Final Exam

Chris Austin

You may use this RMD file to work out the final exam solutions and then transfer your answers to the D2L final exam. \*\* Do not distribute or share this file in any way \*\*

## Problem 1

**Scenario**: A survey was conducted to find out how teenagers think about the future and barriers they think will hinder their career. The researcher would like to determine if the 15 survey items can be summarized more efficiently by a smaller set of latent factors.

You will need these files:

* **careerbarrier.rda**: Data for the 15 survey items for a random sample of 76 teens.
* **Career Barrier Survey.docx**: Descriptions of the variables in the data file.

### Part A

Conduct Bartlett's test for sphericity on the responses for the 15 survey questions.

#### Question 1 - Insert your R code here.

require(DS705data)  
require(psych)  
setwd("~/DS705 - Statistical Methods/Week15/Final Exam Datasets")  
load("./careerbarrier.rda")  
  
corMatrix <- cor(careerbarrier)  
cortest.bartlett(corMatrix, n=nrow(careerbarrier))

## $chisq  
## [1] 287.4985  
##   
## $p.value  
## [1] 6.756016e-19  
##   
## $df  
## [1] 105

# Correlation combinations having absolute magnitude of .5 or greater   
# (eliminating 1.0 diagonals and dividing by two since in both halves of the matrix)  
sum(abs(corMatrix) > .5 & abs(corMatrix) < 1)/2

## [1] 3

# Correlation combinations having absolute value .6 or greater   
# (eliminating 1.0 diagonals and dividing by two since in both halves of the matrix)  
sum(abs(corMatrix) > .6 & abs(corMatrix) < 1)/2

## [1] 1

#### Question 2 - State the null and alternative hypothesis

The null hypothesis is that the correlation matrix is the identity matrix (values along the diagonals).

The alternate hypothesis is that the correlation matrix is not the identity matrix.

#### Question 3 - State your conclusion at a 5% level of significance and respond with whether factor analysis is warranted based on this measure.

With a .05 level of significance, we can reject the null hypothesis. There is sufficient evidence, given a p-value of nearly zero, that the correlations of the variables are such that the dataset is suitable for factor analysis.

Of the correlation pairs calculated, 3 are at 0.5 and above, and 1 is at 0.6 and above.

#### Question 4 - Round the p-value to four decimal places (enter 0 if P < 0.00005).

0

### Part B1

Compute the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) for the responses for the 16 survey questions.

#### Question 5 - Insert your R code here.

KMO(careerbarrier)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = careerbarrier)  
## Overall MSA = 0.68  
## MSA for each item =   
## money lazy law noexp math support health reading   
## 0.36 0.67 0.67 0.61 0.67 0.66 0.67 0.73   
## english aoda grades disc social relatshp looks   
## 0.69 0.82 0.67 0.66 0.75 0.72 0.58

#### Question 6 - Report the overall MSA value.

.68

#### Question 7 - Is the overall MSA value acceptable for factor analysis?

Yes

#### Question 8 - Should any questionnaire items be dropped from the factor analysis because of MSA values under 0.50?

Yes

#### Question 9 - If so which one(s)? (if there aren't any, write "none")

money

### Part B2

Compute the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy (MSA) for the responses for the remaining survey questions **after you have dropped the items from Part B1**. Use the following questions to document your findings.

#### Question 10 - Insert your R code here:

careerbarrier2 <- careerbarrier[,2:14]  
KMO(careerbarrier2)

## Kaiser-Meyer-Olkin factor adequacy  
## Call: KMO(r = careerbarrier2)  
## Overall MSA = 0.71  
## MSA for each item =   
## lazy law noexp math support health reading english   
## 0.67 0.67 0.59 0.83 0.66 0.69 0.74 0.70   
## aoda grades disc social relatshp   
## 0.82 0.69 0.62 0.73 0.75

#### Question 11 - Report the overall MSA value.

.71

#### Question 12 - Is the new overall MSA value acceptable for factor analysis?

Yes

#### Question 13 - Should any questionnaire items be dropped from the factor analysis because of MSA values under 0.50?

No

#### Question 14 - If so which one(s)? (if there aren't any say "none")

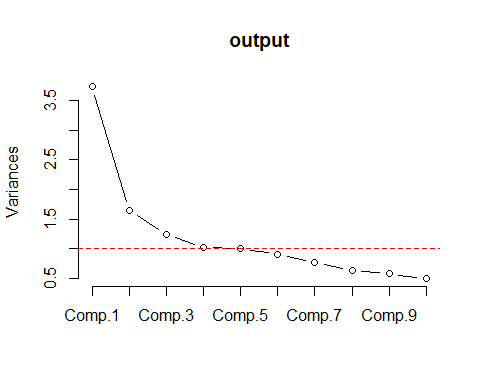
none

### Part C

Use R to create a scree plot for the questionnaire items that you deemed to be appropriate for the factor analysis from Part B.

