Scenario 1 *(maximum 40 points)*

Eagle Farm is a famous horse racing venue in Brisbane, Australia. The study was conducted to find out what influences the finishing positions of the horses in the race. The study focusses on a sequence of 8 races; and the following data was collected: the number of horses in a race, the results of the last race, the number of days since the last race, the weight carried by a horse, the identifying number of a horse in a race, and the age of a horse in years. The Multivariate Regression analysis was used. The results are presented in Scenario 1 – Tables Multivariate Regression Analysis 1

Question 1

a) What are outliers?
b) What are the causes of outliers?

a) Observations that are substantially different from the rest of the observations on one or more characteristics (variables). (1 point)

b) Procedural error (1 point)
Extraordinary event or extraordinary observation (1 point)
Unique in combination (1 point)

Question 2:

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

Ratio of observations per variable is 14.57/1 (number of variables =7, n=102); the threshold of 5:1 is met. Descriptives table is used. (1 point)
Furthermore, all variables are metric; and there are several multiple independent and one dependent variable. (1 point)

(continued on next page)

Question 3:

a) Indicate which assumptions should be checked for Multivariate Regression Analysis and define each of them.

b) Test all the variables for normality. Use 5% level of significance. Indicate exactly which tables/graphs you use in your answer.

c) In case you use 1% level of significance, would your answers regarding normality change for any of the variables tested at point b?

a) *normality*, the variables involved should be normally distributed (1 point)
linearity, the relationship between the dependent and each of the independent variables should be liner, in other words the regression equation should be linear (1 point)
homoscedasticity, the variance of the dependent variable should not depend on any of the independent variables (1 point)
multicollinearity, there should be no linear relationship among the independent variables (1 point)

b) Normality Check statistically using Kolmogorov-Smirnov Test:
H0: Variable is normally distributed
H1: variable is not normally distributed (1 point)
Alpha=0.05
Finishing position: statistics=0.092, p=0.034; p<alpha, so reject H0 (1 point)
NumberOfHorsesInRace: statistic=0.255, p=0.000; p<alpha, so reject H0 (1 point)
FinishingPositionInLastRace: statistic=0.156, p=0.000; p<alpha, so reject H0 (1 point)
DaysSinceLastRace: statistic=0.347, p=0.000; p<alpha, so reject H0 (1 point)
IdentifyingNumberOfHorseInRace: statistic=0.100, p=0.014; p<alpha, so reject H0 (1 point)
WeightCarried: statistic=0.241, p=0.000; p<alpha, so reject H0 (1 point)
AgeOfHorseInYears: statistic=0.143, p=0.000; p<alpha, so reject H0 (1 point)

c) The answers would change for the following variables
Finishing position
IdentifyingNumberOfHorseInRace
For both variables the p-value is higher than 0.01. (1 point)

(continued on next page)

Question 4:

Use the output of scenario 1-Multivariate Regression Analysis 1, to correctly finish (add the missing text or strip out the incorrect option) the text underneath. Unless specified differently, assume tests are two-tailed with $\alpha = 0.05$.

36.3 % of the variance of the dependent variable is explained by the six independent variables. This percentage is significant/~~not significant~~, indicated by the **F**-test. The null-hypothesis for this test is **R^2 = 0**. The alternative hypothesis is $\mathbf{R^2 > 0}$.The test value is **9.008**, with a significance level **0.000**. This significance level is ~~higher~~/lower than α. We reject/~~fail to reject~~ the null-hypothesis. (9 points)

The regression equation from the regression model in Multivariate Regression Analysis 1section is:

%FinishingPosition= 4.591 + 0.257 NumberOfHorsesInRace + 0.293*FinishingPositionInLastRace + 0.012* DaysSinceLastRace + 0.325* IdentifyingNumberOfHorseInRace - 0.143* WeightCarried+ 0.463* AgeOfHorseInYears+ e

(1 point)

In case the constant term is missing subtract 0.5 points. In case the error term is missing subtract 0.5 points.

