Examiners' commentaries 2019

ST3188 Statistical methods for market research

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2018–19. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2018). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

General remarks

Learning outcomes

At the end of the course and having completed the essential reading and activities you should be able to:

- define a market research problem and create an appropriate research design
- perform independent data analysis in a market research setting
- determine which statistical method is appropriate in a given situation and be able to discuss the merits and limitations of a particular method
- use statistical software to analyse datasets and be able to interpret output
- draw appropriate conclusions following empirical analysis and use to form the basis of managerial decision-making
- demonstrate greater commercial awareness.

Format of the examination

The examination is two hours long and you must answer the question in Section A and two questions out of three in Section B. The examination is worth 70% of the final grade. The other 30% is determined by the coursework component. (The coursework comprised the production of a market research proposal – see the 'Assessment' section in the VLE for details.)

As a new course examined for the first time in 2019, the examination papers were deliberately similar in nature to the specimen and mock examination papers given the inevitable uncertainties faced by candidates due to no past examination papers being in existence.

In 2020, given past examination papers now exist (i.e. from 2019), candidates can expect more variation in the topics which could appear, remembering that all blocks of the course are examinable.

Overall performance

The performance of candidates in the examination was generally pleasing, with some excellent answers. Since there is no access to a computer in the examination, actual use of SPSS is not feasible to be directly examined, rather some questions in Section B required the interpretation of SPSS output. Some answers lacked sufficient depth of explanation – remember to comment in detail on the output statistically. For example, when reporting on p-values explicitly right out the hypotheses being tested, i.e. H_0 and H_1. An excellent answer would also state the test statistic being used and relate this to the relevant test statistic value.

Although this is an applied statistics course, candidates are reminded that commercial insight is also important. Always think about which business decisions could be taken as a consequence of the market research, justifying the decision(s) based on the results – the course is about market research after all! It is likely any decision will relate to one (or more) of the marketing mix variables – the four 'p's (product, price, placement and promotion).

Examination revision strategy

Many candidates are disappointed to find that their examination performance is poorer than they expected. This may be due to a number of reasons, but one particular failing is '**question spotting**', that is, confining your examination preparation to a few questions and/or topics which have come up in past papers for the course. This can have serious consequences.

We recognise that candidates might not cover all topics in the syllabus in the same depth, but you need to be aware that examiners are free to set questions on **any aspect** of the syllabus. This means that you need to study enough of the syllabus to enable you to answer the required number of examination questions.

The syllabus can be found in the Course information sheet available on the VLE. You should read the syllabus carefully and ensure that you cover sufficient material in preparation for the examination. Examiners will vary the topics and questions from year to year and may well set questions that have not appeared in past papers. Examination papers may legitimately include questions on any topic in the syllabus. So, although past papers can be helpful during your revision, you cannot assume that topics or specific questions that have come up in past examinations will occur again.

If you rely on a question-spotting strategy, it is likely you will find yourself in difficulties when you sit the examination. We strongly advise you not to adopt this strategy.

Examiners' commentaries 2019

ST3188 Statistical methods for market research

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2018–19. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2018). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone A

This paper contains two sections. **Section A** is compulsory and contains one question. **Section B** contains three questions.

Answer **THE** question in **Section A** and **TWO** questions from **Section B**.

Section A carries 40 marks and **Section B** carries 60 marks. All questions in **Section B** will be given equal weight (30 marks).

Section A: Compulsory

Question 1

(a) **A government is considering the need for additional airport capacity due to projections of increased demand for air travel in the years ahead. The government is deciding how the expected demand can be met in the long term. Short-listed options for increasing airport capacity include expansion of one of two existing airports, A or B. The decision to expand either site involves numerous trade-offs.**

You have been asked to devise an appropriate sampling scheme of airport A and airport B users (passengers as well as non-passengers, such as staff and local residents to the airport) to research their views of expanding one or other site. Explain how each of the following sampling methods could be applied to the overall sampling strategy for this study. Make sure you describe the merits and limitations of each as well as how each would be applied in practice.

i. Judgemental sampling.

ii. Snowball sampling.

iii. Systematic sampling.

iv. Cluster sampling.

(20 marks)

Reading for this question

Block 9 on the VLE covers sampling – design and procedures.

Approaching the question

Candidates should avoid generic 'textbook' descriptions of the named sampling techniques. Rather, it is necessary to *explain* how each sampling scheme may be used in the specified application (i.e. researching additional airport capacity in this question).

