Statistics II for IB&M
PRACTICE EXAM
Suggested Answers

Scenario 1

NOC=number of companions
LOTH= length of thorax, in mm
POEDSS= percentage of each day spent sleeping

Question 1:
Explain if Multivariate Regression Analysis is allowed for the given dataset. Indicate exactly which tables you use in your answer.

Answer 1: Ratio of observations per variable is 31.25/1 (number of variables =4, n=125); the threshold of 5:1 is met. Descriptive table on page 5 is used.
Furthermore, all variables are metric; and there are several multiple independent and one dependent variable.

Question 2:
Use the SPSS output in section Scenario 1 - Tables Multivariate Regression Analysis 1, to correctly finish the text underneath. Unless specified differently, assume tests are two-tailed with α = 0.05.

Answer 2:
44 % of the variance of the dependent variable is explained by the three independent variables. This percentage is **significant** / ~~is not significant~~, indicated by the **F(ANOVA)-test**. The null-hypothesis for this test is **$R^2 = 0$** The alternative hypothesis is **$R^2 > 0$**. The test value is **31.726**, with significance level **0.000**. This significance level is ~~higher~~ / **lower** than α. We **reject** / ~~fail~~ to reject the null-hypothesis.
The independent variable **length of thorax** has the highest influence on the dependent variable. This is indicated by the **Beta (standardized coefficients)** which has value **0.603.**

Question 3:
Explain which independent variables have a significant contribution in the prediction of the dependent variable.

Answer 3:
H0: bi = 0
H1: bi ≠ 0
α = 0.05

From coefficients table, page 7
NOC: t=-2.701 p=0.008,
LOTH: t=8.679 p=0.000
POEDSS: t= -0.571 p=0.569

Interpretation:
NOC: p < alpha , so reject H0 and assume H1 to be valid
LOTH: p < alpha, so reject H0 and assume H1 to be valid
POEDSS: p > alpha, so do not reject H0
Hence, NOC and LOTH have significant contribution in the prediction of the dependent variable.

Question 4:
Explain which independent variables have a unique, significant contribution to the prediction of the dependent variable. Indicate exactly which table you use in your explanation.

Answer 4:
In forward method only variables, which have a unique, significant contribution to predicting the behavior of the dependent variable, are added. Hence, all of them have a unique, significant contribution in the prediction of the dependent variable.

H0: $b_i = 0$
H1: $b_i \neq 0$
$\alpha = 0.05$

From coefficients table, page 8:
Model 1: LOTH: t=9.152 p=0.000.
Model 2: NOC: t=-2.705 p=0.008,
LOTH: t=8.684 p=0.000.

Interpretation:
NOC: p < alpha , so reject H0 and assume H1 to be valid
LOTH: p < alpha, so reject H0 and assume H1 to be valid
Hence, NOC and LOTH have significant contribution in the prediction of the dependent variable.

Question 5:
Does multicollinearity cause a problem in the regression analysis that is presented in the SPSS output in section Scenario 1 – Multivariate Regression Analysis 2 (page 8)? Explain your answer.

Answer 5:
From coefficients table, page 8:
Model 2: NOC: tolerance = 0.963, VIF=1.093
LOTH: tolerance = 0.963, VIF=1.093.

Since, VIF is <10, there are no problems with multicollinearity.

Question 6:
Explain which model you would select for predicting the dependent variable LIFESPAN. (Use analysis results presented in sections Scenario 1 - Tables Multivariate Regression Analysis 1 (pages 5-7), 2 (page 8), and 3 (page 9)). Indicate exactly which model from which table you select.

Answer 6:
Adjusted R-squares of all models:
Enter method (model summary table, page 7) R^2= 42.6%
Forward method (model summary table, page 8) model 1, R^2=40%

Forward method (model summary table, page 8) model 2, R^2=43%
Enter method with dummy variables (model summary, page 9), R^2=61.1%

The last model (enter method with dummy variables) is selected, since the adjusted R-square is the highest and the objective of the study is explanation. When explanation is the objective, the model that is rich, and contains all variables derived from theory should be chosen.

Scenario 2

Question 1:
In the scenario presented above, is Principal Component Analysis the correct extraction method? Motivate your answer.

Answer 1:
No, because the company wants to get "a good understanding" of the factors that would influence their business strategy. PCA is used for data reduction, while PFA is used for data explanation. Explanation is what the company wants, therefore it would be better to use PFA.

