DA410_Assignment4_MattGraham

This is our first project, analyzing air pollution, mortality rates, and relevant parameters.

```
library(nnspat)  # used for dist2full()
library("dplyr")  # used to select numeric datatypes
library("ggplot2")
library(reshape)  # used for melting matricies
library(klaR)
```

8.7

For the psychological data in Table 5.1, the discriminant function coefficient vector was given in Example 5.5.

```
psych <- read.table("C:/mattgraham93.github.io/school/22_3_DA410/data/T5_1_PSYCH.DAT", header
=FALSE)
colnames(psych) <- c('gender', 'y1', 'y2', 'y3', 'y4')

# create subsets and remove gender column
male <- psych[psych$gender==1,-1]
female <- psych[psych$gender==2,-1]

as.data.frame(psych)
```

gender <int>	y1 <int>	y2 <int>	y3 <int>	y4 <int>
1	15	17	24	14
1	17	15	32	26
1	15	14	29	23
1	13	12	10	16
1	20	17	26	28
1	15	21	26	21
1	15	13	26	22
1	13	5	22	22
1	14	7	30	17
1	17	15	30	27

1-10 of 64 rows Previous **1** 2 3 4 5 6 7 Next

(a) Find the standardized coefficients.

Getting covariance matricies for subsets

```
cov.male <- cov(male)
cov.female <- cov(female)
```

Male covariance matrix

```
as.data.frame(cov.male)
```

	y1 <dbl>	y2 <dbl>	y3 <dbl>	y4 <dbl>
y1	5.192540	4.545363	6.522177	5.250000
y2	4.545363	13.184476	6.760081	6.266129
y3	6.522177	6.760081	28.673387	14.467742
y4	5.250000	6.266129	14.467742	16.645161

4 rows

Female covariance matrix

```
as.data.frame(cov.female)
```

	y1 <dbl>	y2 <dbl>	y3 <dbl>	y4 <dbl>
y1	9.136089	7.549395	4.863911	4.151210
y2	7.549395	18.603831	10.224798	5.445565
y3	4.863911	10.224798	30.039315	13.493952
y4	4.151210	5.445565	13.493952	27.995968

4 rows

Calculating pooled variance:

```
male.n <- nrow(male)
male.df <- male.n - 1

female.n <- nrow(female)
female.df <- female.n - 1

psych.n <- nrow(psych)

psych.pooled.var <- ( (male.df*cov.male) %*% (female.df*cov.female)) / psych.n
as.data.frame(psych.pooled.var)
```

	y1 <dbl>	y2 <dbl>	y3 <dbl>	y4 <dbl>

	y1	y2	y3	y4
	<dbl>	<dbl>	<dbl>	<dbl>
y1	2031.185	3289.005	5082.736	4223.838
y2	3002.436	5748.575	6675.044	5365.266
y3	4657.023	8213.044	17379.079	12851.019
y4	3524.728	5927.874	11243.933	10768.318

4 rows

```
a <- matrix(c(.5104, -.2033, .4660, -.3097))

a.star <- sqrt(diag(psych.pooled.var)) %*% a
a.star
```

```
##         [,1]
## [1,] 36.8839
```

The result is not as expected/given in the book. I expected to return a matrix and was given one single value.

b. Calculate í-tests for the individual variables.

```
lapply(psych[-1], function(x) t.test(x~psych$gender))
```

```
## $y1
##
##  Welch Two Sample t-test
##
## data:  x by psych$gender
## t = 5.4173, df = 57.634, p-value = 1.234e-06
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal
to 0
## 95 percent confidence interval:
##  2.285358 4.964642
## sample estimates:
## mean in group 1 mean in group 2
##        15.96875        12.34375
##
##
## $y2
##
##  Welch Two Sample t-test
##
## data:  x by psych$gender
## t = 2.0066, df = 60.249, p-value = 0.04928
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal
to 0
## 95 percent confidence interval:
##  0.006498957 3.993501043
## sample estimates:
## mean in group 1 mean in group 2
##        15.90625        13.90625
##
##
## $y3
##
##  Welch Two Sample t-test
##
## data:  x by psych$gender
## t = 7.7748, df = 61.966, p-value = 9.765e-11
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal
to 0
## 95 percent confidence interval:
##   7.823539 13.238961
## sample estimates:
## mean in group 1 mean in group 2
##        27.18750        16.65625
##
##
## $y4
##
##  Welch Two Sample t-test
##
## data:  x by psych$gender
## t = 0.68791, df = 58.235, p-value = 0.4942
## alternative hypothesis: true difference in means between group 1 and group 2 is not equal
to 0
## 95 percent confidence interval:
##  -1.551558  3.176558
## sample estimates:
## mean in group 1 mean in group 2
##         22.7500         21.9375
```