#### Question 15 - Use R to create a scree plot for the questionnaire items that you deemed to be appropriate for the factor analysis from Part B. Insert your R code here.

output <- princomp(careerbarrier2, cor=TRUE)  
plot(output, type="lines")  
abline(h=1, lty=2, col="red")



output$sdev

## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7   
## 1.9347082 1.2836446 1.1140495 1.0102436 0.9993705 0.9497331 0.8792791   
## Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13   
## 0.7983645 0.7567968 0.7043358 0.6103553 0.5795020 0.5089437

#### Question 16 - Using the knee in the scree plot, how many factors should be extracted?

3

#### Question 17 -How many components have eigenvalues (aka variances, latent roots) greater than 1 and how many factors does this suggest extracting?

4

### Part D

Use a principal components extraction with the varimax rotation to extract 5 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings.

#### Question 18 - Use a principal components extraction with the varimax rotation to extract 5 factors. Print the output with factor loadings under 0.5 suppressed and sort the loadings. Insert your R code here.

factors <- principal(careerbarrier2, nfactors=5, rotate="varimax")  
print.psych(factors, cut=.5, sort=TRUE)

## Principal Components Analysis  
## Call: principal(r = careerbarrier2, nfactors = 5, rotate = "varimax")  
## Standardized loadings (pattern matrix) based upon correlation matrix  
## item RC1 RC3 RC2 RC4 RC5 h2 u2 com  
## reading 7 0.84 0.78 0.22 1.2  
## english 8 0.76 0.71 0.29 1.5  
## health 6 0.57 0.59 0.41 2.5  
## aoda 9 0.54 0.61 0.39 2.7  
## lazy 1 0.81 0.76 0.24 1.3  
## law 2 0.78 0.66 0.34 1.2  
## grades 10 0.64 0.53 0.47 1.5  
## relatshp 13 0.75 0.62 0.38 1.2  
## disc 11 0.74 0.55 0.45 1.0  
## social 12 0.69 0.56 0.44 1.3  
## noexp 3 0.90 0.87 0.13 1.1  
## support 5 0.89 0.84 0.16 1.1  
## math 4 0.57 0.43 3.6  
##   
## RC1 RC3 RC2 RC4 RC5  
## SS loadings 2.18 2.14 2.00 1.17 1.16  
## Proportion Var 0.17 0.16 0.15 0.09 0.09  
## Cumulative Var 0.17 0.33 0.49 0.58 0.67  
## Proportion Explained 0.25 0.25 0.23 0.14 0.13  
## Cumulative Proportion 0.25 0.50 0.73 0.87 1.00  
##   
## Mean item complexity = 1.6  
## Test of the hypothesis that 5 components are sufficient.  
##   
## The root mean square of the residuals (RMSR) is 0.08   
## with the empirical chi square 83.38 with prob < 8.9e-09   
##   
## Fit based upon off diagonal values = 0.89

#### Question 19 - What is the cumulative variance explained? Answer as a percent, not a decimal number.

67%

#### Question 20 - Is this considered an acceptable percent of total variation?

Yes (60% is the recognized minimum cutoff.)

## Problem 2

**Scenario**: Water quality variables nitrogen, turbidity, phosphorus, dissolved oxygen, temperature, and conductivity are measured in 31 randomly selected farm ponds in Southeastern Minnesota. Researchers would like to determine if there is an underlying structure that will enable clustering of these 31 ponds into homogeneous groups. You will need this file: **farmpondquality.rda**

### Part A

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.

#### Question 21 - Insert your R code here.

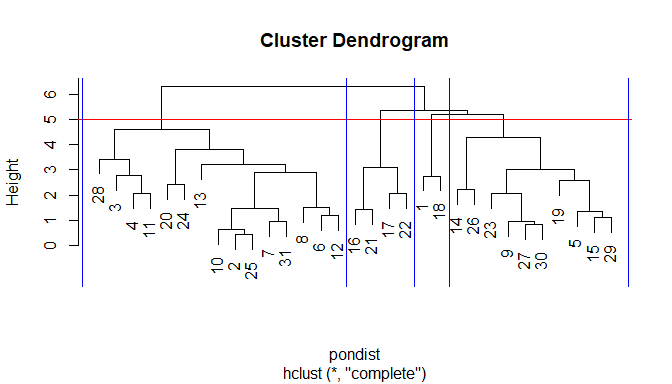
data("farmpondquality")  
pond <- scale(farmpondquality)

### Part B

Plot the dendrogram for hierarchical clustering using complete linkage and add the rectangles by cutting the dendrogram at a height of 5.

#### Question 22 - Insert your R code here.

pondist <- dist(pond)  
  
# "Complete" linkage is hclust default.  
pondOutput <- hclust(pondist)  
plot(pondOutput)  
abline(h=5, col="red")  
  
abline(v=0, col="blue")  
abline(v=15.5, col="blue")  
abline(v=19.5, col="blue")  
abline(v=21.5, col="blue")  
abline(v=32, col="blue")



#### Question 23 - How many clusters does this form?

4

### Part C

Append the original data frame (the unscaled one) with the cluster number from cutting the dendrogram at a height of 5. Find the number of ponds in each cluster and obtain the means of the original variables for each cluster.