Clearly distinguishing between non-probability and probability methods, for the latter the sampling frame should be identified. As explicitly mentioned in the question, as well as the mechanics of each method the merits and limitations should be stated.

While there is no single 'right' answer to such a question, the examiners rewarded responses which directly related to the airport capacity case – in particular the different types of users.

As with any sampling, for the results to be meaningful the objective is to obtain a *representative* sample, which different sampling techniques achieve to varying degrees. The extent of representativeness for each technique should be addressed.

(b) Suppose we are interested in estimating the proportion of a population using a simple random sample of size n.

i. State a suitable estimator of the population proportion as well as its sampling distribution. State clearly any assumptions which you make.

ii. Explain statistically how to determine the minimum sample size necessary to estimate a population proportion to within e units.

iii. Explain what the term '95% confidence interval' means and give a market research example.

iv. Explain what the finite population correction is (including a formula) and when it should be used.

(20 marks)

Reading for this question

Block 10 on the VLE covers sample size determination.

Approaching the question

i. Let $\{X_1, \ldots, X_n\}$ be a simple random sample of size n from a Bernoulli(π) distribution, where the X_is are independent and identically distributed. We have:

$$P = \bar{X} \sim N\left(\pi, \frac{\pi\,(1-\pi)}{n}\right)$$

approximately, by the central limit theorem as $n \to \infty$.

ii. For a $100(1-\alpha)\%$ confidence interval, we have:

$$z_{\alpha/2} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq e$$

and so:

$$\frac{(z_{\alpha/2})^2 \, \pi \, (1-\pi)}{e^2} \leq n.$$

The value of π should be either an assumed value, an estimate based on a pilot study, or set equal to 0.5 as a conservative estimate which provides the maximum standard error.

iii. Formally, a 95% confidence interval covers the unknown parameter with 95% probability over repeated samples. Any reasonable example accepted.

iv. When $n \geq 0.1N$, the standard error will be (non-negligibly) overestimated, hence we require a finite population correction factor defined by:

$$\sqrt{\frac{N-n}{N-1}}.$$

In which case the corrected standard error becomes:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}.$$

Section B: Answer two questions. Each question carries equal weight.

Question 2

(a) A grocery store chain surveyed a set of customers concerning their purchasing habits. Given the survey results and how much each customer spent in the previous month, the store wants to see if gender, denoted as `gender`, (categorised as 'male' and 'female') and usage of coupons, denoted as `usecoup`, (categorised as 'no', 'from newspaper', 'from mailings' and 'from both') are related to the amount they spend in a month.

Analyse the selected SPSS output in Figure 1 (spread over the next two pages) and discuss what conclusions can be drawn from the data. In your analysis, be sure to address at least the following:

- **Determine the strength of the joint effect of the factors.**
- **Test the significance of the variables individually and the interaction between them and interpret the results.**
- **State any other three potential factors or covariates which you think might affect monthly spending in the grocery store chain, justifying your choice.**

(20 marks)

(b) i. Describe the semantic differential scale and the Likert scale. For what purposes are these scales used? Provide an example of each scale.

ii. How does the nature and degree of verbal description affect the response to itemised rating scales?

iii. Construct a simple example of a question with an itemised rating scale.

(10 marks)

Figure 1

Descriptive Statistics

Dependent Variable: Amount spent

Use coupons	Gender	Mean	Std. Deviation	N
No	Male	410.1245	75.71493	49
	Female	340.7190	84.25149	52
	Total	374.3910	87.10401	101
From newspaper	Male	391.9883	57.55010	47
	Female	338.2806	84.69003	35
	Total	369.0643	74.87278	82
From mailings	Male	453.5324	93.20937	54
	Female	396.2361	94.03784	46
	Total	427.1761	97.43944	100
From both	Male	474.1714	125.37160	35
	Female	391.4142	94.99714	33
	Total	434.0099	118.41767	68
Total	Male	430.3043	93.47877	185
	Female	365.6671	92.64058	166
	Total	399.7352	98.40821	351