c. Compare the results of (a) and (b) as to the contribution of the variables to separation of the two groups.

Given part a did not calculate as expected, I can't determine the differences between the two. I can conclude the rank for part b in important are: y3, y1, y2, then y4.

d. Find the partial F for each variable, as in (8.26), and compare with the standardized coefficients.

```
full.psych <- lm(gender ~y1+y2+y3+y4, data=psych)

reduced.psych.y1 <- lm(gender ~y2+y3+y4, data=psych)
reduced.psych.y2 <- lm(gender ~y1+y3+y4, data=psych)
reduced.psych.y3 <- lm(gender ~y1+y2+y4, data=psych)
reduced.psych.y4 <- lm(gender ~y1+y2+y3, data=psych)
```

y1 ANOVA table

```
anova(reduced.psych.y1, full.psych)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	60	7.041790	NA	NA	NA	NA
2	59	6.215481	1	0.8263095	7.843683	0.006884539

2 rows

y2 ANOVA table

```
anova(reduced.psych.y2, full.psych)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	60	6.490669	NA	NA	NA	NA
2	59	6.215481	1	0.2751881	2.612203	0.1113795

2 rows

y3 ANOVA table

```
anova(reduced.psych.y3, full.psych)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	60	10.483454	NA	NA	NA	NA
2	59	6.215481	1	4.267974	40.51343	3.181336e-08

2 rows

y4 ANOVA table

```
anova(reduced.psych.y4, full.psych)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	60	7.262472	NA	NA	NA	NA
2	59	6.215481	1	1.046991	9.93849	0.002544301

2 rows

Observing our linear models, we can see that our ranks across all 4 reduced models = 3. We can also conclude that y2 is not a significant variable when looking to seek differences between genders. While it may have fallen within our initial scope of “reasonable”, it’s lack of variability does not make it a good fit for our model.

8.11 (a and b)

Using the fish data in Table 6.17, do the following:

```
fish <- read.table("C:/mattgraham93.github.io/school/22_3_DA410/data/T6_17_FISH.DAT", header=
FALSE)
colnames(fish) <- c('method', 'y1', 'y2', 'y3', 'y4')

method1 <- fish[fish$method==1, -1]
method2 <- fish[fish$method==2, -1]
method3 <- fish[fish$method==3, -1]

method1.bar <- colMeans(method1)
method2.bar <- colMeans(method2)
method3.bar <- colMeans(method3)
y.bar.all <- colMeans(fish[-1])

as.data.frame(fish)
```

method	y1	y2	y3	y4
<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	5.4	6.0	6.3	6.7
1	5.2	6.5	6.0	5.8

method	y1	y2	y3	y4
<int>	<dbl>	<dbl>	<dbl>	<dbl>
1	6.1	5.9	6.0	7.0
1	4.8	5.0	4.9	5.0
1	5.0	5.7	5.0	6.5
1	5.7	6.1	6.0	6.6
1	6.0	6.0	5.8	6.0
1	4.0	5.0	4.0	5.0
1	5.7	5.4	4.9	5.0
1	5.6	5.2	5.4	5.8

