Lab Assignment: Intro to Factor and Cluster Analysis

Christian Acosta

08/15/2016

Create a Word document from this R Markdown file for the following exercises. Submit the R markdown file and resulting Word document via D2L Dropbox.

## Exercise 1

A researcher was interested in learning what motivates international fans when they watch U.S. sports. A questionnaire was created in which respondents reported their score for 42 "importance factors" about fan motivation by circling the number that indicates why they watch, read, and/or discuss U.S. sports (5=Very High, 4=High, 3=Average, 2=Low, 1=Very Low).

The fans were categorized on issues such as gender, interest in U.S. sports, and the media source from which their information comes. Four hundred surveys were completed for the study.

The data is in the file ifanmot.rda and the survey is in the file IFM\_Survey.docx.

### Part 1a

Conduct Bartlett's test for sphericity on the responses for the 42 survey questions found in columns 1 through 42 of the file ifanmot.rda. State the null and alternative hypothesis and report on the results.

Is factor analysis warranted based on this measure?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1a -|-|-|-|-|-|-|-|-|-|-|-

require(DS705data)

## Loading required package: DS705data

data(ifanmot)  
head(ifanmot)

## Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15 Q16 Q17 Q18 Q19 Q20  
## 1 3 1 4 4 4 4 1 4 3 1 1 4 4 4 4 4 4 4 4 3  
## 2 1 1 1 3 3 4 1 3 3 1 1 5 4 1 3 3 3 3 1 1  
## 3 3 2 4 4 4 4 1 4 4 1 3 4 3 2 3 4 4 1 4 4  
## 4 3 3 4 3 3 4 3 4 3 3 3 3 4 3 3 3 4 2 4 3  
## 5 3 4 5 3 4 4 2 4 3 2 3 3 4 3 3 4 3 3 3 4  
## 6 3 4 3 3 2 4 2 4 3 3 3 4 3 3 4 4 4 3 4 4  
## Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30 Q31 Q32 Q33 Q34 Q35 Q36 Q37 Q38  
## 1 3 2 2 2 2 2 2 4 4 2 3 2 2 2 2 2 2 2  
## 2 1 2 2 3 1 3 3 1 3 2 3 2 2 1 1 1 3 3  
## 3 2 2 2 2 3 3 3 4 3 3 4 4 3 2 3 4 5 4  
## 4 2 2 2 4 3 4 4 4 4 4 5 5 4 3 3 3 3 3  
## 5 3 3 3 4 3 4 4 4 4 2 3 3 2 3 3 2 3 3  
## 6 4 2 2 4 3 3 3 3 3 4 4 4 3 3 2 2 3 3  
## Q39 Q40 Q41 Q42 Gender Interest Source  
## 1 2 2 2 2 Female Weekly TV  
## 2 3 3 1 1 Female Annually TV  
## 3 4 4 1 3 Male Daily Newspapers  
## 4 4 4 3 3 Female Monthly Internet  
## 5 3 4 3 3 Female Monthly TV  
## 6 3 3 2 2 Female Monthly Internet

attach(ifanmot)  
#install.packages("psych")  
require(psych)

## Loading required package: psych

## Warning: package 'psych' was built under R version 3.1.3

#Conduct a test for sphericity on the responses for survey questions.  
test\_subjects <- ifanmot[ , 1:42]  
cortest.bartlett(test\_subjects, n=50) #Is this an identity matrix? i.e. 'perfectly independent'?

## R was not square, finding R from data

## $chisq  
## [1] 10197.66  
##   
## $p.value  
## [1] 0  
##   
## $df  
## [1] 861

Null Hypothesis: The correlation matrix is the identity matrix (shows perfect variable independance).

Alternative Hypothesis: The correlation matrix is NOT the identity matrix.

Reject null hypothesis at alpha = 0.05.

There is significant evidence to suggest that the variables are NOT perfectly independant suggesting we carry out with our test (p = 0).

### Part 1b

Compute the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) for the responses for the 42 survey questions found in columns 1 through 42 of the file ifanmot.rda.

Is the overall MSA value acceptable for factor analysis?

Should any questionnaire items be dropped from the factor analysis because of low MSA values? If so which ones?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1b -|-|-|-|-|-|-|-|-|-|-|-

KMO(test\_subjects)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = test\_subjects)  
## Overall MSA = 0.93  
## MSA for each item =   
## Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12 Q13 Q14 Q15   
## 0.95 0.91 0.96 0.90 0.91 0.97 0.80 0.94 0.96 0.76 0.95 0.92 0.96 0.96 0.94   
## Q16 Q17 Q18 Q19 Q20 Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30   
## 0.91 0.90 0.96 0.94 0.95 0.97 0.87 0.87 0.96 0.96 0.93 0.94 0.96 0.96 0.89   
## Q31 Q32 Q33 Q34 Q35 Q36 Q37 Q38 Q39 Q40 Q41 Q42   
## 0.89 0.89 0.95 0.96 0.96 0.95 0.95 0.94 0.88 0.91 0.79 0.81

The overall MSA value of 0.93 suggests that our current variable selection is a very good fit for factor analysis.