Question 2:
The assumptions required for performing factor analysis have been tested by the research team. Based on the SPSS output, fill in the empty spaces, or correct the underlined text (by stripping out the incorrect answer).

Answer 2:
Testing assumptions means checking **conceptual** and statistical assumptions.
The statistical assumptions to be checked in case of factor analysis are:

1. **Outliers**
2. **Linearity**
3. **Normality**
4. **Homoscedasticity**
5. Intercollinearity.

Intercollinearity is checked, using four measures. First, a sufficient amount of intercorrelations should exist among variables, with correlations above a value of **0.3.**
Second, the **anti-image** matrix should be inspected; this matrix contains **negative values of partial correlations**, which are those correlations unexplained when the effects of the other variables are considered. Third, the Bartlett test of sphericity should be significant. Fourth, the Measure of Sampling Adequacy should be checked, providing the degree of intercorellation, whose values below value of **0.5** are unacceptable.

The conclusion is that ~~there are~~ / **there are no** problems with respect to intercollinearity, based on the following arguments:

- Correlation Matrix page 15: Sufficient correlation (substantial number of correlations), with correlations above the threshold value of 0,3 and significance at .01 level
- Anti-image correlation matrix page 16: all partial correlations < 0.7
- KMO and Bartlett's Test table page 15 :
  H0: correlation matrix **R** = identity matrix **I**

H1: correlation matrix **R** is not equal to the identity matrix **I**
Bartlett's test of sphericity is significant (test value =254.344 , p-value=.000), rjection of H0.

- KMO and Bartlett's Test table page 15 : (KMO) Measure of Sampling Adequacy = .704 and
  Anti-image correlation matrix page 16 : MSA of all variables > .5, thus there are no candidates for exclusion .

Question 3:
How many factors should be extracted? Motivate your answer.

Answer 3:

- Total variance explained table page 16:
  Based on latent root (Kaiser criterion) => 3 factors have eigenvalues > 1
  Percentage of variance criterion Variance > 60% => 3 factors (~62%)
- Scree test - three factors can be extracted
  Hence, three factors

Question 4:
a) What does the `communality of a variable' mean?
b) Assess the communalities of all variables.
c) Why does the "Initial" column of the Communalities table contain only the value "1"?

Answer 4:

a) Communality is amount of variance in each variable that is accounted for as represented in the factor solution.
b) From table Communalities page 15: all communalities >0.5 (half of the variance needs to be taken into account), so all variables are sufficiently explained jointly by factors derived in factor solution.
c) In PCA analysis, the total variance is being used (common, specific and error); initial column has values which are estimates of variance in each variable which is accounted for by all components

Question 5:
Does the un-rotated solution provide a good factor solution? Motivate your answer.

Answer 5:
No, it does not, there is number of cross loadings (Component matrix page 17), so interpretation is difficult.

Question 6:
a) What is a rotation and when should rotation be used?
b) Describe the two types of possible rotation methods.
Use the SPSS output.
c) Does the solution provided by Varimax approach resulted in a good factor solution? Motivate your answer.
d) Which rotation method (Varimax or Oblimin) is the most suitable, and provides the best factor solution? Motivate your answer (specify which matrices you use).

Answer 6:

a) Rotation is redistribution of variance, the reference axes of the factors are turned about the origin until some other position has been reached. Is used to achieve simpler and theoretically more meaningful factor solutions.
b) Orthogonal rotation (Varimax), in which the axes are maintained at 90 degrees, when factors are not correlated.
Oblique rotation (Oblimin), in which the axes are rotated, and 90 degrees are not retained between the reference axes; factors can be correlated.

c) Varimax rotation gives a good factor solution, because there is no cross-loadings (rotated component matrix page 17)
d) Varimax and Oblimin both show no cross-loadings, results are comparable. The Varimax is chosen, since it is easier to interpret

Question 7:
Using the best factor solution, label and interpret the found factors, in the context of the case described in Scenario 2.

Answer 7:
Factor 1: commodity prices, exchange rates, internal regulations and tax system => "Economic indicators".
Factor 2: cultural regulations and impact local communities => "Local culture"
Factor 3: PC Ownership and access to bandwidth => "Technological development".

