Multiple Choice questions – 40 points

Answer key:

1. C
2. B
3. C
4. A
5. B
6. A
7. C
8. A
9. B
10. A
11. C
12. D
13. D
14. D
15. B
16. B
17. C
18. C
19. A
20. A

Scenario 1 – 30 points

A retail company wants to increase their sales. Therefore, they hire you- an expert in Multivariate Regression Analysis- to study important characteristics that can influence the amount of purchases customers make from their company. For the study, data was collected for 100 respondents on 5 separate variables. These variables are listed below:

X22 - Purchase Level
X2 - Industry Type
X10 - Advertising
X11 - Product Line
X13 - Competitive Pricing

Details of the variables are:
X22 is measured as the percentage of purchases from the company
X2 is a dummy variable indicating the type of industry that purchases the company's products (0=manufacturing; 1=services)
Each separate variable X10, X11, and X13 represents the perception of the retail company's performance for this specific attribute, which are considered to be most influential for the selection choices of the customers. Respondents could rate each of the 3 attributes on 0-10 scale with 10 being "Excellent" and 0 being "Poor".

Questions

Appendix A contains the SPSS output required for answering the questions of Scenario 1. When answering these questions, always explicitly mention the table, matrix or graph you used to provide the answer. When test results need to be provided, specify the complete test with correct hypotheses, test values, significance levels, and what you conclude from the test results (interpretation). If no significance level is provided, be sure to specify the level you will use. *Note that not mentioning these details means fewer points!*

Question 1:

Explain if Multivariate Regression Analysis is allowed for the given dataset.

The number of observations is 100, there are five variables (one dependent, four independent). The ratio of observations to variables is 20:1, so the threshold of five observations per variable is met. Furthermore, all variables are metric, and there are multiple independent and one dependent variable. Hence, MRA is allowed.

2 points

Question 2:

Are there any problems with missing data and outliers? Explain your answer.

There are no missing data, there are no problems here.
Based on the boxplots, there are no outliers. In the scatter plots, we see some bivariate outliers in the plots of Advertising – Product line and Product line – Purchase level. However, we conclude that there are no serious problems with outliers.

2 points

Question 3:

Discuss the assumption of normality in this data set. Use a significance level of α=0.05.

Based on the histograms and boxplots, it seems that Competitive Pricing is not normally distributed. The variables Advertising, Product Line and Purchase Level could be normally distributed.
Competitive Pricing has a negative skew and is peaked.
Advertising has two peaks, but it could be normal.
Product Line looks normally distributed.
Purchase level could be normally distributed, however it has several peaks.

Next, the Kolmogorov-Smirnov test is considered. The following hypothesis is tested
H0: The variable is normally distributed.
H1: The variable is not normally distributed.
A significance level of α=0.05 is used.
The test is not significant for Advertising, Product Line and Purchase Level.
Hence we conclude that Purchase Level, Advertising and Product Line are normally distributed.
Competitive pricing follows a non-normal distribution.
2 points

Question 4:

Test for the presence of heteroscedasticity for the variables Advertising, Product Line and Competitive Pricing. What do you conclude?

The Levene test is considered and the following hypothesis is tested
H0: Variance of variable is homogeneic.
H1: Variance of variable is heterogeneic.

A significance level of α=0.05 is used.

For Advertising and Product Line, the significance level is 0.261 and 0.828 respectively. Hence, we fail to reject the null hypothesis.
The significance level for Competitive Pricing is 0.029, which is lower than α=0.05. Hence, we reject the null hypothesis.

We conclude that Advertising and Product Line are homoscedastic and Competitive Pricing is heteroscedastic.

2 points

Questions 5-9 are based on Multivariate Regression Analysis 1.

Question 5:

Provide the regression equation for the regression model.

Total 2 pts

Model
$\hat{X}_{22} = 22.763 + 1.398X_{10}+4.621X_{11}+0.460X_{13}$
or
$X_{22} = 22.763 + 1.398X_{10}+4.621X_{11}+0.460X_{13} + e$

No ^ or e: -1 pt
Both ^ and e: -1 pt
constant forgotten: -1 pt
wrong coefficients: -1 pt

Question 6:

Determine the percentage of variation in the dependent variable that is explained by the regression model . Specify the test used, the hypothesis tested, and whether this percentage is significant.

Total 3 pts

(1 pt)
Look at the Model Summary table, find that the R^2 = 0.455. Therefore, this means that 45.5% of the variation in the dependent variable purchase level (percentage of purchases) is explained by the independent variables.

(1 pt)
The test to determine whether this percentage is significant is the F-test which is in the ANOVA table.
F statistic=26.726, p-value= .000 (close to 0).
The hypothesis tested is:
H0: $R^2 = 0$
H1: $R^2 > 0$

(or this can be more precisely formulated as H0: bi=0; H1: bi≠0 ....in other words all regression coefficients are equal to zero versus there is at least one regression coefficient that is unequal to zero).

