**proc** **print** data = problem3;

Title 'sas data';

WHERE Amount > **1500** AND profit > **4.00**;

FORMAT DATE DATE9.;

FORMAT PROFIT DOLLAR6.;

**run**;

|  |
| --- |
| sas data |

| **Obs** | **ID** | **GENDER** | **SCORE** | **DATE** | **AMOUNT** | **PROFIT** |
| --- | --- | --- | --- | --- | --- | --- |
| **2** | 002 | F | 100 | 12FEB1986 | 9966 | $18 |
| **3** | 003 | M | 89 | 11APR1976 | 1530 | $11 |

Problem 4

**data** mathscores;

input group$ mathscore @@;

datalines;

A 92 A 88 A 76 A 82 A 62 A 74

B 94 B 88 B 73 B 76 B 91 B 78

C 95 C 92 C 84 C 94 C 98 C 82

;

**PROC** **ANOVA** data = mathscores;

Title 'Analysis of math scores';

class group;

model mathscore = group;

means group/SNK;

**RUN**;

|  |
| --- |
| Analysis of math scores |

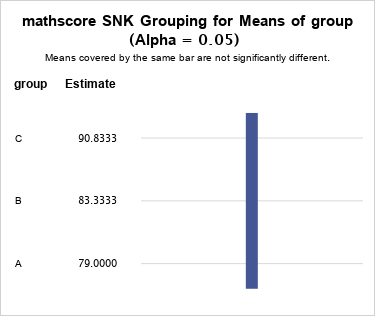
The ANOVA Procedure

Dependent Variable: mathscore

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 2 | 430.111111 | 215.055556 | 2.76 | 0.0956 |
| **Error** | 15 | 1170.166667 | 78.011111 |  |  |
| **Corrected Total** | 17 | 1600.277778 |  |  |  |

The P value is 0.0956 and the mean scores of the 3 groups are not significantly different from each other.

Part B)



As one can see there all the groups are covered by the same bar which means they are not significantly different from each group

Part C)

**data** mathscores;

input group$ mathscore @@;

datalines;

A 92 A 88 A 76 A 82 A 62 A 74

B 94 B 88 B 73 B 76 B 91 B 78

C 95 C 92 C 84 C 94 C 98 C 82

;

**PROC** **GLM** DATA = mathscores;

class group;

Model mathscore = group;

Contrast 'B vs A and C' group **1** -**2** **1**;

**run**;

| **Contrast** | **DF** | **Contrast SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **B vs A and C** | 1 | 10.02777778 | 10.02777778 | 0.13 | 0.7249 |

As you can see here the P value is massive. Fail to reject null hypothesis and conclude B is not significantly different than A and C group.

Part 2)

Part A)

**data** mathscores;

input group$ age mathscore @@;

datalines;

A 2 92 A 2 88 A 1 76 A 2 82 A 1 62 A 1 74

B 2 94 B 1 88 B 1 73 B 2 76 B 2 91 B 1 78

C 2 95 C 1 92 C 1 84 C 2 94 C 2 98 C 1 82

;

**PROC** **anova** data = mathscores;

Title 'Mathscores';

Class group age;

Model mathscore = group | age;

**run**;

| **Source** | **DF** | **Anova SS** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **group** | 2 | 430.1111111 | 215.0555556 | 4.84 | 0.0287 |
| **age** | 1 | 566.7222222 | 566.7222222 | 12.77 | 0.0038 |
| **group\*age** | 2 | 70.7777778 | 35.3888889 | 0.80 | 0.4731 |

The P values Group is higher than 0.01. Fail to reject null hypothesis. Which means group is not significant to the contribution of mean math scores. However Age P value is 0.0038. Age is significant to math scores because we will reject null hypothesis. However the interaction P value is 0.4731. Which means their interaction between group and age is insignificant.

Part B)

**data** mathscores;

input group$ age mathscore @@;

datalines;

A 2 92 A 2 88 A 1 76 A 2 82 A 1 62 A 1 74

B 2 94 B 1 88 B 1 73 B 2 76 B 2 91 B 1 78

C 2 95 C 1 92 C 1 84 C 2 94 C 2 98 C 1 82

;

**PROC** **MEANS** data = mathscores nway noprint;

CLASS group age;

VAR mathscore;

output OUT = MAZE1 MEAN = M;