#### Question 24 - Append the original data frame (the unscaled one) with the cluster number from cutting the dendrogram at a height of 5. Find the number of ponds in each cluster and obtain the means of the original variables for each cluster. Copy your R code *with its output* here.

farmpondquality$cluster <- cutree(pondOutput, h=5)  
table(farmpondquality$cluster)

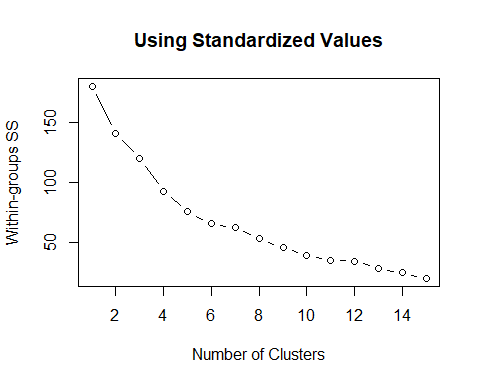
##   
## 1 2 3 4   
## 2 15 10 4

(round(aggregate(farmpondquality[,1:6], by=list(farmpondquality$cluster), FUN=mean),2))

## Group.1 nitro turb phos disoxy temp cond  
## 1 1 0.79 13.77 2.51 9.51 20.62 346.75  
## 2 2 0.30 28.45 1.88 9.22 23.08 149.63  
## 3 3 0.18 9.09 1.63 11.51 20.41 206.72  
## 4 4 0.18 9.18 0.93 8.55 22.57 488.56

#### Question 25 - 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. Use the standardized values. Insert your R code here.

wss <- (nrow(pond)-1) \* sum(apply(pond,2,var))  
for(i in 2:15) wss[i] <- sum(kmeans(pond, centers = i)$withinss)  
plot(1:15, wss, type="b", xlab="Number of Clusters", ylab = "Within-groups SS", main = "Using Standardized Values")



#### Question 26 - How many clusters do you think are best, based on this plot? Why?

Subjectively, the "knee" in the plot looks to level out at around 5 clusters.

#### Question 27 - Perform the k-means clustering on the z-scores of the 6 pond quality variables 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 the raw variables. Insert your R code here.

fit <- kmeans(pond, 5, nstart=25)  
print(round(aggregate(farmpondquality[,1:6], by=list(fit$cluster), FUN=mean),2))

## Group.1 nitro turb phos disoxy temp cond  
## 1 1 0.39 54.28 1.83 6.22 22.72 147.33  
## 2 2 0.18 14.89 1.14 10.90 21.66 131.44  
## 3 3 0.67 11.83 1.85 9.03 22.40 260.64  
## 4 4 0.20 8.75 1.02 9.10 21.80 471.90  
## 5 5 0.22 15.86 3.29 11.56 22.08 231.78

table(fit$cluster)

##   
## 1 2 3 4 5   
## 4 12 4 5 6

## Problem 3

A study was conducted on the relationship of seating position and nausea on buses. The data in the following table classifies each person in a random sample of bus riders by the location of his or her seat and whether or not nausea was reported.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table | Front | Middle | Rear | Total |
| Nausea | 98 | 110 | 161 | 369 |
| No Nausea | 264 | 321 | 280 | 865 |
| Total | 362 | 431 | 441 | 1234 |

### Part A

Test to see whether or not the seat position within a bus is associated with motion sickness.

#### Question 28 - Insert your R code here.

seating <- matrix(c(98,264, 110,321, 161,280), nrow=2)  
rownames(seating) <- c("Nausea", "NoNausea")  
colnames(seating) <- c("Front", "Middle", "Rear")  
  
(seatingChi <- chisq.test(seating, correct = FALSE))

##   
## Pearson's Chi-squared test  
##   
## data: seating  
## X-squared = 14.509, df = 2, p-value = 0.000707

seatingChi$expected

## Front Middle Rear  
## Nausea 108.248 128.8809 131.8712  
## NoNausea 253.752 302.1191 309.1288

seatingChi$observed

## Front Middle Rear  
## Nausea 98 110 161  
## NoNausea 264 321 280

#### Question 29 - State the null and alternative hypothesis.

The null hypothesis is that the row and column variables are independent.

The alternative hypothesis is that the row (indication of nausea) and column (location on bus) variables are associated.

#### Question 30 - State the test statistic. Give the answers to the nearest thousandth decimal.

14.509

#### Question 31 - State the degrees of freedom.

2

#### Question 32 - Round the p-value to four decimal places (enter 0 if P < 0.00005).

.0007

#### Question 33 - State your conclusion.

Within a .05 level of significance, we can reject the null hypothesis.

Enough evidence exists, at a p-value of .0007, to state that there is an association between an indication of nausea and the area in which a passenger sits on buses.

