Final Exam

DS 705

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.

```
# you'll have to cut and paste this R code into D2L
```

Question 2 - State the null and alternative hypothesis

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

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

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.

```
# you'll have to cut and paste this R code into D2L
```

Question 6 - Report the overall MSA value.

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

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

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

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:

```
# you'll have to cut and paste this R code into D2L
```

Question 11 - Report the overall MSA value.

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

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

Question 14 - If so which one(s)? (if there aren't any say "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.

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

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

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.

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

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

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.

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.

Question 23 - How many clusters does this form?

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.

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.

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

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.

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.

Question 29 - State the null and alternative hypothesis.

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

Question 31 - State the degrees of freedom.

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

Question 33 - State your conclusion.

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.

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

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

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

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.

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

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

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.

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?

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.

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

- (a) Normality is rejected for the rents of both the 3-bedroom and 2-bedroom apartments.
- (b) Normality is rejected for the rents of the 2-bedroom apartments but not the 3-bedroom apartments.
- (c) Normality is rejected for the rents of the 3-bedroom apartments but not the 2-bedroom apartments.
- (d) Normality is not rejected for the rents of either the 3-bedroom or 2-bedroom apartments.

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.

Question 46 - State the null and alternative hypotheses.

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

Question 48 - State the conclusion for the test.

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.

Question 50 - State the hypotheses.

Question 51 - State the test statistic.

Question 52 - State the df.

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

Question 54 - State the conclusion.

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.

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

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon$$

Fit this model and report on which coefficients are significantly different from zero (create the interaction terms separately, do not use x1:x3 or x2:x3 in the code). Use the hierarchical approach to model-building.

Question 56 - Insert your R code here.

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

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

Part B

Fit the model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \epsilon$$

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.

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

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.

Part C

Fit the model.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \epsilon$$

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.

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

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

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

Part D

Fit the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \epsilon$$

Question 66 - Interpret the value of $\hat{\beta}_2$ in the context of the problem.

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.

Question 68 - Report the probability to 4 decimal places.

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.

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

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?

Question 72 - Insert your R code here.

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

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 $\alpha = .05$.

Question 74 - Insert your R code here.

Question 75 - State the null and alternative hypothesis.

Question 76 - State the conclusion for the test.

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

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.

Question 79 - Write your conclusion.

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.