Question 5:

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

H0: bi = 0
H1: bi ≠ 0 (1 point)
α = 0.05
NumberOfHorsesInRace: p>alpha, so we do not reject H0
FinishingPositionInLastRace: p<alpha, so reject H0
DaysSinceLastRace: p>alpha, so we do not reject H0
IdentifyingNumberOfHorseInRace: p<alpha, so reject H0
WeightCarried: p>alpha, so we do not reject H0
AgeOfHorseInYears: p>alpha, so reject H0
Therefore, two variables FinishingPositionInLastRace and IdentifyingNumberOfHorseInRace appear to be significant. (2 points)

(continued on next page)

A second Multivariate Regression Analysis was performed, the results can be found in Scenario 1 – Tables Multivariate Regression Analysis 2. Use the results from in Scenario 1 – Tables Multivariate Regression Analysis 2 to answer the following questions.

Question 6:

Indicate which model selection method was used in Multivariate Regression Analysis 2, provide the definition of this method, and write the regression equation of the final model.

In the second part of scenario 1, the *forward method* was used. (1 point)
With forward method the model is built only with independent variables, which have a unique, significant contribution to predicting of the dependent variable. Forward method starts with no variables in the model and adds one variable at a time based on its contribution to prediction. (1 point)
In the model built, final model consists of IdentifyingNumberOfHorseInRace, FinishingPositionInLastRace and NumberOfHorsesInRace.
FinishingPosition= -0.642 + 0.264 NumberOfHorsesInRace +
0.327*FinishingPositionInLastRace + 0. 353* IdentifyingNumberOfHorseInRace + e
(1 point)

Question 7:

Indicate which independent variable has the highest influence on the dependent variable in the final model (Model 3) of the regression analysis. Indicate exactly which tables you use in your answer.

The standardized coefficient column (Beta column) is used.
The variable IdentifyingNumberOfHorseInRace appears to have the highest standardised coefficient B=0.381. (1 point)

(continued on next page)

Question 8:

Explain which independent variables have significant contribution to the prediction of the dependent variable in Model 3.

In forward method only variables, which have a unique, significant contribution to predicting the behaviour of the dependent variable, are added. Hence, all of them have a unique, significant contribution in the prediction of the dependent variable.
H0: bi = 0
H1: bi ≠ 0 (1 point)
α = 0.05
Model 3: IdentifyingNumberOfHorseInRace: t=4.204 p=0.000,
FinishingPositionInLastRace: t=3.553 p=0.001,
NumberOfHorsesInRace: t=2.152 p=0.034,
Interpretation:
IdentifyingNumberOfHorseInRace: p < alpha , so reject H0
FinishingPositionInLastRace: p < alpha, so reject H0
NumberOfHorsesInRace: p < alpha, so reject H0
Hence, all variables have significant contribution in the prediction of the dependent variable.
(2 points)

Question 9

In Multivariate Regression Analysis 2 three models are estimated. Explain which model you would select for predicting the dependent variable FinishingPosition. Explain which criterion you use.

Adjusted R-squared can be used for the comparison of the three models. Model 3 appears to be comparatively the best model, as it gives the highest adjusted R-squared value.
(1 point)

(continued on next page)

Scenario 2 *(maximum 35 points)*

According to current studies, interest in ethical attitudes of business students who are likely to be future business leaders is on the increase (Borkowski and Ugras, 1998). This is probably due to the major corporate scandals that started coming to light in the late 1990s and the first decade of the 21st century. In order to better understand ethical orientation among future business leaders, an exploratory study employing Principal Component Analysis was conducted among 200 business students. The study intended to identify two evaluative dimensions, idealism vs. relativism, based on Forsyth's model.

