midterm

Jason Grahn

2/9/2019

# 4

In the following table, we have a comparison of four reagents. The first reagent is the one presently in use and the other three are less expensive reagents that we wish to compare with the first. All four reagents are used with a blood sample from each patient.

The three variables measured for each reagent are 𝑦1=white blood count, 𝑦2=red blood count, and 𝑦3=hemoglobin count. Compare the four reagents using all four MANOVA tests. State each hypotheses clearly, and interpret the results.

reagent <- readr::read\_table2(file = here::here("/midterm/T6\_19\_REAGENT.DAT"),  
 col\_names = FALSE) %>%   
 rename(reagent = X1,  
 subject = X2,  
 y1 = X3,  
 y2 = X4,  
 y3 = X5) %>%   
 mutate(reagent = factor(reagent),  
 subject = factor(subject))  
  
head(reagent,5)

## # A tibble: 5 x 5  
## reagent subject y1 y2 y3  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 1 1 8 3.96 12.5  
## 2 1 2 4 5.37 16.9  
## 3 1 3 6.3 5.47 17.1  
## 4 1 4 9.4 5.16 16.2  
## 5 1 5 8.2 5.16 17

manova.reagent <- manova(cbind(y1, y2, y3) ~ reagent,  
 data=reagent)  
  
manova.reagent

## Call:  
## manova(cbind(y1, y2, y3) ~ reagent, data = reagent)  
##   
## Terms:  
## reagent Residuals  
## resp 1 0.955 225.163  
## resp 2 0.06295 14.92915  
## resp 3 2.53938 171.67050  
## Deg. of Freedom 3 76  
##   
## Residual standard errors: 1.721241 0.4432111 1.502938  
## Estimated effects may be unbalanced

## The Hypothesis

: The mean vectors between reagents are equal ( ).

: The mean vectors between reagents are not equal ( ).

## The tests

reagent.lambda <- broom::tidy(manova.reagent, test = "Wilks")  
reagent.lambda

## # A tibble: 2 x 7  
## term df wilks statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 reagent 3 0.873 1.15 9 180. 0.328  
## 2 Residuals 76 NA NA NA NA NA

### Wilks interpretation

p-Value of reagent.lambda is 0.3280951. This value is above 0.05; so we do not reject and conclude that the means are equal.

reagent.pillai <- broom::tidy(manova.reagent, test = "Pillai")  
reagent.pillai

## # A tibble: 2 x 7  
## term df pillai statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 reagent 3 0.128 1.13 9 228 0.345  
## 2 Residuals 76 NA NA NA NA NA

### Pillai interpretation

p-Value of reagent.pillai is 0.3453898. This value is above 0.05; so we do not reject and conclude that the means are equal.

reagent.hotel <- broom::tidy(manova.reagent, test = "Hotelling-Lawley")  
reagent.hotel

## # A tibble: 2 x 7  
## term df hl statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 reagent 3 0.146 1.18 9 218 0.312  
## 2 Residuals 76 NA NA NA NA NA

### Hotelling interpretation

p-Value of reagent.hotel is 0.3117863. This value is above 0.05; so we do not reject and conclude that the means are equal.

reagent.roy <- broom::tidy(manova.reagent, test = "Roy",  
 intercept = FALSE)  
  
reagent.roy

## # A tibble: 2 x 7  
## term df roy statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 reagent 3 0.143 3.63 3 76 0.0166  
## 2 Residuals 76 NA NA NA NA NA

roy.theta <- round(reagent.roy$roy[1]/(reagent.roy$roy[1] + 1), 3)  
roy.theta

## [1] 0.125

### Roy interpretation

p-Value of reagent.roy is 0.0165718. While this value is below 0.05; Roy’s Theta is 0.125 so we do not reject and conclude that the means are equal.