### Part B

Construct a 90% confidence interval (without the Yates correction) for the difference in population proportions of all bus riders in the front who report nausea and all bus riders in the rear who report nausea. (Use Diff = Front - Rear)

#### Question 34 - Insert your R code here.

frontRear <- c(98,161)  
frontRearCounts <- c(362,441)  
prop.test(frontRear, frontRearCounts, conf.level = .90, correct = FALSE)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: frontRear out of frontRearCounts  
## X-squared = 8.1012, df = 1, p-value = 0.004424  
## alternative hypothesis: two.sided  
## 90 percent confidence interval:  
## -0.14819088 -0.04053138  
## sample estimates:  
## prop 1 prop 2   
## 0.2707182 0.3650794

#### Question 35 - Enter the lower bound of the 90% CI (round to 3 decimal places).

-.148

#### Question 36 - Enter the upper bound of the 90% CI (round to 3 decimal places).

-.041

#### Question 37 - Write an interpretation for the interval in the context of the problem.

The probability of being nauseous when sitting in the front of the bus is between 4.1% and 14.8% less than being nauseous when sitting in the back of the bus.

### Part C

Compute the odds ratio of having nausea for those in the rear compared to those in the front of the bus. Report the odds ratio to 3 decimal places.

#### Question 38 - Insert your R code here.

# Successes (having nausea) are in column 1  
# Leading treatment of interest (seated in the rear of the bus) is in row 2  
(seatingFrontRear <- matrix(c(98,161,264,280), nrow=2, dimnames=list(c("Front", "Rear"), c("Nausea", "NoNausea"))))

## Nausea NoNausea  
## Front 98 264  
## Rear 161 280

#Compute the odds ratio of having nausea for those in the rear compared to those in the front of the bus.   
# P(A: Having nausea | B: in the rear of the bus) = P(A&B)/P(B)  
(a <- (161/sum(seatingFrontRear))/((161+280)/sum(seatingFrontRear)))

## [1] 0.3650794

# P(A: Having nausea | B: in the front of the bus) = P(A&B)/P(B)  
(b <- (98/sum(seatingFrontRear))/((98+264)/sum(seatingFrontRear)))

## [1] 0.2707182

# Relative risk for having nausea for those in the rear as compared to those in the front:   
round(a/b,3)

## [1] 1.349

# odds of having nausea if sitting in rear:  
round(a/(1-a),3)

## [1] 0.575

# odds of having nausea if sitting in front:   
round(b/(1-b),3)

## [1] 0.371

# odds ratio of nausea from sitting in rear compared to front:  
round((a/(1-a))/(b/(1-b)), 3)

## [1] 1.549

# confirmed by:  
mosaic::oddsRatio(seatingFrontRear, verbose=TRUE)

##   
## Odds Ratio  
##   
## Proportions  
## Prop. 1: 0.2707   
## Prop. 2: 0.3651   
## Rel. Risk: 1.349   
##   
## Odds  
## Odds 1: 0.3712   
## Odds 2: 0.575   
## Odds Ratio: 1.549   
##   
## 95 percent confidence interval:  
## 1.094 < RR < 1.662   
## 1.145 < OR < 2.096   
## NULL

## [1] 1.54898

#### Question 39 - Report the odds ratio to 3 decimal places.

The odds ratio is 1.549

#### Question 40 - Interpret the odds ratio in the context of the problem.

The odds of getting nausea when sitting in the rear of the bus are 54.9% more than the odds of getting nausea when sitting in the front of the bus.

## Problem 4

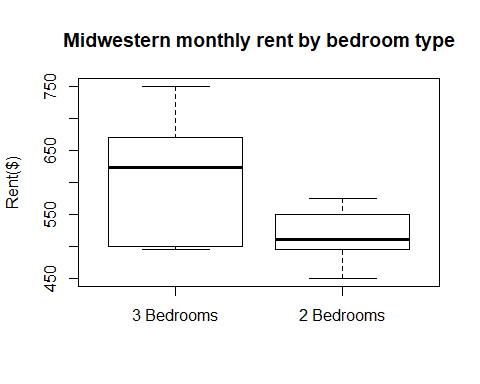
**Scenario** : A random sample of apartments was obtained from mid-sized towns in the Midwest. They are classified as having either "3 bedrooms" or "2 bedrooms" and the monthly rent was recorded. You will need this file: **monthlyrent.rda**

### Part A

Load the data set and create boxplots for the monthly rents for each type of apartment.

#### Question 41 - Insert your R code here.

data("monthlyrent")  
boxplot(rent~type, data=monthlyrent, ylab="Rent($)", main = "Midwestern monthly rent by bedroom type")



#### Question 42 - Comment on the shapes of the boxplots and whether or not they contain outliers. Does there appear to be a difference in the distributions of monthly rent between 2 and 3 bedroom apartments?

The distributions do not contain outliers. Apartments with 3 bedrooms look to have a much wider variance (given the comparative sizes of the first to third-quantile box) and a somewhat rightward-skewed distribution as compared to the 2-bedroom rent distribution.

The 2-bedroom boxplot shows a much smaller range between first and third quartile. These distributions look to be noticably different from one another.