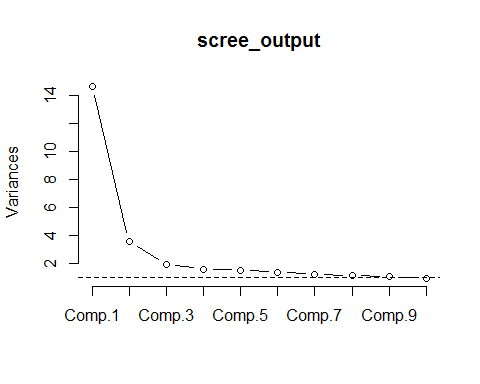
There were no variables that needed to be dropped. The lowest MSA for an individual variable was .79, which is an adequate score.

### Part 1c

Use R to create a scree plot for the questionnaire items that you deemed to be appropriate for the factor analysis from the previous question. Use the scree plot to answer the questions below.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1c -|-|-|-|-|-|-|-|-|-|-|-

scree\_output <- princomp(test\_subjects, cor=T)  
plot(scree\_output, type="lines"); abline(h=1, lty=2)



Where would you say the "knee" is in the scree plot? At the third value.

How many factors does the knee in the scree plot suggest extracting? Three.

How many components have eigenvalues (aka variances, latent roots) greater than 1? Nine.

How many factors does this suggest extracting? Nine.

### Part 1d

Use a principal components extraction with the varimax rotation to extract 3 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings.

Answer the questions below.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1d -|-|-|-|-|-|-|-|-|-|-|-

fa.out <- principal(test\_subjects, nfactors = 3, rotate = "varimax")  
print.psych(fa.out, cut=.5, sort = T)

## Principal Components Analysis  
## Call: principal(r = test\_subjects, nfactors = 3, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item PC1 PC2 PC3 h2 u2 com  
## Q13 13 0.69 0.55 0.45 1.3  
## Q3 3 0.68 0.53 0.47 1.3  
## Q5 5 0.66 0.44 0.56 1.0  
## Q6 6 0.66 0.50 0.50 1.3  
## Q9 9 0.66 0.55 0.45 1.5  
## Q15 15 0.65 0.47 0.53 1.2  
## Q16 16 0.65 0.48 0.52 1.2  
## Q37 37 0.64 0.53 0.47 1.6  
## Q38 38 0.62 0.42 0.58 1.2  
## Q28 28 0.61 0.54 0.46 1.8  
## Q26 26 0.59 0.58 0.42 1.9  
## Q8 8 0.58 0.46 0.54 1.7  
## Q4 4 0.58 0.35 0.65 1.1  
## Q36 36 0.58 0.51 0.61 0.39 2.0  
## Q17 17 0.58 0.38 0.62 1.3  
## Q1 1 0.57 0.46 0.54 1.8  
## Q20 20 0.54 0.58 0.42 2.6  
## Q40 40 0.53 0.39 0.61 1.6  
## Q27 27 0.53 0.52 0.56 0.44 2.1  
## Q29 29 0.52 0.49 0.51 2.4  
## Q39 39 0.38 0.62 1.9  
## Q12 12 0.24 0.76 1.1  
## Q14 14 0.42 0.58 2.7  
## Q19 19 0.42 0.58 2.9  
## Q10 10 0.75 0.58 0.42 1.0  
## Q7 7 0.72 0.52 0.48 1.0  
## Q41 41 0.66 0.44 0.56 1.0  
## Q22 22 0.64 0.52 0.48 1.5  
## Q42 42 0.64 0.41 0.59 1.0  
## Q2 2 0.58 0.36 0.64 1.1  
## Q23 23 0.55 0.45 0.55 1.9  
## Q35 35 0.53 0.56 0.44 2.2  
## Q18 18 0.50 0.38 0.62 2.0  
## Q11 11 0.42 0.58 2.3  
## Q34 34 0.49 0.51 2.8  
## Q21 21 0.44 0.56 2.8  
## Q32 32 0.82 0.71 0.29 1.1  
## Q31 31 0.78 0.67 0.33 1.2  
## Q30 30 0.73 0.55 0.45 1.1  
## Q33 33 0.64 0.50 0.50 1.5  
## Q25 25 0.42 0.58 2.4  
## Q24 24 0.38 0.62 2.9  
##   
## PC1 PC2 PC3  
## SS loadings 9.63 5.53 4.96  
## Proportion Var 0.23 0.13 0.12  
## Cumulative Var 0.23 0.36 0.48  
## Proportion Explained 0.48 0.27 0.25  
## Cumulative Proportion 0.48 0.75 1.00  
##   
## Mean item complexity = 1.7  
## Test of the hypothesis that 3 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.06   
## with the empirical chi square 2531.01 with prob < 1.2e-194   
##   
## Fit based upon off diagonal values = 0.97

What is the cumulative variance explained (as a percent)? 48% of cumulative variance explained.