# 5

The table below displays scores on math, English, and art tests for 5 students. Note that data from the table is represented in matrix A, where each column in the matrix shows scores on a test and each row shows scores for a student:

#input data  
student <- as.factor(c(1, 2, 3, 4, 5))  
math <- c(90,90,60,60,30)   
english <- c(60,90,60,60,30)  
art <- c(90,30,60,90,30)  
  
#make a table  
students <- tibble(math, english, art)  
students

## # A tibble: 5 x 3  
## math english art  
## <dbl> <dbl> <dbl>  
## 1 90 60 90  
## 2 90 90 30  
## 3 60 60 60  
## 4 60 60 90  
## 5 30 30 30

## (a) Calculate the sample covariance matrix S

S <- round(var(students), 2)  
S

## math english art  
## math 630 450 225  
## english 450 450 0  
## art 225 0 900

## (b) Calculate the sample correlation matrix R.

R <- round(cor(S), 2)  
R

## math english art  
## math 1.00 0.90 -0.76  
## english 0.90 1.00 -0.97  
## art -0.76 -0.97 1.00

## (c) Now let’s define 𝑧 = −2𝑦1 + 3𝑦2 + 𝑦3,where 𝑦1denotes Math scores, 𝑦2 denotes English scores, and 𝑦3 denotes Art scores. Find the sample mean vector 𝑧 and the sample variance 𝑠2.

#build the constant matrix  
a <- matrix(c(-2, 3, 1))  
  
#gather the column means  
y <- colMeans(students)  
  
#sample mean vector  
z <- y %\*% t(a)  
z

## [,1] [,2] [,3]  
## [1,] -132 198 66  
## [2,] -120 180 60  
## [3,] -120 180 60

#sample variance  
s2 <- t(a) %\*% as.matrix(S) %\*% a  
s2

## [,1]  
## [1,] 1170

# 6

Use the beetle data, do the following:

beetles <- readr::read\_table2(file = here::here("/midterm/T5\_5\_FBEETLES.DAT"),  
 col\_names = FALSE) %>%   
 rename(exp.no = X1,  
 species = X2,  
 y1 = X3,  
 y2 = X4,  
 y3 = X5,  
 y4 = X6) %>%   
 dplyr::select(y1, y2, y3, y4, species) %>%   
 mutate(species = factor(species))

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## X2 = col\_double(),  
## X3 = col\_double(),  
## X4 = col\_double(),  
## X5 = col\_double(),  
## X6 = col\_double()  
## )

head(beetles,5)

## # A tibble: 5 x 5  
## y1 y2 y3 y4 species  
## <dbl> <dbl> <dbl> <dbl> <fct>   
## 1 189 245 137 163 1   
## 2 192 260 132 217 1   
## 3 217 276 141 192 1   
## 4 221 299 142 213 1   
## 5 171 239 128 158 1

## (a) Find the classification function and cutoff point.

# group by species  
beet1 <- subset(x = beetles, subset = (species == 1), y1:y4)  
beet2 <- subset(x = beetles, subset = (species == 2), y1:y4)  
# find group means for all variables for each species  
mean1 <- apply(X = beet1, FUN = mean, MARGIN = 2)  
#mean1  
mean2 <- apply(X = beet2, FUN = mean, MARGIN = 2)  
#mean2  
# find covariance matrix for each species  
var1 <- var(beet1)  
#var1  
var2 <- var(beet2)  
#var2  
  
beetles.manova <- stats::manova(cbind(y1, y2, y3, y4) ~ species,   
 data = beetles)  
  
# find \_pooled\_ sample variance for each species  
## length of each set  
length1 <- nrow(beet1)  
length2 <- nrow(beet2)  
## variance of each set  
pvar1 <- (length1 - 1) \* var1  
pvar2 <- (length2 - 1) \* var2   
## use length and variance to get pooled variance  
sp1 <- 1 / (length1 + length2 - 2) \* (pvar1 + pvar2)  
#sp1  
  
cutoff <- .5\*(mean1 - mean2) %\*% solve(sp1) %\*% (mean1 + mean2)  
cutoff <- round(cutoff,2)  
cutoff

## [,1]  
## [1,] -15.81

species\_prediction <- apply(beetles[,1:4], 1, function(y) {  
 z <- (mean1 - mean2) %\*% solve(sp1) %\*% y })  
species\_prediction

## [1] -4.640975 -12.769954 -3.599007 -8.333783 -8.398459 -8.319691  
## [7] -5.998789 -7.104974 -14.269659 -8.795039 -5.377660 -6.685556  
## [13] -7.531402 -11.533464 -10.922504 -11.168223 -7.646623 -12.531815  
## [19] -14.524616 -26.822874 -17.458688 -26.373153 -24.334824 -25.694962  
## [25] -26.666573 -27.486750 -15.047912 -21.363037 -23.600898 -28.771377  
## [31] -22.022042 -20.074530 -26.994104 -21.101538 -23.623558 -16.050876  
## [37] -23.304407 -17.341090 -18.974629