### Part B

Conduct the Shapiro-Wilk test for normality for each sample.

#### Question 43 - Conduct the Shapiro-Wilk test for normality for each sample. Insert your R code here.

shapiro.test(monthlyrent$rent[monthlyrent$type == "2 Bedrooms"])

##   
## Shapiro-Wilk normality test  
##   
## data: monthlyrent$rent[monthlyrent$type == "2 Bedrooms"]  
## W = 0.96512, p-value = 0.8423

shapiro.test(monthlyrent$rent[monthlyrent$type == "3 Bedrooms"])

##   
## Shapiro-Wilk normality test  
##   
## data: monthlyrent$rent[monthlyrent$type == "3 Bedrooms"]  
## W = 0.90679, p-value = 0.2596

#### Question 44 - Using a 5% level of significance for each test individually, choose the option that describes conclusions for each distribution.

* 1. Normality is rejected for the rents of both the 3-bedroom and 2-bedroom apartments.
  2. Normality is rejected for the rents of the 2-bedroom apartments but not the 3-bedroom apartments.
  3. Normality is rejected for the rents of the 3-bedroom apartments but not the 2-bedroom apartments.
  4. Normality is not rejected for the rents of either the 3-bedroom or 2-bedroom apartments.

D

### Part C

Test the for equality of population variances using a 5% level of significance.

#### Question 45 - Test for equality of population variances using a 5% level of significance. Insert your R code here.

require(car)  
leveneTest(rent~type, data=monthlyrent)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 1 7.5154 0.01342 \*  
## 18   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### Question 46 - State the null and alternative hypotheses.

The null hypothesis is that the population variances are equal.

The alternative hypothesis is that the population variances are not equal.

#### Question 47 - Round the p-value to four decimal places (enter 0 if P < 0.00005).

.0134

#### Question 48 - State the conclusion for the test.

At a .05 level of significance, we can reject the null hypothesis. At a p-value of .0134, there is enough evidence to conclude that the population variance of the rents for two-bedroom apartments in the Midwest is different than the population variance of the rents for three-bedroom apartments in the Midwest.

### Part D

Conduct the appropriate hypothesis test (*for two samples only - not ANOVA*) to compare the population mean rents for these two types of apartments in mid-sized Midwestern towns. Use a 10% level of significance.

#### Question 49 - Insert your R code here.

# Satisfaction of normality leads us to a t-test.  
# Since the test is to compare rather than provide direction of difference, a two-tailed test is used.  
# The failure of the Levene test to show equal variances requires us to use var.equal = FALSE.  
  
t.test(rent~type, data=monthlyrent, var.equal=FALSE, conf.level = .90)

##   
## Welch Two Sample t-test  
##   
## data: rent by type  
## t = 3.0105, df = 12.32, p-value = 0.01057  
## alternative hypothesis: true difference in means is not equal to 0  
## 90 percent confidence interval:  
## 37.65164 146.34836  
## sample estimates:  
## mean in group 3 Bedrooms mean in group 2 Bedrooms   
## 608 516

#### Question 50 - State the hypotheses.

The null hypothesis is that the mean of two-bedroom rent prices is the same is the mean of three-bedroom rent prices (the difference between the two is zero.)

The alternate hypothesis is that the mean of two-bedroom rent prices is the different from the mean of three-bedroom rent prices (the difference between the two is non-zero.)

#### Question 51 - State the test statistic.

t = 3.0105

#### Question 52 - State the df.

df = 12.32

#### Question 53 - Round the p-value to four decimal places (enter 0 if P < 0.00005).

p = .0106

#### Question 54 - State the conclusion.

At a .10 level of significance, we can reject the null hypothesis. Given a p-value of .0106, we have enough evidence to state that the mean rents for two and three-bedroom apartments in the Midwest are different.

#### Question 55 - Also obtain the appropriate 90% confidence interval for the difference in population mean rents for these two types of apartments in mid-sized Midwestern towns. (you can add your R code in Question 49). Write an interpretation for the interval in the context of the problem.

With a .90 level of confidence, we can say that the difference in the means of 2-bedroom and 3-bedroom apartments in the Midwest is between $37.65 and $146.35.

## Problem 5

**Scenario**: A psychologist is interested in the relationship between academic performance and “self-concept” as well as the student’s IQ and gender for 39 seventh grade students in a rural school district. Academic performance is measured as a grade point average (let y = GPA). Self-concept (x1) is measured by the student’s score on the Piers-Harris Children’s Self-Concept Scale, the IQ (x2) and the type of learner (x3, 0=Visual, 1=Auditory) of each student is also recorded. You will need this file: **gpa7th.rda**

### Part A

Consider the model

Fit this model and report on which coefficients are significantly different from zero. Use the hierarchical approach to model-building.