Is this considered an acceptable percent of total variation? No. We generally set the 'reasonable' threshold for explanation at 60% of cumulative variance.

### Part 1e

Use a principal components extraction with the varimax rotation to extract 9 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings.

Answer the questions below.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1e -|-|-|-|-|-|-|-|-|-|-|-

fa.out.nine <- principal(test\_subjects, nfactors = 9, rotate = "varimax")  
print.psych(fa.out.nine, cut = .5, sort = T)

## Principal Components Analysis  
## Call: principal(r = test\_subjects, nfactors = 9, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item PC1 PC9 PC5 PC3 PC4 PC7 PC2 PC6 PC8 h2 u2  
## Q6 6 0.70 0.62 0.38  
## Q3 3 0.67 0.62 0.38  
## Q9 9 0.64 0.64 0.36  
## Q28 28 0.62 0.65 0.35  
## Q15 15 0.62 0.61 0.39  
## Q8 8 0.57 0.57 0.43  
## Q13 13 0.55 0.62 0.38  
## Q36 36 0.54 0.72 0.28  
## Q37 37 0.53 0.63 0.37  
## Q1 1 0.52 0.55 0.45  
## Q29 29 0.51 0.56 0.44  
## Q19 19 0.69 0.71 0.29  
## Q20 20 0.63 0.73 0.27  
## Q14 14 0.60 0.56 0.44  
## Q18 18 0.54 0.54 0.46  
## Q11 11 0.53 0.54 0.46  
## Q35 35 0.53 0.63 0.37  
## Q23 23 0.75 0.75 0.25  
## Q22 22 0.71 0.75 0.25  
## Q24 24 0.63 0.60 0.40  
## Q27 27 0.57 0.69 0.31  
## Q26 26 0.57 0.72 0.28  
## Q25 25 0.52 0.58 0.42  
## Q21 21 0.49 0.51  
## Q32 32 0.83 0.80 0.20  
## Q31 31 0.80 0.75 0.25  
## Q30 30 0.76 0.67 0.33  
## Q33 33 0.59 0.58 0.42  
## Q4 4 0.82 0.76 0.24  
## Q5 5 0.75 0.74 0.26  
## Q12 12 0.74 0.62 0.38  
## Q39 39 0.82 0.83 0.17  
## Q40 40 0.75 0.73 0.27  
## Q38 38 0.64 0.67 0.33  
## Q10 10 0.83 0.79 0.21  
## Q7 7 0.80 0.76 0.24  
## Q2 2 0.70 0.64 0.36  
## Q41 41 0.82 0.74 0.26  
## Q42 42 0.81 0.74 0.26  
## Q34 34 0.56 0.44  
## Q17 17 0.75 0.74 0.26  
## Q16 16 0.69 0.76 0.24  
## com  
## Q6 1.6  
## Q3 1.9  
## Q9 2.3  
## Q28 2.6  
## Q15 2.2  
## Q8 2.8  
## Q13 3.3  
## Q36 3.9  
## Q37 3.6  
## Q1 3.0  
## Q29 3.2  
## Q19 2.1  
## Q20 2.9  
## Q14 2.2  
## Q18 3.0  
## Q11 2.9  
## Q35 3.7  
## Q23 1.7  
## Q22 2.1  
## Q24 2.1  
## Q27 3.5  
## Q26 3.5  
## Q25 3.5  
## Q21 4.6  
## Q32 1.4  
## Q31 1.4  
## Q30 1.3  
## Q33 2.5  
## Q4 1.3  
## Q5 1.7  
## Q12 1.3  
## Q39 1.5  
## Q40 1.6  
## Q38 2.4  
## Q10 1.3  
## Q7 1.4  
## Q2 1.6  
## Q41 1.2  
## Q42 1.2  
## Q34 5.4  
## Q17 1.7  
## Q16 2.2  
##   
## PC1 PC9 PC5 PC3 PC4 PC7 PC2 PC6 PC8  
## SS loadings 5.65 3.67 3.66 3.31 2.49 2.45 2.38 2.26 2.11  
## Proportion Var 0.13 0.09 0.09 0.08 0.06 0.06 0.06 0.05 0.05  
## Cumulative Var 0.13 0.22 0.31 0.39 0.45 0.51 0.56 0.62 0.67  
## Proportion Explained 0.20 0.13 0.13 0.12 0.09 0.09 0.09 0.08 0.08  
## Cumulative Proportion 0.20 0.33 0.46 0.58 0.67 0.76 0.84 0.92 1.00  
##   
## Mean item complexity = 2.4  
## Test of the hypothesis that 9 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.04   
## with the empirical chi square 1014.38 with prob < 2.3e-34   
##   
## Fit based upon off diagonal values = 0.99

What is the cumulative variance explained (as a percent)? 67%

Is this considered an acceptable percent of total variation? Yes. Generally, the minimum threshold for cumulative variance explained is at 60%.

