Lab Assignment: Intro to Factor Analysis

Spencer Swartz

April 18, 2017

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

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

## Loading required package: psych

## Warning: package 'psych' was built under R version 3.3.3

data("ifanmot")  
mat <- cor(ifanmot[,1:42])  
cortest.bartlett(mat,n=400)

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

Hypothesis: : The correlation matrix is the identity matrix : The correlation matrix is not the identity matrix

Result: We reject H\_0 (p=0), there is significant evidence that the correlation matrix is not the identity matrix.

### 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(mat)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = mat)  
## 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

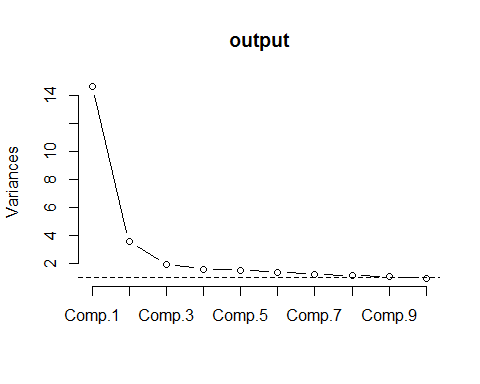
The overall MSA is Superb for this data set, in addition it does not look like any of the questions should be dropped as the lowest individual MSA values is .76(Q10).

### 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 -|-|-|-|-|-|-|-|-|-|-|-

output <- princomp(ifanmot[,1:42], cor=TRUE)  
plot(output,type="lines") # scree plot   
abline(h=1,lty=2) # add horizonal dotted line at 1



output$sdev^2

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6   
## 14.6299304 3.5432306 1.9437153 1.5963579 1.5219162 1.3529379   
## Comp.7 Comp.8 Comp.9 Comp.10 Comp.11 Comp.12   
## 1.2186202 1.1171978 1.0584621 0.9296358 0.8727509 0.8261235   
## Comp.13 Comp.14 Comp.15 Comp.16 Comp.17 Comp.18   
## 0.7768238 0.7289769 0.6499980 0.5977686 0.5710810 0.5531995   
## Comp.19 Comp.20 Comp.21 Comp.22 Comp.23 Comp.24   
## 0.5177140 0.5132560 0.4691781 0.4532074 0.4282455 0.4164713   
## Comp.25 Comp.26 Comp.27 Comp.28 Comp.29 Comp.30   
## 0.3953515 0.3878640 0.3693198 0.3519037 0.3316491 0.3133673   
## Comp.31 Comp.32 Comp.33 Comp.34 Comp.35 Comp.36   
## 0.2854463 0.2750360 0.2636268 0.2451903 0.2338960 0.2245353   
## Comp.37 Comp.38 Comp.39 Comp.40 Comp.41 Comp.42   
## 0.2132297 0.1930199 0.1807200 0.1746080 0.1568402 0.1175977

Where would you say the "knee" is in the scree plot?

I would say the knee is at Comp.3

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

3

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

9

How many factors does this suggest extracting?

9

### 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(ifanmot[,1:42],nfactors=3,rotate="varimax")  
print.psych(fa.out,cut=.5,sort=TRUE)

## Principal Components Analysis  
## Call: principal(r = ifanmot[, 1:42], nfactors = 3, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item RC1 RC2 RC3 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  
##   
## RC1 RC2 RC3  
## 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)?

the first 3 factors explained 48% of the total variation among the 39 remaining variables.

Is this considered an acceptable percent of total variation?

No we are looking for at least 60%

### 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 <- principal(ifanmot[,1:42],nfactors=9,rotate="varimax")  
print.psych(fa.out,cut=.5,sort=TRUE)

## Principal Components Analysis  
## Call: principal(r = ifanmot[, 1:42], nfactors = 9, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item RC1 RC9 RC5 RC3 RC4 RC7 RC2 RC6 RC8 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  
##   
## RC1 RC9 RC5 RC3 RC4 RC7 RC2 RC6 RC8  
## 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)?

The first 9 factors explained 67% of the total variation among the 32 remaining variables.

Is this considered an acceptable percent of total variation?

Yes this is above 60%C

### Part 1f

Read the questions in the survey (IFM Survey.docx) for the groups of items that load onto each factor and put a comprehensive label on each of the 9 factors from the most recent factor analysis (extracting 9 factors with principal components and a varimax rotation).

For consistency assign the following 9 labels to the most appropriate factors:

**Artistic, Boredom, Entertainment, Fun, Gambling, Identification, Loyalty, Patriotism, Social**

Factors 1 through 9 move from left to right in the sorted output (even though the columns labels PC1-PC9 in the output are not in order).

I have labeled the second factor (labeled PC9 in the R output) for you as "Social". Look at the survey items that correspond to the second factor (Q11, Q14, Q18, Q19, Q20, and Q35) and see if that label makes sense.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1f -|-|-|-|-|-|-|-|-|-|-|-

Factor 1: Entertainment Factor 2: Social Factor 3: Identification Factor 4: Patriotism Factor 5: Artistic Factor 6: Loyalty Factor 7: Gambling Factor 8: Boredom Factor 9: Fun

### Part 1g

Combine the factor scores produced by the 9-factor solution with the original data frame. Also, rename the factor scores using the labels you assigned previously. Some R code to begin this has been provided. Add to it to complete this request.