#### Question 56 - Insert your R code here.

data("gpa7th")  
model1 <- lm(y ~ x1 + x2 + x3 + x1:x3 + x2:x3, data = gpa7th)  
summary(model1)

##   
## Call:  
## lm(formula = y ~ x1 + x2 + x3 + x1:x3 + x2:x3, data = gpa7th)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.20709 -0.30514 0.03727 0.25057 1.15223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.963561 1.221225 -2.427 0.0209 \*   
## x1 0.040486 0.008774 4.614 5.72e-05 \*\*\*  
## x2 0.031657 0.011863 2.669 0.0117 \*   
## x3Auditory 2.074237 1.652304 1.255 0.2182   
## x1:x3Auditory -0.034406 0.015210 -2.262 0.0304 \*   
## x2:x3Auditory 0.002440 0.015614 0.156 0.8768   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4805 on 33 degrees of freedom  
## Multiple R-squared: 0.6653, Adjusted R-squared: 0.6146   
## F-statistic: 13.12 on 5 and 33 DF, p-value: 4.572e-07

#### Question 57 - Should any terms be dropped from this model at a 5% level of significance?

Yes

#### Question 58 - Select the term(s) that should be dropped.

Only the x2:x3 cross-term should be dropped at a .05 level of signinficance. Even though the x3 factor is above the .05 significance threshold, because x3 is part of a significant x1:x3 relationship, x3 itself cannot be eliminated from the model.

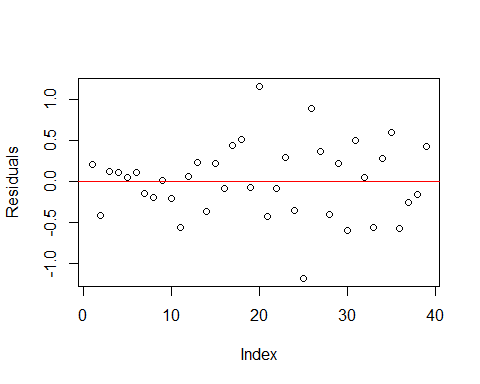
### Part B

Fit the model.

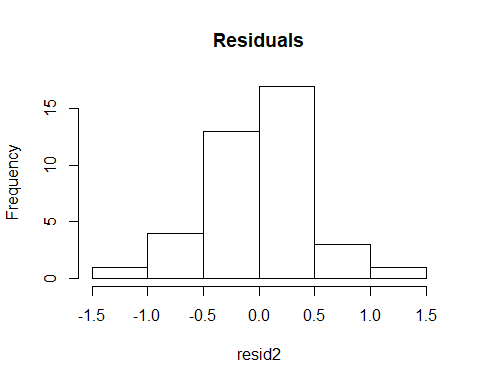
Obtain the residuals for this model and evaluate the residual plots using the "plot" function. Also create a histogram of the residuals.

#### Question 59 - Insert your R code here.

require(lmtest)  
model2 <- lm(y ~ x1 + x2 + x3 + x1:x3, data = gpa7th)  
resid2 <- model2$residuals  
plot(resid2, ylab="Residuals")  
abline(h=0, col="red")



hist(resid2, main = "Residuals")



bptest(model2)

##   
## studentized Breusch-Pagan test  
##   
## data: model2  
## BP = 9.0755, df = 4, p-value = 0.05924

#### Question 60 - Does a visual inspection of the residual plots and histogram indicate that the model assumptions appear to be satisfied? Explain your answer.

The scatterplot of residuals does not note any serially-correlated patterns. Neither does it note any patterns in the magnitude either above or below the zero line. As such, the scatterplot would lead one to believe that the residuals are homoscadastistic.

The results of the histogram are also in line with this assumption. The plot peaks around zero and is nearly symmetric with no obvious outliers.

#### Question 61 - Also perform a Bruesch-Pagan test for homogeneity of variance among the residuals. Use a 5% level of significance (you can include the code in your answer to Question 59). Comment on the results of the Bruesch-Pagan test.

The check for homogeneity of variance looks not to be violated at a .05 level of significance, but just barely. The "status quo" of the Breusch-Pagan test assumes homoscedasticity.

### Part C

Fit the model.

Construct a 95% confidence interval for the mean gpa of all 7th-graders with Piers-Harris Children’s Self-Concept Score of 50, IQ of 105, and who are auditory learners.

#### Question 62 - Insert your R code here.

model3 <- lm(y ~ x1 + x2 + x3 + x1:x3, data = gpa7th)  
newdata3 <- data.frame(x1= 50, x2=105, x3=factor("Auditory"))  
(predict3 <- predict.lm(model3, newdata = newdata3, interval="confidence"))

## fit lwr upr  
## 1 2.997526 2.696134 3.298919

round(predict3[1,"lwr"],3)

## [1] 2.696

round(predict3[1,"upr"],3)

## [1] 3.299

#### Question 63 - Enter the lower bound of the 95% CI (round to 3 decimal places).

2.696

#### Question 64 - Enter the upper bound of the 95% CI (round to 3 decimal places).

3.299

#### Question 65 - Write an interpretation for the interval in the context of the problem.

With a 95% level of confidence, the mean GPA for children achieving a Self-Concept Score of 50, having IQ of 105, and who are auditory learners can expect a predicted range of values for their GPA to be between 2.696 and 3.299.