1-10 of 36 rows Previous **1** 2 3 4 Next

Calculate E and H

```
method1.bar.diff <- method1.bar - y.bar.all
method2.bar.diff <- method2.bar - y.bar.all
method3.bar.diff <- method3.bar - y.bar.all

H <- 12 * unname(method1.bar.diff %*% t(method1.bar.diff)
                 + method2.bar.diff %*% t(method2.bar.diff)
                 + method3.bar.diff %*% t(method3.bar.diff)
                 )

"compute.within.matrix" <- function(data, mean) {
  ret <- matrix(as.numeric(0), nrow=4, ncol=4)

  for (i in 1:12) {
      diff <- as.numeric(data[i,] - mean)
      ret <- ret + diff %*% t(diff)
  }
  return(ret)
}

E <- compute.within.matrix(method1, method1.bar) + compute.within.matrix(method2, method2.ba
r) +
compute.within.matrix(method3, method3.bar)

E.H <- solve(E) %*% H
```

Matrix E

```
as.data.frame(E)
```

V1	V2	V3	V4
<dbl>	<dbl>	<dbl>	<dbl>
13.408333	7.723333	8.675000	5.864167
7.723333	8.480000	7.526667	6.213333
8.675000	7.526667	11.607500	7.037500
5.864167	6.213333	7.037500	10.565833

4 rows

Matrix H

```
as.data.frame(H)
```

V1	V2	V3	V4
<dbl>	<dbl>	<dbl>	<dbl>
1.0505556	2.173333	-1.375556	-0.7602778
2.1733333	4.880000	-2.373333	-1.2566667
-1.3755556	-2.373333	2.382222	1.3844444
-0.7602778	-1.256667	1.384444	0.8105556

4 rows

Matrix E*H

```
as.data.frame(E.H)
```

V1	V2	V3	V4
<dbl>	<dbl>	<dbl>	<dbl>
0.03627194	0.004839798	-0.1338574	-0.08405638
0.93949060	2.008612237	-1.1501075	-0.62633818
-0.61058920	-1.208796623	0.8663577	0.48664081
-0.23787326	-0.497671979	0.3046060	0.16755795

4 rows

a. Find the eigenvectors of E^-1 * H .

```
n <- 12  # total records
k <- 3  # methods of cooking
p <- 4  # dependent variables (judges)

eigen(E.H)
```

```
## eigen() decomposition
## $values
## [1]  2.951475e+00  1.273244e-01 -1.853023e-16  6.170595e-17
##
## $vectors
##             [,1]        [,2]       [,3]       [,4]
## [1,] -0.03181703 -0.63526646  0.9232879  0.3615806
## [2,] -0.81967777  0.59729861 -0.3047669 -0.1552483
## [3,]  0.53294806  0.48673081  0.1094385 -0.4208771
## [4,]  0.20756299 -0.05257385  0.2065908  0.8173249
```

b. Carry out tests of significance for the discriminant functions and find the relative importance of each as in (8.13). Do these two procedures agree as to the number of important discriminant functions?

```
vals <- eigen(E.H)[1]
eigen_mean <- sapply(vals, mean)
sprintf("Eigenvalue mean: %s", eigen_mean)
```

```
## [1] "Eigenvalue mean: 0.769699950326934"
```

```
sapply(vals, FUN = '/', FUN.VALUE = eigen_mean)
```

```
##             values
## [1,]  3.834579e+00
## [2,]  1.654208e-01
## [3,] -2.407462e-16
## [4,]  8.016884e-17
```

Looking at our eigenvalues and eigenvectors, we can conclude our judge, V1, does not obtain agreement between our two procedures.

8.15

Carry out a stepwise selection of variables on the fish data of Table 6.17.

```
fish.model <- greedy.wilks(fish[-1],fish$method, "lda", niveau = .1)
fish.model
```

```
## Formula containing included variables:
##
## fish$method ~ y2 + y3
## <environment: 0x00000207cc00a558>
##
##
## Values calculated in each step of the selection procedure:
##
##   vars Wilks.lambda F.statistics.overall p.value.overall F.statistics.diff
## 1   y2    0.6347305             9.495283    5.529998e-04          9.495283
## 2   y3    0.2605673            15.344403    7.525472e-09         22.975295
##   p.value.diff
## 1 5.529998e-04
## 2 5.611982e-07
```

When looking at our stepwise selection, we are returned with 2 significant variables. Both y2 and y3 are significant and should be included in our linear models.