### -|-|-|-|-|-|-|-|-|-|-|- Answer 1g -|-|-|-|-|-|-|-|-|-|-|-

fan <- principal(ifanmot[,1:42],nfactors=9,rotate="varimax")  
fulldata <- cbind(ifanmot,fan$scores)  
require(plyr)

## Loading required package: plyr

fulldata <- rename(fulldata,c("RC1"="Entertainment","RC9"="Social","RC5"="Identification","RC3"="Patriotism","RC4"="Artistic","RC7"="Loyalty","RC2"="Gambling","RC6"="Boredom","RC8"="Fun")) # REPLACE the ? with your factor labels

## 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")  
typeA <- as.matrix(hotdogs[hotdogs$type=="A", 1:2 ])  
#typeA <- as.numeric(typeA)  
typeB <- as.matrix(hotdogs[hotdogs$type=="B", 1:2 ])  
#typeB <- as.numeric(typeB)  
typeC <- as.matrix(hotdogs[hotdogs$type=="C", 1:2 ])  
#typeC <- as.numeric(typeC)  
#install.packages('MVN')  
require(MVN)

## Loading required package: MVN

## Warning: package 'MVN' was built under R version 3.3.3

## sROC 0.1-2 loaded

##   
## Attaching package: 'MVN'

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

hzTest(typeA)

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeA   
##   
## HZ : 0.2193613   
## p-value : 0.9150411   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

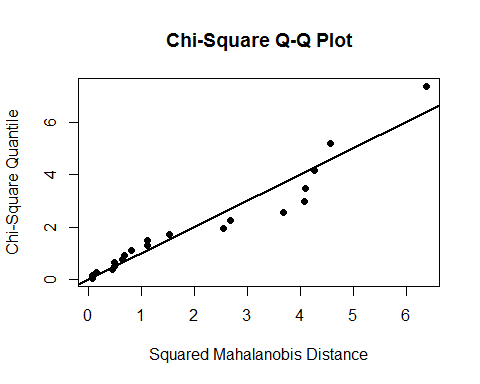
mardiaTest(typeA)

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : typeA   
##   
## 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.   
## ---------------------------------------

roystonTest(typeA)

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeA   
##   
## H : 0.3074662   
## p-value : 0.8622045   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(typeA,qqplot=TRUE)



## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeA   
##   
## HZ : 0.2193613   
## p-value : 0.9150411   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(typeB)

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeB   
##   
## HZ : 0.2757632   
## p-value : 0.7535134   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

mardiaTest(typeB)

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : typeB   
##   
## 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.   
## ---------------------------------------

roystonTest(typeB)

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeB   
##   
## H : 1.956662   
## p-value : 0.3812906   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(typeB,qqplot=TRUE)

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeB   
##   
## HZ : 0.2757632   
## p-value : 0.7535134   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(typeC)

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeC   
##   
## HZ : 0.584752   
## p-value : 0.1241476   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

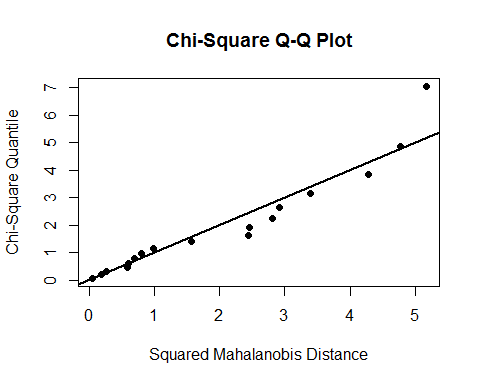
mardiaTest(typeC)

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : typeC   
##   
## 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.   
## ---------------------------------------

roystonTest(typeC)

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeC   
##   
## H : 2.52345   
## p-value : 0.288548   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

hzTest(typeB,qqplot=TRUE)



## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : typeB   
##   
## HZ : 0.2757632   
## p-value : 0.7535134   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

No, there is not sufficient evidence based on these tests to conclude that sodium and calories are not multivariate normal.

### 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 evidence to show the population covariance matrices are different.
2. All three criteria are met for a MANOVA Test, so yes we may proceed.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

lmodel <- lm(cbind(calories,sodium)~type,data=hotdogs)  
m.out <- manova(lmodel)  
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

(i)

$$ H\_0: \pmb{\mu}\_{typeA} = \pmb{\mu}\_{typeB} = \pmb{\mu}\_{typeC}$$

1. Pvalue = 0.00006
2. Reject . There is strong evidence to show the three types of hotdogs are different in terms of population mean calories and sodium.

### 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 -|-|-|-|-|-|-|-|-|-|-|-

summary.aov(m.out)

## Response calories :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## type 2 15425 7712.5 6.2784 0.003651 \*\*  
## Residuals 51 62649 1228.4   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response sodium :  
## Df Sum Sq Mean Sq F value Pr(>F)  
## type 2 47043 23522 2.226 0.1183  
## Residuals 51 538911 10567

The univariate ANOVAs suggest that is a significant difference in the population mean in calories but no significant differences in population mean sodium when considered separately.

### 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) # fit model  
plda <- predict(object=fit, newdata=hotdogs) # compute combinations  
ld1 <- plda$x[,1] # extract most separating combination

### 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 -|-|-|-|-|-|-|-|-|-|-|-

linear.model<-aov(ld1~type,data=hotdogs)  
TukeyHSD(linear.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 5% significance level there is strong evidence that the population mean calories and sodium for the hotdogs for type A and B both differ from those of type C. There is not a significant difference in population mean calories and sodium for the apes at locations A and B.