## (b) Find the classification table using the nearest neighbor method by setting k = 3.

## Normalization  
normalize <- function(x) {  
 return((x - min(x)) / (max(x) - min(x))) }  
  
normalized.beetles <- as.data.frame(lapply(beetles[1:4], normalize))  
  
## Splitting Data Set into Training and Test Sets  
## r subset datasets  
set.seed(1234)  
ind <- sample(2,   
 nrow(beetles),   
 replace=TRUE,   
 prob=c(0.67, 0.33))  
  
# make training set and labels  
beetles.training <- beetles[ind==1, 1:4]  
beetles.trainLabels <- as.matrix(beetles[ind==1, 5])  
  
# make testing set and labels  
beetles.test <- beetles[ind==2, 1:4]  
beetles.testLabels <- as.matrix(beetles[ind==2, 5])

### K-Nearest Neighbor in R

beetles.knn <- knn(train = beetles.training,  
 cl = beetles.trainLabels,  
 k = 3,  
 test = beetles.test)  
  
table(beetles.knn)

## beetles.knn  
## 1 2   
## 4 5

## (c) Calculate misclassification rate.

table(beetles.testLabels, beetles.knn)

## beetles.knn  
## beetles.testLabels 1 2  
## 1 4 0  
## 2 0 5

accuracy.table <- beetles %>%   
 mutate(ind = ind) %>%   
 filter(ind == 2) %>%   
 mutate(beetles.knn = beetles.knn,  
 predict.true = if\_else(species == beetles.knn, TRUE, FALSE))   
  
accuracy.rate <- accuracy.table %>%   
 group\_by(predict.true) %>%   
 summarize(count = n(),  
 rate = count / nrow(accuracy.table) \* 100,  
 rate = round(rate,2))  
  
accuracy.rate

## # A tibble: 1 x 3  
## predict.true count rate  
## <lgl> <int> <dbl>  
## 1 TRUE 9 100

# 7

Use the above beetle data, do the following: ## (a) Use LDA by setting probability of 50% and 50% to train model.

model\_1 <- lda(formula = species ~ .,   
 data = beetles,   
 prior = c(1,1)/2)  
model\_1

## Call:  
## lda(species ~ ., data = beetles, prior = c(1, 1)/2)  
##   
## Prior probabilities of groups:  
## 1 2   
## 0.5 0.5   
##   
## Group means:  
## y1 y2 y3 y4  
## 1 194.4737 267.0526 137.3684 185.9474  
## 2 179.5500 290.8000 157.2000 209.2500  
##   
## Coefficients of linear discriminants:  
## LD1  
## y1 -0.09327642  
## y2 0.03522706  
## y3 0.02875538  
## y4 0.03872998

## (b) Predict new observation (189,245,138,164).

y1 <- 189   
y2 <- 245   
y3 <- 138  
y4 <- 164  
predict\_me <- tibble(y1, y2, y3, y4)  
  
predict\_subgroup <- predict(model\_1, # predictions  
 newdata = predict\_me)  
  
predict\_subgroup$class

## [1] 1  
## Levels: 1 2

The LDA method predicts the new observation will be beetle species 1, which is *Haltica oleracea*.

## (c) Calculate misclassification rate.

predict\_beetles <- predict(model\_1, # predictions  
 data = beetles)  
  
predict\_class <- predict\_beetles$class  
  
beetle\_predictions <- beetles %>%   
 mutate(predictions = predict\_class,  
 correct\_species = factor((species == predictions))) %>%   
 group\_by(correct\_species) %>%   
 summarise(classified\_count = n(),   
 total = nrow(beetles)) %>%   
 mutate(percentage = round(classified\_count / total \* 100, 2))   
  
beetle\_predictions

## # A tibble: 2 x 4  
## correct\_species classified\_count total percentage  
## <fct> <int> <int> <dbl>  
## 1 FALSE 1 39 2.56  
## 2 TRUE 38 39 97.4

It’s not clear if this question is asking about LDA for the overall beetle prediction or just the new observation. Assuming this question is asking about the overall beetle population, the misclassification rate is 2.56 %.