(1 pt)
Interpretation of findings:
p-value<alpha, reject H0. Conclude that the R^2 is significantly different from zero.

Question 7:

Explain which independent variables have a significant contribution in the prediction of the dependent variable in the regression model. Use a 5% significance level for your test.

Total 3 pts
(1 pt)
H0: bi=0
H1: bi≠0
alpha=5%

(1 pt)
t-test , look at coefficients table.
X10: t-value=2.334, p-value=0.022
X11: t-value=7.895, p-value=0.000
X13: t-value=0.914, p-value=0.363

(1 pt)
Interpretation of findings
X10: p<alpha, reject H0 X11: p<alpha, reject H0
X13: p>alpha, fail to reject H0
Conclusion: X10 and X11 are significantly different from 0 (i.e., these two independent variables have a significant contribution in the prediction of the dependent variable).

Question 8:

Indicate and explain which independent variable has the highest influence on the dependent variable of the regression equation.

Total 2 pts

(Coefficients table)
The independent variable that has the highest influence on the dependent variable is X11 since it has the highest standardized regression coefficient (0.686).

Question 9:

Explain the difference between zero- order correlation and partial correlation.

Total 3 points

The zero-order correlations are the normal correlations and are a combination of the unique correlation of a specific independent variable and the correlation of other variables with that particular independent variable. (-1 pt)

The partial correlation is the correlation that is uniquely determined by a specific independent variable X excluding correlations from any of the independent variables on that specific independent variable X and on the dependent variable. (-1 pt)

Hence, the difference between zero correlation and partial correlation is that in the partial correlation the correlations from any of the independent variables on that specific independent variable X and on the dependent variable are excluded. (-1 pt)

Questions 10-11 are based on Multivariate Regression Analysis 2.

Question 10:

Provide the regression equation of the regression model.

Total 2 points

$\hat{X}_{22} = 23.644 + 1.382X_{10} + 4.582X_{11} + 0.586\,X_{13} - 3.069\,X_2$
or
$X_{22} = .644 + 1.382X_{10} + 4.582X_{11} + 0.586\,X_{13} - 3.069\,X_2 + e$

No ^ or e: -1 pt
Both ^ and e: -1 pt
constant forgotten: -1 pt
wrong coefficients: -1 pt

Question 11:

How would you interpret the influence of the dummy variable?

Total 3 points

If the type of industry that purchases HBAT's paper products is the magazine industry, the dummy variable X_2 will attain the value 0. And if the type of industry is the newsprint industry, X_2 will attain the value 1. (-1 pt)

The coefficient of X_2, -3.069, is the difference in the percentage of purchases from HBAT between the newsprint industry and the magazine industry. (-1 pt) While keeping all other independent variables and the error term constant. (-1 pt)

Question 12

Explain which model you would select for predicting the dependent variable X_{22}. (Use multivariate analysis 1 and 2 results presented in Scenario ?). Indicate exactly which model you would select.

total 3 points

The adjusted R^2 is the relevant indicator to look at. This indicates the amount of variance explained of the dependent variable by the independent variables, corrected for the amount of independent variables in the equation. (-1 pt)

Table for model summary of scenario 1, multivariate regression analysis 1, indicates an adjusted R^2 of .438, while the adjusted R^2 is .463 from multivariate regression analysis 2. (-1 pt)

Since the adjusted R^2 of the model from multivariate regression analysis 2 is highest, this model is selected. (-1 pt)

Scenario 2 – 30 points

Question 1: 5pts

a) What is factor analysis and what is the goal of factor analysis? 2pts

b) What is the difference between principal component analysis and common factor analysis and when do you use which method? 2 pts

c) What is the difference between R and Q-type factor analysis? 1pt

a) **Factor analysis is an interdependence technique, whose primary purpose is to define the underlying structure among the variables in the analysis. 1pt**
The goal is to reduce or summarize the data. 1pt

b) **1) Principal Component Analysis (PCA), which is used to reduce the dimensionality of data. The total variance is redistributed in p observed variables over p principal components. The first principal component has largest contribution to total variance. The second has the second largest contribution, etc. (1 point)**

2) Common factor analysis or PFA, which is used to summarize/explain the data. The method reproduces observed correlations as good as possible, using small number of common factors. Common factor analysis considers only the common variance among variables. The observed relations between the variables are describing underlying constructs (i.e. the common factors), which may serve further as theoretical deepening. (1 point)

c) **R factor analysis analyzes a set of variables to identify the dimensions that are latent, whereas Q factor analysis analyzes individual cases to group respondents together. 1pt**

An important tool for many retail firms is training of its sales force, for which several different techniques are available. In order to gain more insight into the effectiveness of certain types of training techniques, a sales manager wishes to reduce the data he has to be used in further multivariate analysis. The sales training data were collected via a mail questionnaire which was sent to 80 sales training managers at various firms throughout the United States. The questionnaire addresses the usage of various training methods by the respondent's firm. There are 9 methods in total. The respondent rates all methods on a five point scale for present frequency, that is, for how often each technique is being used at present.