## Exercise 2

People who are concerned about their health may prefer hot dogs that are low in sodium and calories. The data file contains sample data on the sodium and calories contained in each of 54 major hot dog brands. The hot dogs are classified by type: beef, poultry, and meat (identified as types A, B, and C)

The data file called hotdogs.rda contains the sodium and calorie content for random samples of each type of hot dog. This data set is included in the DS705data package.

### Part 2a

Use the three multivariate normality tests presented this week to test for multivariate normality among the two response variables: sodium and calories. Include a chi-square quantile plot in your analysis and use a 5% level of significance for each individual hypothesis test.

According to these test, is there sufficient evidence to conclude that sodium and calories are not multivariate normal?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2a -|-|-|-|-|-|-|-|-|-|-|-

data(hotdogs)  
attach(hotdogs)  
head(hotdogs)

## calories sodium type  
## 1 93 323 A  
## 2 189 414 A  
## 3 142 559 A  
## 4 92 300 A  
## 5 156 393 A  
## 6 218 404 A

#Perform tests for multivariate normality between sodium and calories.  
#install.packages("MVN")  
require(MVN)

## Loading required package: MVN

## Warning: package 'MVN' was built under R version 3.1.3

## sROC 0.1-2 loaded

##   
## Attaching package: 'MVN'

## The following object is masked from 'package:psych':  
##   
## mardia

#1) Henze-Zinkler hzTest()  
hzTest(hotdogs[type=='A', c('calories', 'sodium')])

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "A", c("calories", "sodium")]   
##   
## HZ : 0.2193613   
## p-value : 0.9150411   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(hotdogs[type=='B', c('calories', 'sodium')])

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "B", c("calories", "sodium")]   
##   
## HZ : 0.2757632   
## p-value : 0.7535134   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(hotdogs[type=='C', c('calories', 'sodium')])

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "C", c("calories", "sodium")]   
##   
## HZ : 0.584752   
## p-value : 0.1241476   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

#2) Royston roystonTest()  
roystonTest(hotdogs[type=='A', c('calories', 'sodium')])

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "A", c("calories", "sodium")]   
##   
## H : 0.3074662   
## p-value : 0.8622045   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

roystonTest(hotdogs[type=='B', c('calories', 'sodium')])

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "B", c("calories", "sodium")]   
##   
## H : 1.956662   
## p-value : 0.3812906   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

roystonTest(hotdogs[type=='C', c('calories', 'sodium')])

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : hotdogs[type == "C", c("calories", "sodium")]   
##   
## H : 2.52345   
## p-value : 0.288548   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

#3) Mardia mardiaTest()  
mardiaTest(hotdogs[type=='A', c('calories', 'sodium')])

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : hotdogs[type == "A", c("calories", "sodium")]   
##   
## g1p : 0.2528302   
## chi.skew : 0.8427675   
## p.value.skew : 0.9326243   
##   
## g2p : 7.381651   
## z.kurtosis : -0.3456678   
## p.value.kurt : 0.7295924   
##   
## chi.small.skew : 1.071202   
## p.value.small : 0.898814   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------

mardiaTest(hotdogs[type=='B', c('calories', 'sodium')])

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : hotdogs[type == "B", c("calories", "sodium")]   
##   
## g1p : 0.3168806   
## chi.skew : 0.8978283   
## p.value.skew : 0.9248722   
##   
## g2p : 6.651096   
## z.kurtosis : -0.6952092   
## p.value.kurt : 0.4869242   
##   
## chi.small.skew : 1.188302   
## p.value.small : 0.8800208   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------

mardiaTest(hotdogs[type=='C', c('calories', 'sodium')])

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : hotdogs[type == "C", c("calories", "sodium")]   
##   
## g1p : 0.3799771   
## chi.skew : 1.076602   
## p.value.skew : 0.8979666   
##   
## g2p : 5.181197   
## z.kurtosis : -1.452778   
## p.value.kurt : 0.1462855   
##   
## chi.small.skew : 1.424914   
## p.value.small : 0.8398535   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------

#Create a chi-square quantile plot.  
#install.packages('mvoutlier')  
require(mvoutlier)