The prediction parameters are well within the ranges of the sample dataset - both individually and in conjunction - for all three predictive variables (two continuous and one categorical).

### Part D

Fit the model

#### Question 66 - Interpret the value of in the context of the problem.

For every unit change in IQ, GPA should be expected to increase by .033 providing that the other variables (Self-concept score and Auditory/Visual learner) remain constant.

### Part E

Fit the first-order logistic regression model to predict that a randomly selected student is an auditory learner from gpa, the Piers-Harris Children’s Self-Concept Score, and the IQ.

Use it to predict the probability of being an auditory learner for a 7th-grader with a 3.5 gpa, Piers-Harris Children’s Self-Concept Score of 50, and IQ of 105. Report the probability to 4 decimal places.

#### Question 67 - Insert your R code here.

gpa7th2 <- data.frame(gpa7th, isAuditory = gpa7th$x3 == "Auditory")  
model4 <- glm(isAuditory ~ y + x1 + x2, data = gpa7th2, family="binomial")  
newdata4 <- data.frame(y = 3.5, x1= 50, x2=105)  
round(predict(model4, newdata=newdata4, type="response"),4)

## 1   
## 0.7666

#### Question 68 - Report the probability to 4 decimal places.

.7666

## Problem 6

**Scenario**: A study was conducted to evaluate the quality of beef after storage times (STORAGE) of 10, 40, 80, and 120 days. The beef quality variables assessed were beefy aroma (BEEFY), bloody aroma (BLOODY), and a grassy aroma (GRASSY), which were all measured on a rating scale ranging from 0 to 15. Thirty samples were evaluated by several beef quality specialists and the average rating was obtained for this data file. You will need this file: **beef.rda**

### Part A

Use the Henze-Zirkler Multivariate Normality Test to test for multivariate normality among the three response variables: BEEFY, BLOODY, and GRASSY. Include a chi-square quantile plot in your analysis and use a 1% level of significance for each individual hypothesis test.

#### Question 69 - Insert your R code here.

require(MVN)  
require(mvoutlier)  
data("beef")  
  
# Multivariate tests over BEEFY, BLOODY, GRASSY  
hzTest(as.matrix(beef[, 2:4]))

## Henze-Zirkler's Multivariate Normality Test   
## ---------------------------------------------   
## data : as.matrix(beef[, 2:4])   
##   
## HZ : 0.7762501   
## p-value : 0.1296601   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------------

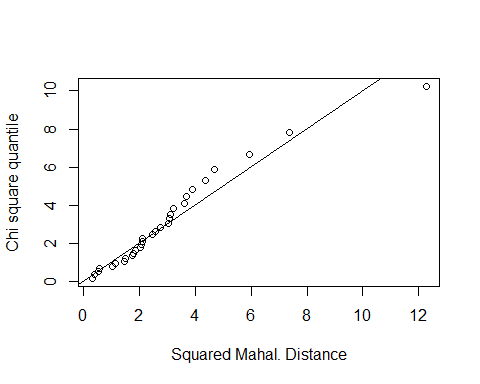
mardiaTest(as.matrix(beef[, 2:4]))

## Mardia's Multivariate Normality Test   
## ---------------------------------------   
## data : as.matrix(beef[, 2:4])   
##   
## g1p : 2.466879   
## chi.skew : 12.33439   
## p.value.skew : 0.2632999   
##   
## g2p : 14.90046   
## z.kurtosis : -0.04976883   
## p.value.kurt : 0.9603066   
##   
## chi.small.skew : 14.25772   
## p.value.small : 0.1615573   
##   
## Result : Data are multivariate normal.   
## ---------------------------------------

roystonTest(as.matrix(beef[, 2:4]))

## Royston's Multivariate Normality Test   
## ---------------------------------------------   
## data : as.matrix(beef[, 2:4])   
##   
## H : 5.324645   
## p-value : 0.01202528   
##   
## Result : Data are not multivariate normal.   
## ---------------------------------------------

means <- colMeans(beef[,2:4])  
covariances <- cov(beef[,2:4])  
dsquare <- mahalanobis(beef[,2:4], means, covariances)  
plot(sort(dsquare), qchisq(ppoints(nrow(beef)), df=3), ylab="Chi square quantile", xlab="Squared Mahal. Distance")  
abline(a=0, b=1)



#### Question 70 - According to this test, is there sufficient evidence to conclude that BEEFY, BLOODY, and GRASSY are not multivariate normal? Explain.

No. The Henze-Zinkler test for multivariate normality was within the .01 significance level (p ~ .130) for the trio of variables, suggesting normality was satisfied.

Also, the Mardia Test (p ~ .162) and Royston test (p ~ .012) suggested normality at a .01 level of significance, despite the verbiage from the Royston test to the contrary (it's assuming .05, I'm guessing...).

### Part B

Conduct Box's M Test to test for equality of covariances. Use a 1% level of significance.