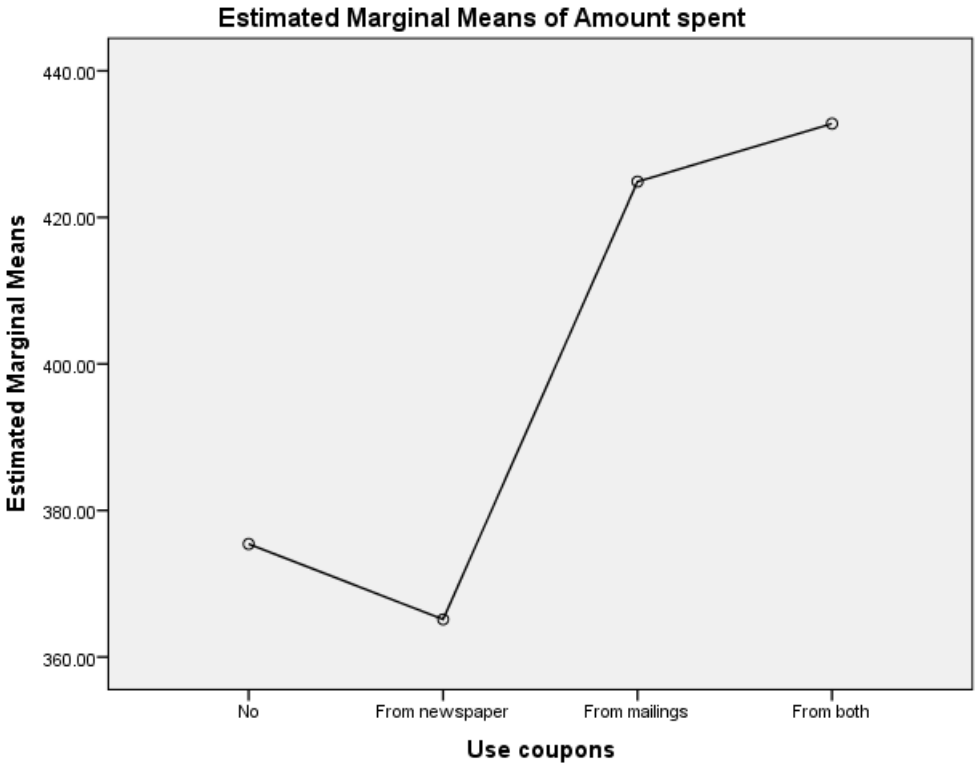
Tests of Between-Subjects Effects

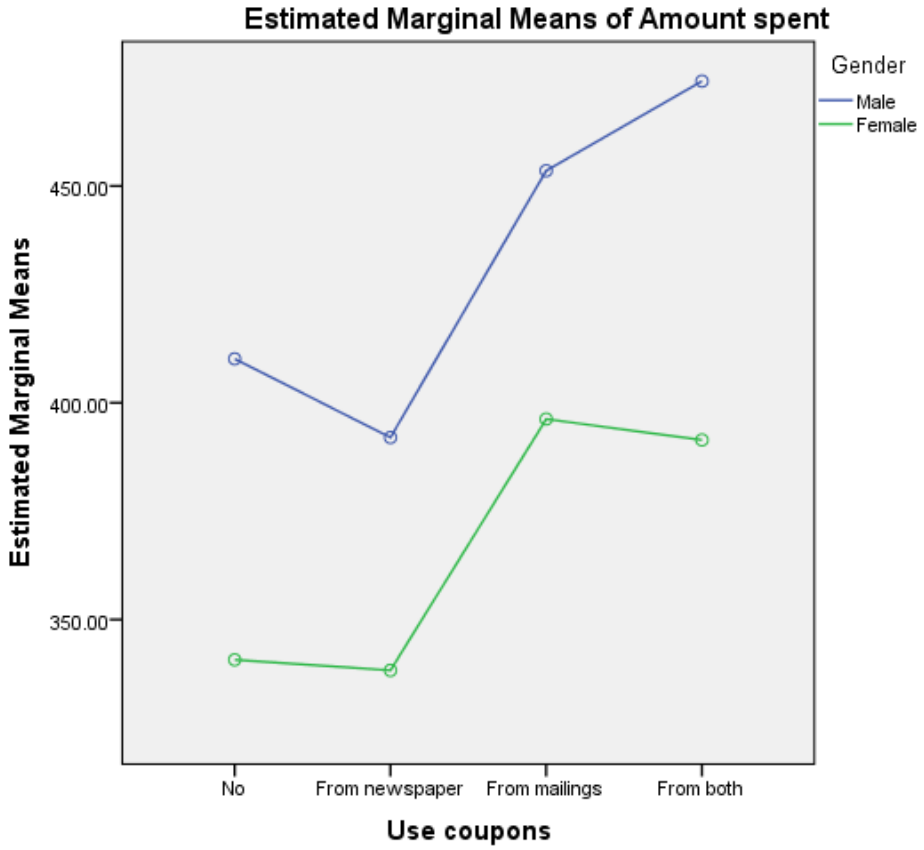
Dependent Variable: Amount spent

Source	Sum of Squares	df	Mean Square	F	Sig.
usecoup	56382853.167	4	14095713.292	1780.785	.000
gender	367388.346	1	367388.346	46.414	.000
usecoup * gender	9877.335	3	3292.445	.416	.742
Error	2714999.592	343	7915.451		
Total	59475118.440	351			

a. R Squared = .954 (Adjusted R Squared = .953)