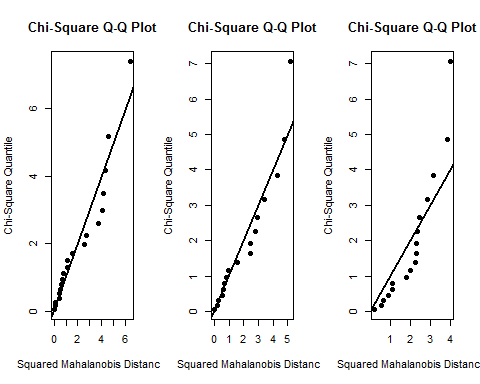
## Loading required package: mvoutlier

## Warning: package 'mvoutlier' was built under R version 3.1.3

## Loading required package: sgeostat

## Warning: package 'sgeostat' was built under R version 3.1.3

old.par <- par()  
par(mfrow=c(1,3))  
out <- with(hotdogs, hzTest(hotdogs[type=='A', c('calories', 'sodium')], qqplot = T))  
out <- with(hotdogs, hzTest(hotdogs[type=='B', c('calories', 'sodium')], qqplot = T))  
out <- with(hotdogs, hzTest(hotdogs[type=='C', c('calories', 'sodium')], qqplot = T))



par(old.par)

## Warning in par(old.par): graphical parameter "cin" cannot be set

## Warning in par(old.par): graphical parameter "cra" cannot be set

## Warning in par(old.par): graphical parameter "csi" cannot be set

## Warning in par(old.par): graphical parameter "cxy" cannot be set

## Warning in par(old.par): graphical parameter "din" cannot be set

## Warning in par(old.par): graphical parameter "page" cannot be set

The data appears to be multivariate normal as a result of the three multivariance normality tests (alpha = 0.05): Henze-Zinkler, Royston, and Mardia.

### Part 2b

Conduct Box's M Test to test for equality of covariances. Use a 5% level of significance.

1. Is there sufficient evidence to conclude that the covariance matrices are not equal?
2. Based on the criteria of multivariate normality and equal covariance matrices, is it appropriate to proceed with MANOVA?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2b -|-|-|-|-|-|-|-|-|-|-|-

source('BoxMTest.R')  
out <- BoxMTest(as.matrix(hotdogs[,1:2]), hotdogs$type)

## ------------------------------------------------  
## MBox Chi-sqr. df P  
## ------------------------------------------------  
## 2.6592 2.5074 6 0.8676  
## ------------------------------------------------  
## Covariance matrices are not significantly different.

1. There is not enough evidence to show that the population covariance matrices are different between hot dog types A, B, and C.
2. Yes.

### Part 2c

Regardless of the outcomes of the previous hypothesis tests, conduct a MANOVA to determine if there are differences between beef, poultry, and meat hot dogs for the population mean vectors when sodium and calorie content are considered together. Use the Wilk's Lambda statistic and let .

Provide the R code, output, and state the following:

1. Null and alternative hypotheses
2. P-value,
3. State the conclusion for the test

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2c -|-|-|-|-|-|-|-|-|-|-|-

hotdog\_model <- lm(cbind(sodium, calories)~type, data=hotdogs)  
m.out <- manova(hotdog\_model)  
summary(m.out, test = "Pillai")

## Df Pillai approx F num Df den Df Pr(>F)   
## type 2 0.42114 6.8018 4 102 6.802e-05 \*\*\*  
## Residuals 51   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

1. Null Hypothesis: The three types of hot dogs are similar in terms of calories and sodium content in the overall population. Alternative Hypothesis: The three types of hot dogs differ in terms of calories and sodium content in the overall population.
2. P-Value: .000068
3. There is strong evidence to suggest that the three hot dog types differ in terms of population mean calories and sodium content.

### Part 2d

Follow up with univariate ANOVAs and Tukey multiple comparisons on the response variables to see which the population means differ. Use a 5% level of significance for each univariate ANOVA and each Tukey procedure.

Write a few sentences summarizing the differences that you find (in the context of the problem).

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2d -|-|-|-|-|-|-|-|-|-|-|-

aov.model.sodium <- aov(sodium~type, data=hotdogs)  
summary(aov.model.sodium)

## Df Sum Sq Mean Sq F value Pr(>F)  
## type 2 47043 23522 2.226 0.118  
## Residuals 51 538911 10567

aov.model.calories <- aov(calories~type, data=hotdogs)  
summary(aov.model.calories)

## Df Sum Sq Mean Sq F value Pr(>F)   
## type 2 15425 7712 6.278 0.00365 \*\*  
## Residuals 51 62649 1228   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

TukeyHSD(aov.model.sodium)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = sodium ~ type, data = hotdogs)  
##   
## $type  
## diff lwr upr p adj  
## B-A -1.85000 -83.70936 80.00936 0.9983605  
## C-A 62.67941 -19.17995 144.53877 0.1643830  
## C-B 64.52941 -20.58390 149.64272 0.1700632

TukeyHSD(aov.model.calories)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = calories ~ type, data = hotdogs)  
##   
## $type  
## diff lwr upr p adj  
## B-A -9.144118 -37.05448 18.766240 0.7102899  
## C-A -39.673529 -67.58389 -11.763171 0.0033799  
## C-B -30.529412 -59.54922 -1.509603 0.0371484

Analysis of variance performed on sodium levels in the three types of hot dogs revealed there were NOT significant differences between the three types (alpha = .05). This finding was confirmed by a Tukey test which showed that there was not a significant difference between any of the mean sodium contents of hot dog types A, B and C (familywise error rate of .05).