**RUN**;

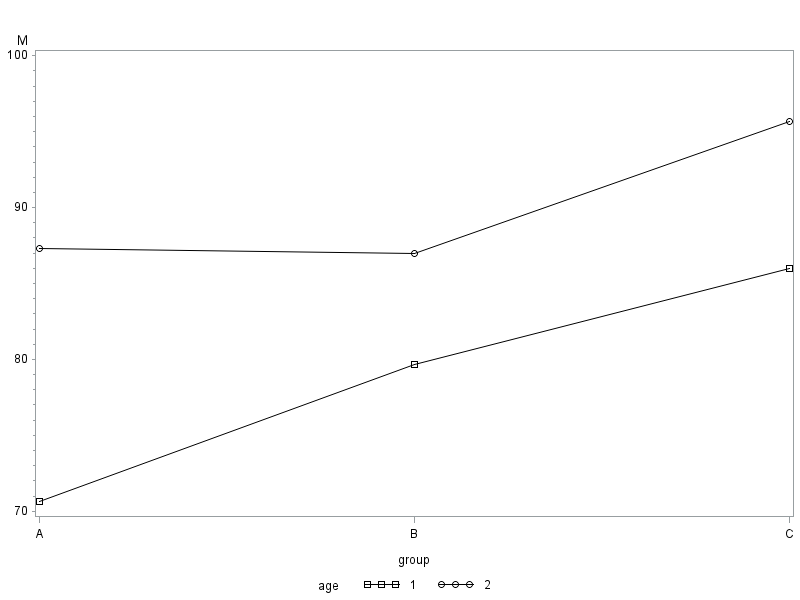
SYMBOL1 VALUE = square COLOR = BLACK I = JOIN;

SYMBOL2 VALUE = CIRCLE COLOR = BLACK I = JOIN;

**PROC** **GPLOT** data = MAZE1;

PLOT M\*group = age;

**RUN**;



Part C)

**data** mathscores;

input cond $ mathscore @@;

datalines;

A\_12-14 76 A\_12-14 62 A\_12-14 74

A\_15-18 92 A\_15-18 88 A\_15-18 82

B\_12-14 88 B\_12-14 73 B\_12-14 78

B\_15-18 94 B\_15-18 76 B\_15-18 91

C\_12-14 92 C\_12-14 84 C\_12-14 82

C\_15-18 95 C\_15-18 94 C\_15-18 98

;

**proc** **anova** data = mathscores;

Title 'Anaysis of 6 levels';

CLASS cond;

model mathscore = cond;

means cond/snk alpha = 0.01;

**run**;

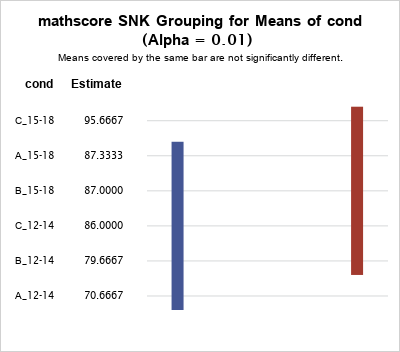
|  |
| --- |
| Anaysis of 6 levels |

The ANOVA Procedure

Dependent Variable: mathscore

| **Source** | **DF** | **Sum of Squares** | **Mean Square** | **F Value** | **Pr > F** |
| --- | --- | --- | --- | --- | --- |
| **Model** | 5 | 1067.611111 | 213.522222 | 4.81 | 0.0121 |
| **Error** | 12 | 532.666667 | 44.388889 |  |  |
| **Corrected Total** | 17 | 1600.277778 |  |  |  |

0.0121 is barely greater than 0.01. Accept that even with classifying with the 6 classfications. The mean mathscores are not significantly different from each other. Had we select a better significance such as 0.05 then we would’ve gotten there a conclusion that the groups means are significantly different from each other.



C-15-18,A-15-18,B-15-18,C-12-14,B-12-14 are not significantly different but it is different in from A\_12-14. A\_12-14 lowers means is lower than all of them.

A-15-18,B-15-18,C-12-14,B-12-14, A\_12-14 are not different from each other but it is different in from from C\_15-18. C\_15-18 means is higher than all of them.

Question 5)

**data** string;

input line $;

State = substr(line,**1**,**2**);

N = substr(line,**3**,LENGTH(line)-**3**);

group = substr(line,LENGTH(line),**1**);

if group = 'B' then NEWN = N \* **10**;

if group = 'A' then NEWN = N \* **100**;

datalines;

NY113A

NY15B

NJ10A

NY34A

NJ25B

NY87A

NJ23B

NJ9A

;

**proc** **print** data = string;

**run**;

|  |
| --- |
| Anaysis of 6 levels |

| **Obs** | **Line** | **State** | **N** | **group** | **NEWN** |
| --- | --- | --- | --- | --- | --- |
| **1** | NY113A | NY | 113 | A | 11300 |
| **2** | NY15B | NY | 15 | B | 150 |
| **3** | NJ10A | NJ | 10 | A | 1000 |
| **4** | NY34A | NY | 34 | A | 3400 |
| **5** | NJ25B | NJ | 25 | B | 250 |
| **6** | NY87A | NY | 87 | A | 8700 |
| **7** | NJ23B | NJ | 23 | B | 230 |
| **8** | NJ9A | NJ | 9 | A | 900 |