Figure 1 (continued)





Reading for this question

For part (a), Block 13 on the VLE covers analysis of variance. For part (b), Block 7 on the VLE covers measurement and scaling.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Calculation of multiple η^2 (called 'R Squared' in the output):

$$\frac{56382853.167 + 367388.346 + 9877.335}{59475118.440} = 0.9544$$

  and the subsequent interpretation of this value in terms of the proportion of total variation explained.

- Explaining why both factors are highly significant, and why the interaction is insignificant. Hypotheses, test statistics, test statistic values and p-values should all be cited and correctly interpreted.

- Discussion of the plots, including disordinal interaction with non-crossover (although this is insignificant, as noted above, candidates could comment on any *practical* significance despite the lack of *statistical* significance).

- Any sensibly justified factors and covariates accepted.

(b) i. A semantic differential scale is a seven-point rating scale with bipolar labels which have semantic meaning. In a typical application, respondents rate objects on a series of itemised, seven-point rating scales, bounded at each end by one of two bipolar adjectives, such as 'powerful' or 'weak'. The Likert scale typically has five response categories ranging from 'strongly disagree' to 'strongly agree'. Respondents are required to indicate a degree of agreement or disagreement with each of a series of statements related to the stimulus objects.

These scales are used to measure the strength of feeling about the individual constructs or components of marketing phenomena such as brand, product and company images, feelings about advertising and promotion strategies, new product development studies and in a variety of other applications.

ii. The nature and degree of verbal description associated with scale categories varies considerably and can affect the responses. Scale categories may have verbal, numerical or even pictorial descriptions. Furthermore, the researcher must decide whether to label every scale category, scale some categories or scale only extreme categories. Providing a verbal description for each category may not improve the accuracy or reliability of the data. Yet, an argument can be made for labelling all or many scale categories to reduce scale ambiguity. The category descriptions should be located as close to the response categories as possible.

The strength of the adjectives used to anchor the scale may influence the distribution of the responses. With strong anchors (1 = completely disagree, 7 = completely agree), respondents are less likely to use the extreme scale categories. This results in less variable and more peaked response distributions. Weak anchors (1 = generally disagree, 7 = generally agree), in contrast, produce uniform or flat distributions.

iii. Any sensible example accepted.

Question 3

(a) **A loan officer at a bank wants to be able to identify characteristics which are indicative of people who are likely to default on loans, and the officer wants to use these characteristics to identify good and bad credit risks.**

Records of 700 existing loan customers were analysed using discriminant analysis in an attempt to develop a predictive model for the credit risk of new customers. Each of these 700 customers had either previously defaulted on at least one loan or they had not.

The following information is also available about each of the 700 customers:

- **Age in years.**
- **Years with current employer.**
- **Years at current address.**
- **Household income in thousands.**
- **Debt to income ratio ($\times$100).**
- **Credit card debt in thousands.**
- **Other debt in thousands.**

Analyse the selected SPSS output in Figure 2 (spread over the next two pages) and discuss what conclusions can be drawn from the data.

In your analysis, be sure to address at least the following:

- **State the theoretical and estimated discriminant analysis models.**
- **Comment on the relative importance of the predictor variables.**
- **Discuss how the loan officer should assess the default risk of a new customer. Comment on how accurate any such assessment might be.**

(20 marks)

(b) **Define cross-sectional and longitudinal designs, providing an example of each. Discuss the relative advantages and disadvantages of these two designs.**

(10 marks)

Figure 2

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.404[a]	100.0	100.0	.536

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.712	235.447	7	.000

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age in years	.122
Years with current employer	-.829
Years at current address	-.310
Household income in thousands	.215
Debt to income ratio (x100)	.603
Credit card debt in thousands	.564
Other debt in thousands	-.178