Forsyth (1980) developed two dimensions, idealism and relativism, to classify an individual's ethical and moral judgments. Idealism refers to the degree to which an individual believes that the right decision can be made in an ethically questionable situation. Idealistic individuals believe that there is a morally correct alternative that will not harm others. Less idealistic individuals may make decisions irrespective of the impact on others. Relativism, on the other hand, refers to the rejection of universal rules in making ethical judgments and focuses on the social consequences of behaviour. High relativists evaluate the current situation and use this as the basis for making a judgment. Low relativists believe that standard rules can be applied regardless of the issue at hand.
Forsyth's dimensions are based on 20 items. In the present study a selection of 9 items is used, see below:

1. People should make certain that their actions never intentionally harm others even to a small degree.
2. Risks to another should never be tolerated, irrespective of how small the risks might be.
3. The existence of potential harm to others is always wrong, irrespective of the benefits to be gained.
4. One should never psychologically or physically harm another person.
5. One should not perform an action which might in any way threaten the dignity and welfare of another individual.
13. Moral standards should be seen as being individualistic; what one person considers to be moral may be judged to be immoral by another person.
14. Different types of moralities cannot be compared as to "rightness".
15. Questions of what is ethical for everyone can never be resolved since what is moral or immoral is up to the individual.
17. Ethical considerations in interpersonal relations are so complex that individuals should be allowed to formulate their own individual codes.

References:

Borkowski, S. C. & Ugras, Y. J. 1998. Business students and ethics: A meta-analysis. *Journal of Business Ethics*, 17 (8), 117-127.

Questions

You have to answer a couple of questions. For most of the questions, you have to check the SPSS output given in the appendix. 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!).

(continued on next page)

Question 1:

a) Provide the definition of factor analysis.
b) What is the difference between Principal Component Analysis – PCA (also called Component Analysis) and Common Factor Analysis (PFA)?

a) Factor analysis is an interdependence technique, whose primary purpose is to define the underlying structure among the variables in the analysis. (1 point)

FA can be used either for data summarization or data simplification/reduction. (1 point)

b) PCA 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)

PFA is used to summarize/explain the data. The method reproduces observed correlations as good as possible, using small number of common factors ($m<<p$). The observed relations between the variables are describing underlying constructs (i.e. the common factors), which may serve further as theoretical deepening. (1 point)

Question 2:

Intercollinearity is one of the statistical assumptions to be checked in case of factor analysis, and is checked using four measures. Describe each measure, based on the SPSS output (see Appendix Scenario 2, Section questions 2 – 4).

total 8 points

i) sufficient amount of correlation among variables, $|r| > 0.3$ (1 point)

15 out of 36 above 0.3 (1 point)

ii) anti-image correlation matrix
- matrix with (negative) partial correlation (1 point)
- all partial correlations are small < 0.7 (1 point)

iii) Barlett's test of sphericity (1 point)
- test is significant, chi-square value = 368.828, p-value=.000 (1 point)

iv) Measure of Sampling Adequacy (MSA)
- overall Kaiser-Meyer-Olkin MSA = .731 > .5 (1 point)
- variable specific MSA's: all above .5 (1 point)

(continued on next page)

Question 3:

Based on the Scenario 2 description, and SPSS output, how many factors should be extracted (see Scenario 2, Section questions 2 – 4)? Motivate your answer.

Based on the Total variance explained, 55,88% of variance will be explained by 2 factors model, and 65,90% by 3 factors models, thus 3 factors would explain more variance. (1 point)

Based on Kaiser criterion, the 2 factors solution provide a solution above the threshold of 1. (1 point)

Based on the scree-plot, the 3 factors solution qualifies, because the point where the curve first begins to straighten is after 3 factors. (1 point)

However, corroborating with the goal of the study, one would only be interested in extracting 2 factors (idealism/relativism). (1 point)

Question 4:

a) The total variance of a variable can be divided into three types of variance. Describe each variance type.

b) Based on the SPSS output, assess the communalities of all variables (see Scenario 2, Section questions 2 – 4).

a)

(i) *common variance* is the variance in a variable that is shared with all other variables in the analysis. This variance is accounted for based on a variable's correlations with all other variables in analysis. (1 point)

(ii) *specific variance* (unique variance) is that variance associated with only a specific variable. This variance cannot be explained by the correlations to the other variables but is still associated uniquely with a single variable. (1 point)

(iii) *error variance* is variance that cannot be explained by correlations with other variables, but is due to unreliability in the data-gathering process, measurement error, or a random component in the measured phenomenon. (1 point)

b) All variables have sufficient communality values, above the threshold of 0.5. (1 point)

(continued on next page)

Question 5:

What is rotation and what is the difference between an Orthogonal and an Oblique rotation? Give examples of both oblique and orthogonal methods.