Scenario 3

Question 1:
The researcher starts his analysis using a Factor Analysis (see Scenario 3 Tables Factor Analysis 1, pages 21-24).
a) Given the research objective stated above, and the results that are shown in the tables, explain why the researcher uses factor analysis.
b) Is Factor Analysis allowed? Motivate your answer.

Answer 1:
a) The factor analysis is done because the researcher wants to condense the information contained in original variables into smaller set of factors, the simplification or data reduction is the purpose, so PCA is used.
b) Ratio of observations per variable is 4.54/1 (number of variables =11, n=50); the threshold of 5:1 is not met; factor analysis is not allowed.

Question 2:
Considering the outcome of the first Factor Analysis, the researcher decides to do a second Factor Analysis (see Scenario 3 - Tables Factor Analysis 2, pages 25-27).
Explain why the researcher executes a second Factor Analysis.

Answer 2:
One of the communalities was low, for variable "I don't like to see children's toys lying about", which was 0.272 and in rotated solution had insufficient loadings. So, the model needed to be re-specified.

Question 3:
The researcher considers the results from the second Factor Analysis to proceed with the analysis (section Scenario 3 – Tables Factor Analysis 2, pages 25-27). Do you agree with this decision? Explain your position.

Answer 3:
Yes, since in the second factor analysis (after the model is re-specified and variable *"I don't like to see children's toys lying about"* is ignored) with Varimax rotation, the factor solution derived is good. All the communalities >0.5 (from table communalities page 25) and there are no cross loadings (rotated component matrix page 27).

Question 4:
Consider the factor solution from Factor Analysis 2 (see section Scenario 3 – Tables Factor Analysis 2, pages 25-27). How would you label the three factors that were found in this solution? Explain why.

Answer 4:
Factor 1: Bargain searcher, because all these variables are related to people looking for bargains, sales, specials, low prices or being aware of prices in (grocery) shops.
Factor 2: Importance of nutrition, because all the statements are about the importance of nutrition, and that people are concerned about it.
Factor 3: Unimportance of Nutrition, because these variables shows that people are not concerned about nutritious food and it is not important for them.

Question 5:
The next analysis the researcher applies is Multivariate Regression Analysis (see section Scenario 3 - Tables Multivariate Regression Analysis, pages 27-28). He builds a regression model using the factors extracted in Factor Analysis 2 and the remaining variables from the questionnaire. Provide the regression equation for the regression model.

Answer 5:
%Advertisêment appraisal= 76,320 + 2,971*"REGR factor score 1" + 2,595*"REGR factor score 2" + 1,867*"REGR factor score 3" + 2,420*I don't like to see children's toys lying about" + e

Question 6:
Indicate how good the regression model is that the research found, and if the model is significant.

Answer 6:
R^2 =34.5% (Model summary table page 28)
H0: $R^2 = 0$
H1: $R^2 > 0$
$\alpha = 0.05$,
p = 0.001 (from ANOVA table page 28)
p < alpha , so reject H0 and assume H1 to be valid
This percentage is significant.

Question 7:
Explain which independent variables have a significant contribution to the explanation of the dependent variable y 'advertisement appraisal'.

Answer 7:
H0: bi = 0
H1: bi ≠ 0
α = 0.05

From coefficients table, page 28:
"REGR factor score 1": t=2.249 p= 0.029
"REGR factor score 2": t=1.994 p=0.052
"REGR factor score 3": t=1.477 p=0.147
I don't like to see children's toys lying about: t=2.069 p=0.044

Interpretation:
"REGR factor score 1": p < alpha , so reject H0 and assume H1 to be valid
"REGR factor score 2": p > alpha, so do not reject H0
"REGR factor score 3": p > alpha, so do not reject H0
"I don't like to see children's toys lying about": p < alpha , so reject H0 and assume H1 to be valid

Hence, "REGR factor score 1" and "I don't like to see children's toys lying about" have significant contribution to the explanation of the dependent variable.

Question 8:
Does multicollinearity provide problems in the found regression model? Motivate your answer.

Answer 8:
From coefficients table, page 28:
"REGR factor score 1": tolerance = 0.900, VIF=1.111
"REGR factor score 2": tolerance = 0.928, VIF=1.078
"REGR factor score 3": tolerance = 0.984, VIF=1.016
"I don't like to see children's toys lying about": tolerance = 0.830, VIF=1.205

Since, VIF is <10, there are no problems with multicollinearity.