#### Question 71 - Is there sufficient evidence to conclude that the covariance matrices are not equal at the 1% level of significance?

No. Do not reject the null hypothesis that the covariance matrices are equal.

With a p-value of .0255, there is not sufficient evidence to conclude that the covariance matrices are unequal at a .01 level of significance.

#### Question 72 - Insert your R code here.

source('./BoxMTest.R')  
BoxMTest(as.matrix(beef[, 2:4]), beef$STORAGE, alpha = .01)

##   
## ------------------------------------------------------------  
## MBox F df1 df2 P  
## ------------------------------------------------------------  
## 40.7426 1.7548 18 1901 0.0255  
## ------------------------------------------------------------  
## Covariance matrices are not significantly different.

## $MBox  
## 10   
## 40.74261   
##   
## $F  
## 10   
## 1.754844   
##   
## $df1  
## [1] 18  
##   
## $df2  
## [1] 1901  
##   
## $pValue  
## 10   
## 0.02545674

#### Question 73 - Based on the criteria of multivariate normality and equal covariance matrices, is it appropriate to proceed with MANOVA?

Yes.

### Part C

Regardless of the outcomes of the previous hypothesis tests, conduct a MANOVA to determine if there are differences between the different storage times for the population mean vectors when beefy, bloody, and grassy aromas are considered together. Use the Wilk's Lambda statistic and let

#### Question 74 - Insert your R code here.

lmodel <- lm(cbind(BEEFY, BLOODY, GRASSY) ~ STORAGE, data=beef)  
lmodel.out <- manova(lmodel)  
summary(lmodel.out, test="Wilks")

## Df Wilks approx F num Df den Df Pr(>F)   
## STORAGE 3 0.1269 8.6897 9 58.56 3.816e-08 \*\*\*  
## Residuals 26   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### Question 75 - State the null and alternative hypothesis.

The null hypothesis is that the means of beefy, bloody, and grassy aromas are the same between groups of storage times (10, 40, 80, and 120 days).

The alternative hypothesis is that the means of beefy, bloody, and grassy aromas differ between groups of storage times (10, 40, 80, and 120 days).

#### Question 76 - State the conclusion for the test.

At a significance level of .05, we can reject the null hypothesis. With a p-value approaching zero, there is sufficient evidence to state that the means of beefy, bloody, and grassy aromas differ between groups of storage times.

#### Question 77 - Round the p-value to four decimal places (enter 0 if P < 0.00005).

0

### Part D

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

#### Question 78 - Insert your R code here.

summary.aov(lmodel)

## Response BEEFY :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## STORAGE 3 322.55 107.516 37.192 1.491e-09 \*\*\*  
## Residuals 26 75.16 2.891   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response BLOODY :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## STORAGE 3 91.744 30.5814 33.948 3.837e-09 \*\*\*  
## Residuals 26 23.422 0.9008   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Response GRASSY :  
## Df Sum Sq Mean Sq F value Pr(>F)   
## STORAGE 3 51.727 17.2422 45.563 1.713e-10 \*\*\*  
## Residuals 26 9.839 0.3784   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#BEEFY  
onewayComp(BEEFY ~ STORAGE, data = beef, var.equal = TRUE)$comp[, c(6,7)]

## p adj rej H\_0  
## 40-10 5.668196e-01 0  
## 80-10 3.323837e-06 1  
## 120-10 2.685747e-08 1  
## 80-40 1.891019e-05 1  
## 120-40 7.728097e-08 1  
## 120-80 2.537189e-01 0

#BLOODY  
onewayComp(BLOODY ~ STORAGE, data = beef, var.equal = TRUE)$comp[, c(6,7)]

## p adj rej H\_0  
## 40-10 7.351889e-01 0  
## 80-10 9.169770e-06 1  
## 120-10 8.373159e-08 1  
## 80-40 2.785917e-05 1  
## 120-40 1.378152e-07 1  
## 120-80 2.907963e-01 0

#GRASSY  
onewayComp(GRASSY ~ STORAGE, data = beef, var.equal = TRUE)$comp[, c(6,7)]

## p adj rej H\_0  
## 40-10 2.090224e-01 0  
## 80-10 1.609428e-06 1  
## 120-10 1.063290e-09 1  
## 80-40 5.561082e-05 1  
## 120-40 1.027262e-08 1  
## 120-80 2.039515e-02 1

#### Question 79 - Write your conclusion.

The univariate ANOVA tests suggest - at a significance level of .05 - That each of the three variables (Beefy, Bloody, and Grassy aromas) have different means across groups of storage times (10, 40, 80, and 120 days). All three p-values approach zero.

The Tukey tests identify the following combinations as significant (having different means):

*For all three aroma types:*

* 10 as compared to 80 days.
* 10 as compared to 120 days.
* 40 as compared to 80 days.
* 40 as compared to 120 days.

*For GRASSY aromas only:*

* 80 as compared to 120 days.

#### Questions 80-86 - There are 7 more questions in the D2L version of the final. None of these questions require any R so they are not included in the RMD version of the final exam.