Analysis of variance performed on calorie levels in the three types of hot dogs revealed there was a significant difference between the three types (p-value: .00365). We are 95% confident that the population mean calories for hot dog type C is between 67.6 to 11.8 calories lower than hot dog type A. In addition, the population mean calories for hot dog type C is between 59.5 and 1.5 lower than hot dog type A.

The ANOVA procedure does not account for interactions between the variables sodium and calories.

### Part 2e

Using linear discriminant analysis, construct a linear combination of sodium and calories given by LD1.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2e -|-|-|-|-|-|-|-|-|-|-|-

require('MASS')

## Loading required package: MASS

fit <- lda(type~calories+sodium, data=hotdogs)  
plda <- predict(fit, hotdogs)  
LD1 <- plda$x[,1]  
#LD2 <- plda$x[,2]

### Part 2f

Using the linear combination of sodium and calories given by LD1, conduct the Tukey HSD multiple comparisons procedure to determine which hot dog types differ with respect to this most discriminating linear combination. Use . Which pairs of manufacturers produces significantly different hotdogs on average?

Report on your results.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 2f -|-|-|-|-|-|-|-|-|-|-|-

LD1.model <- aov(LD1 ~ type, data=hotdogs)  
TukeyHSD(LD1.model)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = LD1 ~ type, data = hotdogs)  
##   
## $type  
## diff lwr upr p adj  
## B-A 0.2787548 -0.5175787 1.075088 0.6770098  
## C-A 1.8856669 1.0893335 2.682000 0.0000017  
## C-B 1.6069121 0.7789240 2.434900 0.0000621

At the 95% significance level, we can claim that the manufacturers of hot dog type A & C produce different hot dogs (p-value: 0.000002). In addition, hot dog types B & C produce different hot dogs (p-value: 0.00006). Hot dog manufacturers A & B do NOT produce significantly different hot dogs.

## Exercise 3

The depth, area, and perimeter of 31 randomly selected farm ponds in Southeastern Minnesota are recorded in the data file farmpondsize.rda.

### Part 3a

Load the data set and standardize the variables in the file (i.e. find the z-scores for each value). Store the z-scores in a new data frame.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3a -|-|-|-|-|-|-|-|-|-|-|-

data(farmpondsize)  
attach(farmpondsize)  
head(farmpondsize) #depth, area, perimeter

## depth area perim  
## 1 13.95 0.48 295.8  
## 2 16.66 0.12 129.7  
## 3 10.67 0.32 237.8  
## 4 7.13 0.39 271.0  
## 5 16.22 0.53 302.3  
## 6 10.60 0.09 122.0

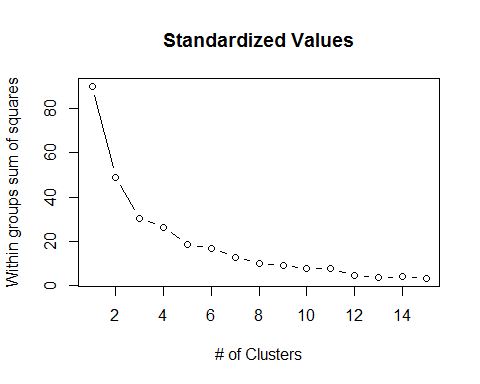
fps.z <- scale(farmpondsize)

### Part 3b

For k-means clustering, plot the within sum of squares for the first 15 clusters against the cluster number and use the plot to determine a good number of clusters to partition the cases into.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3b -|-|-|-|-|-|-|-|-|-|-|-

wss <- (nrow(fps.z)-1)\*sum(apply(fps.z, 2, var))  
for (i in 2:15) wss[i] <- sum(kmeans(fps.z, centers = i)$withinss)  
par(mfrow=c(1,1))  
plot(1:15, wss, type = 'b', xlab = "# of Clusters", ylab="Within groups sum of squares", main="Standardized Values")



The 'knee' appears to be at the third cluster - so the cases will be partitioned into three clusters.