Structure Matrix

	Function 1
Debt to income ratio (x100)	.666
Years with current employer	-.464
Credit card debt in thousands	.397
Years at current address	-.262
Other debt in thousands	.232
Age in years	-.219
Household income in thousands	-.112

Figure 2 (continued)

Canonical Discriminant Function Coefficients

	Function 1
Age in years	.015
Years with current employer	-.130
Years at current address	-.046
Household income in thousands	.006
Debt to income ratio (x100)	.096
Credit card debt in thousands	.275
Other debt in thousands	-.055
(Constant)	-.576

Unstandardized coefficients

Functions at Group Centroids

Previously defaulted	Function 1
No	-.377
Yes	1.066

Unstandardized canonical discriminant functions evaluated at group means

Classification Results

		Previously defaulted	Predicted Group Membership		
			No	Yes	Total
Original	Count	No	393	124	517
		Yes	44	139	183
	%	No	76.0	24.0	100.0
		Yes	24.0	76.0	100.0
Cross-validated[a]	Count	No	391	126	517
		Yes	47	136	183
	%	No	75.6	24.4	100.0
		Yes	25.7	74.3	100.0

Reading for this question

For part (a), Block 15 on the VLE covers discriminant analysis. For part (b), Block 2 on the VLE covers research design.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Theoretical and estimated discriminant analysis models with all terms defined.
- Comment on the relative importance of the predictor variables using the standardised coefficients and the structure matrix.
- Comment on the hit ratio and the development of a classification rule based on the group centroids.

(b) Cross-sectional studies measure units from a sample of the population at only one point in time. Longitudinal studies repeatedly meaasure the same sample units of a population over time. Cross-sectional studies are better in terms of representative sampling and reducing response bias. Longitudinal studies are better at detecting change, collecting large amounts of data and accuracy. Any reasonable examples accepted.

Question 4

(a) An industry analyst would like to predict car sales from a set of predictors. However, many of the predictors are correlated, and the analyst fears that this might adversely affect her results. The analyst uses factor analysis to focus the analysis on a manageable subset of the predictors. Observed variables used in the factor analysis were the following car characteristics:

- **Price in thousands (`price`).**
- **4-year resale value (`resale`).**
- **Engine size (`engine_s`).**
- **Horsepower (`horsepow`).**
- **Width (`width`).**
- **Length (`length`).**
- **Fuel capacity (`fuel_cap`).**
- **Fuel efficiency (`mpg`).**

Figure 3 (spread over the next two pages) presents selected SPSS output from a factor analysis with principal components extraction, using the varimax rotation procedure.

In your analysis, be sure to address at least the following:

- **Explain how you determine the number of factors and interpret the extracted factors.**
- **Explain qualitatively how the fit of the factor analysis model should be examined.**
- **Briefly discuss the relative merits of using 'hard' variables (such as these) versus 'soft' variables (such as attitudinal variables) in factor analysis.**

(20 marks)

(b) Describe the similarities and differences between discriminant analysis, regression and ANOVA.

(10 marks)

Figure 3

KMO and Bartlett's Test

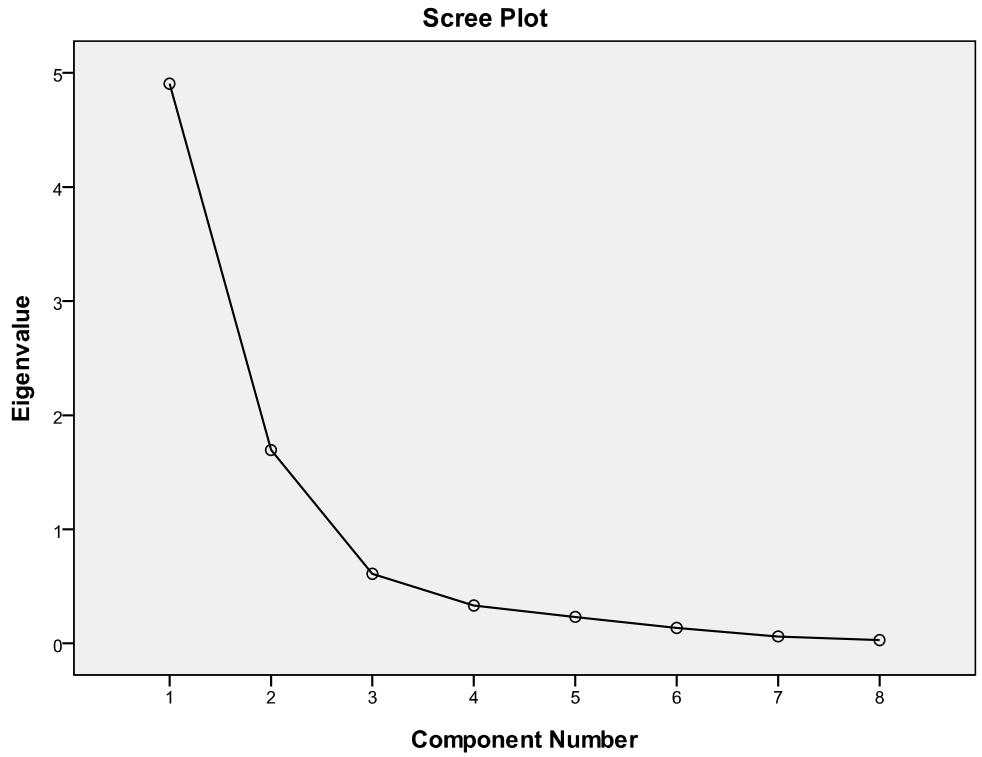
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.792
Bartlett's Test of Sphericity	Approx. Chi-Square	1051.073
	df	28
	Sig.	.000