A1 'CONF/DISCUSSION-PRESENT'
A2 'LECTURE METHOD-PRESENT'
A3 'CASE STUDY-PRESENT'
A4 'TV-LECTURE-PRESENT'
A5 'FILM VIEWING-PRESENT'
A6 'VIDEO TAPE/DISC-PRESENT'
A7 'INTERACTIVE VIDEO-PRESENT'
A8 'ROLE PLAY:VIDEO TAPE-PRESENT'
A9 'BUSINESS GAMES-PRESENT'

Questions
You have to answer a couple of questions. For some of the questions, you have to check the SPSS output given in Appendix B. When answering to these questions, always mention explicitly which table, matrix or graph you used to provide the answer (not mentioning this means fewer points!).

Question 2: **8pts**

a) Is factor analysis allowed on this dataset? Motivate your answer. 1pt
b) Intercorrelation is an important statistical assumption that has to be met for factor analysis. Describe three possible measures to test this assumption. 3pts
c) Based on the SPSS output in Appendix B, do the data in factor analysis 1 meet the required assumptions? If not, what remedy would you propose? 4pts

a) Only metric values and a ratio of observations per variable of at least 8 : 1 which is greater than 5 : 1 -> FA allowed (1 point)

b + c) i) anti-image correlation matrix
- matrix with (negative) partial correlation, correlation that is unexplained when the effects of other variables are taken into account (1 point)
- all partial correlations are small < 0.7 (1 point)
ii) Barlett's test of sphericity, which test for the present of correlations among the variables. It provides the statistical significance that the correlation matrix has significant correlations among at least some of the variables. (1 point)
- test is significant, p-value=.000, assumption met (1 point)
iii) Measure of Sampling Adequacy (MSA), which is 1 if the variable is perfectly predicted without error by the other variables (1 point)
- overall Kaiser-Meyer-Olkin MSA = .650 > .5

- However, variable specific MSA's: only INTERACTIVE VIDEO not >0.50, assumption not met(1 point)
Remedy: delete this variable and redo the factor analysis after excluding this variable (1point)

Question 3: **9pts**

a) Give three different criteria to determine how many factors should be extracted. 3pts
b) Based on the SPSS output in Appendix B for factor analysis 2, how many factors should be extracted and based on what criterion? 1pt
c) After obtaining the factor solution, the researcher has to evaluate the factor solution. In which three cases is respecification of the model needed? 3pts
d) Does the unrotated factor solution in factor analysis 2 provide a good factor solution? Motivate your answer . 2pts

a) **1) Latent root criterion** **1pt**
2) Total variance explained criterion **1pt**
3) Scree test criterion **1pt**
(4 A priori criterion)

b) **3 factors based on the latent root criterion** **1pt**

c) **1) no significant factor loadings**
1pt
2) cross-loadings **1pt**
3) communalities for variables <0.5 **1pt**

d) **No, cross-loadings > 0.4 for TV-LECTURE, VIDEOTAPE/DISC, ROLEPLAY:VIDEO TAPE and BUSINESS GAMES** **1pt**
Communality for FILM VIEWING <0.5 **1pt**

Question 4: **2pts**

a) What is factor rotation? Use the SPSS output in factor analysis 3 to explain the concept. 2pt

a) **Rotation is a method of redistributing variance from factors previously obtained (the unrotated solution). It is used in the interpretation phase, to achieve a simpler, theoretically more meaningful factors pattern. (1 point)**
In the SPSS output, it can be inferred that the variance explained by the first factor is lower after rotation, whereas factors 2 and 3 now explain more of the total variance. That indicates that variance has been redistributed from earlier factors to later factors. (1 point)

Question 5: **6pts**

a) Consider the SPSS output in Appendix B for factor analysis 3. Does the orthogonal factor rotation solution provide a good factor solution? Motivate your answer. 1pt

b) Does the oblique factor rotation give a satisfactory solution? Motivate your answer. 1pt

c) Based on the SPSS and the goal of the researcher, which factor solution would be preferred in this case? 2pts

d) If no satisfactory solutions are found, what remedies would suggest? 2pts

a) **No, still cross-loading for TV LECTURE and communality for FILM VIEWING <0.5** **1pt**

b) **No, from pattern matrix: still cross-loading for TV LECTURE and communality for FILM VIEWING <0.5**
1pt

c) **The component correlation matrix indicates low correlation among the factors and considering that the researcher intends to use the reduced data in further multivariate analysis, the orthogonal rotation method is preferred, which creates uncorrelated factors.**
2pts

d) **Possible remedies are:** **1pt per remedy**

e) **1) Delete the FILM VIEWING variable**
2) Still no good solution? Try deleting the TV LECTURE variable
3) Extract different number of factors
4)Different extraction method, ie common factors