### Part 3c

Perform the k-means clustering on the z-scores for pond depth, area, and perimeter using the number of clusters you determined from the plot. Find the number of cases in each cluster as well as the cluster means for each variable.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3c -|-|-|-|-|-|-|-|-|-|-|-

fit <- kmeans(fps.z, 3, nstart=25)  
round(aggregate(farmpondsize, by=list(fit$cluster), FUN=mean), 2)

## Group.1 depth area perim  
## 1 1 7.86 0.44 317.62  
## 2 2 10.10 0.10 123.37  
## 3 3 22.89 0.25 202.00

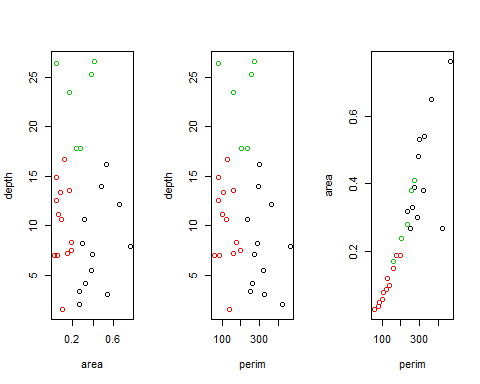
### Part 3d

Create three scatterplots, each with separate plotting colors for each cluster: plot 1: area against depth plot 2: perimeter against depth plot 3: perimeter against area.

Write a brief summary of your observations about how the ponds have been clustered with regard to the pond size variables.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3d -|-|-|-|-|-|-|-|-|-|-|-

par(mfrow=c(1,3))  
plot(farmpondsize[c("area", "depth")], col = fit$cluster)  
plot(farmpondsize[c("perim", "depth")], col = fit$cluster)  
plot(farmpondsize[c("perim", "area")], col = fit$cluster)



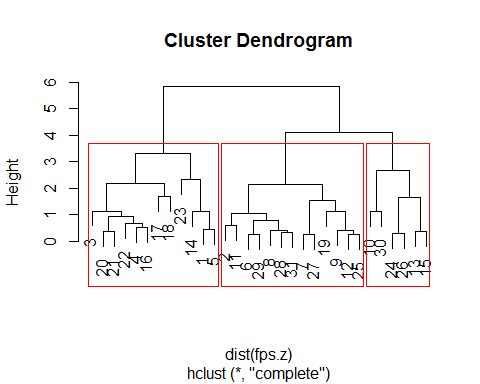
There appear to be distinct clusters when comparing the variables area vs. depth and perimeter vs. depth. While there appears to be a linear relationship between perimeter and area (as to be expected), the clustering does not seem as 'crisp.'

### Part 3e

Plot the dendrogram for hierarchical clustering using complete linkage and add the rectangles for the number of clusters you chose in part 1b above.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 3e -|-|-|-|-|-|-|-|-|-|-|-

output <- hclust(dist(fps.z))  
par(mfrow=c(1,1))  
plot(output)  
rect.hclust(output, h=4)



## Exercise 4

The data set police.rda contains 15 anthropometric and physical fitness measurements for 50 white male applicants to the police department of a major metropolitan city. Note that the first column is simple the ID number of each applicant and is not a variable.

We saw in week 13 how factor analysis could be used to reduce the number of variables for this data set. Now we will see how cluster analysis will group the 50 applicants according to these 15 variables.

### Part 4a

Load the data set and scale the variables in the file by dividing each value by the standard deviation. Store the scaled values in a new data frame.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4a -|-|-|-|-|-|-|-|-|-|-|-

data(police)  
head(police)

## ID REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST CHNUP BREATH  
## 1 1 0.310 179.6 74.20 41.7 27.3 82.4 19.0 64 64 2 158  
## 2 2 0.345 175.6 62.04 37.5 29.1 84.1 5.5 88 78 20 166  
## 3 3 0.293 166.2 72.96 39.4 26.8 88.1 22.0 100 88 7 167  
## 4 4 0.254 173.8 85.92 41.2 27.6 97.6 19.5 64 62 4 220  
## 5 5 0.384 184.8 65.88 39.8 26.1 88.2 14.5 80 68 9 210  
## 6 6 0.406 189.1 102.26 43.3 30.1 101.2 22.0 60 68 4 188  
## RECVR SPEED ENDUR FAT  
## 1 108 5.5 4 11.91  
## 2 108 5.5 4 3.13  
## 3 116 5.5 4 16.89  
## 4 120 5.5 4 19.59  
## 5 120 5.5 5 7.74  
## 6 91 6.0 4 30.42

police.clean <- police[,2:16]  
head(police.clean)

## REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST CHNUP BREATH  
## 1 0.310 179.6 74.20 41.7 27.3 82.4 19.0 64 64 2 158  
## 2 0.345 175.6 62.04 37.5 29.1 84.1 5.5 88 78 20 166  
## 3 0.293 166.2 72.96 39.4 26.8 88.1 22.0 100 88 7 167  
## 4 0.254 173.8 85.92 41.2 27.6 97.6 19.5 64 62 4 220  
## 5 0.384 184.8 65.88 39.8 26.1 88.2 14.5 80 68 9 210  
## 6 0.406 189.1 102.26 43.3 30.1 101.2 22.0 60 68 4 188  
## RECVR SPEED ENDUR FAT  
## 1 108 5.5 4 11.91  
## 2 108 5.5 4 3.13  
## 3 116 5.5 4 16.89  
## 4 120 5.5 4 19.59  
## 5 120 5.5 5 7.74  
## 6 91 6.0 4 30.42

police.s <- scale(police.clean, center = F, scale = apply(police.clean, 2, sd, na.rm = T))  
round(head(police.s), 2)

## REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST CHNUP BREATH  
## [1,] 6.43 26.79 6.47 26.29 19.00 13.80 3.11 4.96 7.89 0.46 6.21  
## [2,] 7.16 26.20 5.41 23.64 20.25 14.08 0.90 6.82 9.61 4.57 6.52  
## [3,] 6.08 24.79 6.36 24.84 18.65 14.75 3.61 7.75 10.85 1.60 6.56  
## [4,] 5.27 25.93 7.49 25.98 19.20 16.35 3.20 4.96 7.64 0.91 8.64  
## [5,] 7.96 27.57 5.74 25.09 18.16 14.77 2.38 6.20 8.38 2.05 8.25  
## [6,] 8.42 28.21 8.92 27.30 20.94 16.95 3.61 4.65 8.38 0.91 7.39  
## RECVR SPEED ENDUR FAT  
## [1,] 9.66 15.12 7.09 1.62  
## [2,] 9.66 15.12 7.09 0.43  
## [3,] 10.38 15.12 7.09 2.30  
## [4,] 10.73 15.12 7.09 2.67  
## [5,] 10.73 15.12 8.86 1.05  
## [6,] 8.14 16.50 7.09 4.14

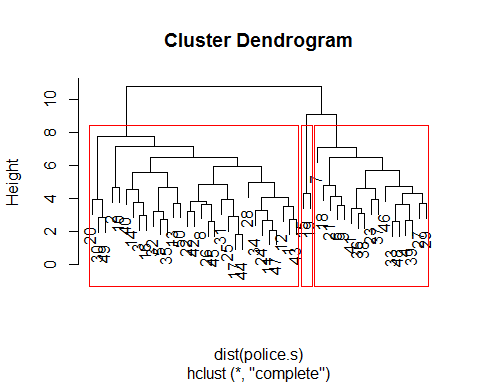
### Part 4b

Plot the dendrogram for hierarchical clustering on the scaled values using complete linkage and add the rectangles for the number of clusters given by cutting the dendrogram at a height of 8.

How many clusters does this create?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4b -|-|-|-|-|-|-|-|-|-|-|-

output <- hclust(dist(police.s))  
plot(output)  
rect.hclust(output, h=8)



Three clusters were created, although one is relatively small compared to the other two.

### Part 4c

Append the original data frame (the unscaled one) with the cluster number from cutting the dendrogram at a height of 8. Find the number of applicants in each cluster.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4c -|-|-|-|-|-|-|-|-|-|-|-

police.clean$clusternumber <- cutree(output, h=8)  
summary(as.factor(police.clean$clusternumber))

## 1 2 3   
## 31 17 2

### Part 4d

Display the sample means rounded to 2 decimal places for the 15 variables for each cluster. Recall that body fat (FAT) was an important variable in the factor analysis. What can you say about the sample means for the variable FAT for each cluster?

### -|-|-|-|-|-|-|-|-|-|-|- Answer 4d -|-|-|-|-|-|-|-|-|-|-|-

round(aggregate(police.clean, by=list(police.clean$clusternumber), FUN = mean), 2)

## Group.1 REACT HEIGHT WEIGHT SHLDR PELVIC CHEST THIGH PULSE DIAST CHNUP  
## 1 1 0.31 175.14 71.23 40.29 27.52 87.19 13.97 75.10 75.48 8.00  
## 2 2 0.32 182.79 90.01 42.17 29.07 96.01 18.88 68.12 71.29 3.76  
## 3 3 0.32 179.30 89.68 40.90 28.95 97.95 26.25 84.00 89.00 1.00  
## BREATH RECVR SPEED ENDUR FAT clusternumber  
## 1 183.68 115.29 5.48 4.38 9.89 1  
## 2 208.18 112.76 5.59 3.88 19.27 2  
## 3 217.00 143.00 4.50 3.50 27.51 3

The sample means for the variable "FAT" differ significantly between clusters 1, 2 and 3 (9.89, 19.27, and 27.51 respectively).