# 8

The following table contains data from O’Sullivan and Mahan with measurements of blood glucose levels on three occasions for 30 women. The 𝑦’s represent fasting glucose measurements on the three occasions; the 𝑥’s are glucose measurements 1 hour after sugar intake.

glucose <- read\_csv(here::here("midterm/glucose.csv"))

## Parsed with column specification:  
## cols(  
## y1 = col\_double(),  
## y2 = col\_double(),  
## y3 = col\_double(),  
## x1 = col\_double(),  
## x2 = col\_double(),  
## x3 = col\_double()  
## )

head(glucose, 5)

## # A tibble: 5 x 6  
## y1 y2 y3 x1 x2 x3  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 62 75 68 116 130 91  
## 2 74 64 70 109 101 103  
## 3 64 71 66 77 102 130  
## 4 73 70 64 115 110 109  
## 5 68 67 75 76 85 119

## Find the mean vector

glucose\_means <- matrix(colMeans(glucose))  
glucose\_means

## [,1]  
## [1,] 72.20000  
## [2,] 72.73333  
## [3,] 73.30000  
## [4,] 108.46667  
## [5,] 102.46667  
## [6,] 108.46667

## … and covariance matrix for all six variables

S\_glucose <- round(var(glucose), 2)  
S\_glucose

## y1 y2 y3 x1 x2 x3  
## y1 77.61 0.99 23.73 100.08 4.87 34.32  
## y2 0.99 36.20 15.22 -46.46 30.37 -32.08  
## y3 23.73 15.22 57.46 13.41 -6.42 1.48  
## x1 100.08 -46.46 13.41 959.50 299.36 232.64  
## x2 4.87 30.37 -6.42 299.36 500.19 61.81  
## x3 34.32 -32.08 1.48 232.64 61.81 527.02

# 9

Various aspects of economic cycles were measured for consumer goods and producer goods by Tintner. The variables are:

𝑦1=length of cycle  
𝑦2=percentage of rising prices  
𝑦3=cyclical amplitude  
𝑦4=rate of change

The data for several items are given in the following table:

goods <- readr::read\_table2(file = here::here("/midterm/T5\_8\_GOODS.DAT"),  
 col\_names = FALSE) %>%   
 rename(item = X1,  
 goods\_type = X2,  
 y1 = X3,  
 y2 = X4,  
 y3 = X5,  
 y4 = X6) %>%   
 mutate(item = factor(item),  
 goods\_type = factor(goods\_type))

## Parsed with column specification:  
## cols(  
## X1 = col\_double(),  
## X2 = col\_double(),  
## X3 = col\_double(),  
## X4 = col\_double(),  
## X5 = col\_double(),  
## X6 = col\_double()  
## )

head(goods,5)

## # A tibble: 5 x 6  
## item goods\_type y1 y2 y3 y4  
## <fct> <fct> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1 72 50 8 0.5  
## 2 2 1 66.5 48 15 1   
## 3 3 1 54 57 14 1   
## 4 4 1 67 60 15 0.9  
## 5 5 1 44 57 14 0.3

## Use Hotelling’s test to test for a difference in the mean measurements vector of the Consumers Goods and the mean vector of the Producer Goods. State each hypotheses clearly, and interpret the results.

### The Hypothesis

: The mean vectors between good types are equal ( ).

: The mean vectors between good types are not equal ( ).

manova.goods <- manova(cbind(y1, y2, y3, y4) ~ goods\_type,  
 data=goods)  
  
manova.goods

## Call:  
## manova(cbind(y1, y2, y3, y4) ~ goods\_type, data = goods)  
##   
## Terms:  
## goods\_type Residuals  
## resp 1 8232.458 10150.989  
## resp 2 22.237 1252.500  
## resp 3 190.667 304.622  
## resp 4 0.103 3.357  
## Deg. of Freedom 1 17  
##   
## Residual standard errors: 24.43598 8.5835 4.233079 0.4443472  
## Estimated effects may be unbalanced

### The Test

goods.hotel <- broom::tidy(manova.goods, test = "Hotelling-Lawley")  
goods.hotel

## # A tibble: 2 x 7  
## term df hl statistic num.df den.df p.value  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 goods\_type 1 1.09 3.80 4 14 0.0270  
## 2 Residuals 17 NA NA NA NA NA

### The Interpretation

p-Value of reagent.hotel is 0.027. This value is above 0.05; so we reject and conclude that the means vectors are *not* equal between consumer and producer goods.