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)

Orthogonal rotation methods: produce factors which are uncorrelated. This is done either by simplifying rows, thus making as many values in each row as close to zero as possible (e.g. maximizing a variable's loading on a single factor), or by simplifying columns, thus by making as many values in each column as close to zero as possible (e.g. making the number of high loadings as few as possible). (1 point)

Oblique rotation methods: produce factors which are correlated. Similar to orthogonal rotations. (1 point)

Varimax rotation is an orthogonal method, thus factors remains uncorrelated through rotation process. Oblimin is an oblique approach, thus factors resulting are correlated. (1 point)

Question 6:

Do the un-rotated 2 factors and 3 factors solutions provide a good factor solution (see Appendix Scenario 2, Section questions 6 – 8)? Motivate your answer for each case.

2 factors solution

Yes, it is a good solution, no cross-loadings, each factor has loadings with values above the threshold of 0.4 (sample size 200, alpha .05, power level 80%). (1 point)

3 factors solution

Not such a good solution, one cross-loading for Q4. (1 point)

(continued on next page)

Question 7:

a) Do the rotated 2 factors and 3 factors solutions provided by Varimax approach result in a good factor solution (see Appendix Scenario 2, Section questions 5 – 7)? Motivate your answer.

b) Which method (Unrotated, Varimax or Oblimin) is the most suitable, and provides the best factor solution? Motivate your answer (specify which matrices you use; see Appendix Scenario 2, Section questions 6 – 8).

a)
2 factors

Varimax: good solution, no cross-loadings (1 point)

Oblimin: good solution, no cross-loadings (1 point)

3 factors

Varimax: good solution, no cross-loadings (1 point)

Oblimin: not good solution, cross-loadings Q17 (1 point)

b)
Unrotated and rotated Varimax, 2 factors solutions are the best
Rotated Varimax 2 factor solution is preferable to Oblimin 2 factor solution, because of small correlations among components (1 point)

Unrotated, and rotated Varimax 3 factor solutions are not suitable, because we search for 2 conceptually different factors (idealism vs. realism) (1 point)

(continued on next page)

Question 8:

A possible use of Factor Analysis results is to summate scales (see Appendix Scenario 2, Section questions 6 – 8).

a) What goals are sought when summated scales are created?

b) In the attempt of assessing the reliability of the summated scale, which measure is used? Based on the SPSS output, is the scale reliable or not? Argument your answer.

a)

(i) simplification: representing multiple aspects of a concept into one measure (1 point)

(ii) overcoming measurement error (1 point)

b) Cronbach's alpha = 0.68

This value is sufficient in case of exploratory research, above 0.6 (1 point)

(continued on next page)

Scenario 3 *(maximum 25 points)*
Nutrint is a large manufacturer of nutrients. To inform customers about its products, Nutrint's marketing spending per year is high. The marketing department of Nutrint thinks that opinion and attitude towards nutrition is an important predictor of the appreciation of commercials of Nutrint.
A researcher was delegated to study the effect of various consumer attitudes on advertisement appraisal. For this, he has sent out a number of questionnaires. The researcher uses various statistical techniques to analyze the collected data. The results of the analysis are provided in the SPSS outputs in the Appendix (Scenario 3).

Question 1:

The researcher starts his analysis using a Factor Analysis (see Scenario 3, Section questions 1 – 2).
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.

a) MRA should be used, for dependence is the research objective	(1 point)
ratio observations to variables = 50 : 11, which is lower than 5:1	(1 point)
FA used to simplify the model	(1 point)
b) observations : variables = 50:11 < 5:1	
Therefore FA not allowed	(1 point)

Question 2:

Considering the outcome of the first Factor Analysis, the researcher decides to do a second Factor Analysis (see Scenario 3, Section questions 1 – 2). Explain why the researcher executes a second Factor Analysis.

Communalities of variables 'I don't like to see children's toys lying about' < .5.	(1 point)
Therefore insufficiently explained by factor solution. In second FA, this variable is omitted.	(1 point)

(continued on next page)

Question 3:

The researcher considers the results from the second Factor Analysis to proceed with the analysis (see Scenario 3, Section questions 3 – 4). Do you agree with this decision? Explain your position.

Yes, because, (1 point) All assumptions are met: All individual MSA's > .5, total MSA > .5, Bartlett test significant Solution good: communalities all > .5, no crossloadings (in the rotated solution) (1 point)

Question 4:

Consider the factor solution from Factor Analysis 2 (see Scenario 3, Section questions 3 – 4). How would you label the three factors that were found in this solution? Explain why.

Factor 1 - label: … **price conscious**, because **all variables comprising this factor are about getting the lowest price** (1 point)
Factor 2 – label: **health consciousness**…, because **all variables comprising this factor are about being thoughtful about health implications of food** (1 point)
Factor 3 – label: … **worrying about food**…, because **all variables comprising this factor are about worrying about food** (1 point)

Question 5:

The next analysis the researcher applies is Multivariate Regression Analysis (see Scenario 3, Section questions 5 -- 9). 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.

y = advertisement appraisal x1 = REGR factor score 1 for analysis 1 x2 = REGR factor score 2 for analysis 1 x3 = REGR factor score 3 for analysis 1 x4 = I don't like children's things lying about y = 76.320 + 2.971 * x1 + 2.595 * x2 + 1.867 * x3 + 2.420 * x4 + e (3 points, deduce 1 point if: constant forgotten and/or variable names not specified)

(continued on next page)

Question 6:

Indicate how good the regression model is that the researcher found, and if the model is significant (see Scenario 3, Section questions 5 – 9).

R2 = .345, which means that 34.5 % of variance of dependent variable is explained (1 point)

Alpha=0.05 (or 0.1, 0.01) (1 point)
H0: R2 = 0,
H1: R2 > 0 (1 point)
F(4,45) = 5.931, s.= .001, R2 is significant at alpha = .05, .10, .01 (1 point)

Question 7:

Explain which independent variables have a significant contribution to the explanation of the dependent variable y 'advertisement appraisal' (see Scenario 3, Section questions 5 – 9).

H0: bi = 0;
H1 bi ≠ 0 (1 point)

For alpha = .05: (1 point)
Significant variables are:
REGR factor score 1 (t = 2.249; s = .029),
I don't like to see children's toys lying about (t = 2.069; s = .044) (1point)

OR
For alpha = 0.10:
The above two variables appear to be significant **AND**
REGR factor score 2 (t = 1.994; s = .052) (1point)

OR
For alpha = 0.01:
None of the variables are significant (1point)

(continued on next page)

Question 8:
What would be the predicted value of the 'advertisement appraisal' variable (dependent variable) in the context of Multivariate Regression Analysis (see Scenario 3, Section questions 5 – 9),

a) if all factors are equal to 1 and the variable 'I don't like to see children's toys lying around' is equal to 2
b) if all factors are equal to 2 and the variable 'I don't like to see children's toys lying around' is equal to 0

y = advertisement appraisal
x1 = REGR factor score 1 for analysis 1
x2 = REGR factor score 2 for analysis 1
x3 = REGR factor score 3 for analysis 1
x4 = I don't like children's things lying about

a) x1=x2=x3=1, x4=2
then
y^=76.320 + 2.971 *1 + 2.595 *1 + 1.867 * 1 + 2.420 * 2=88.593 (1 point)

b) x1=x2=x3=2, x4=0
then
y^=76.320 + 2.971 *2 + 2.595 *2 + 1.867 * 2 + 2.420 *0=91.186 (1 point)

Question 9:
Does multicollinearity provide problems in the found regression model? Motivate your answer (see Scenario 3, Section questions 5 – 9).

Multicollinearity does not appear as.
all tolerances > .10, (1 point)
all VIFs < 10 (1 point)

END OF THE EXAM