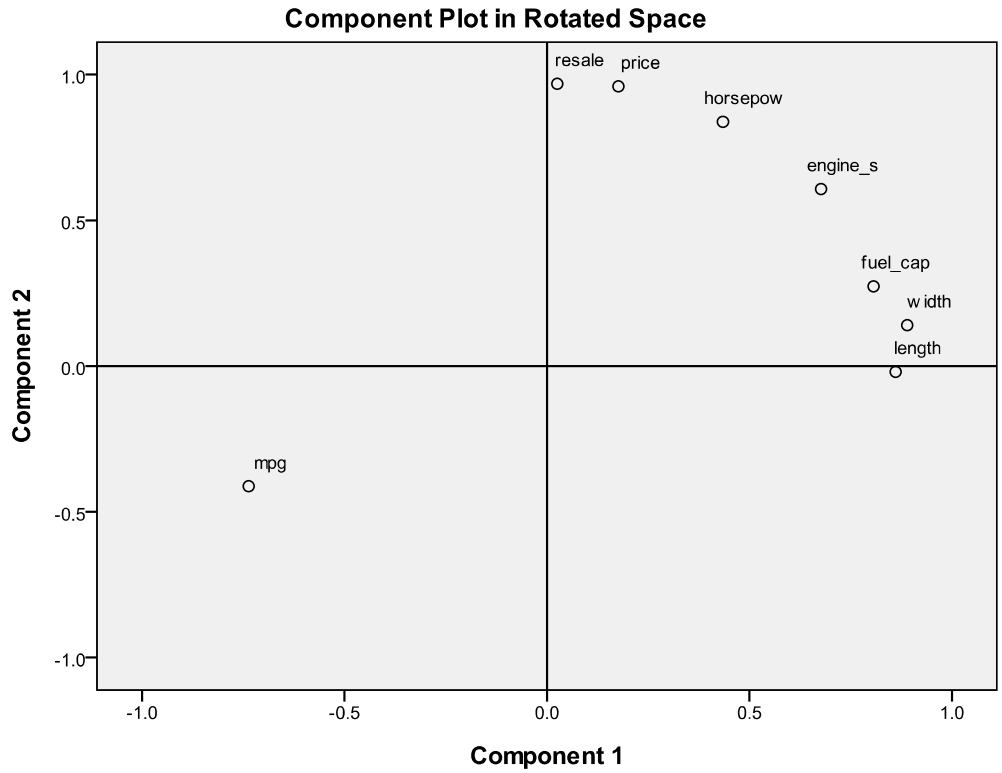
Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.905	61.310	61.310
2	1.694	21.181	82.491
3	.610	7.625	90.117
4	.332	4.148	94.264
5	.232	2.906	97.170
6	.136	1.699	98.870
7	.061	.765	99.635
8	.029	.365	100.000

Rotated Component Matrix[a]

	Component	
	1	2
Price in thousands	.177	.959
4-year resale value	.026	.968
Engine size	.677	.607
Horsepower	.435	.838
Width	.890	.140
Length	.861	-.020
Fuel capacity	.806	.274
Fuel efficiency	-.737	-.412

Figure 3 (continued)





Reading for this question

For part (a), Block 17 on the VLE covers factor analysis. For part (b), Blocks 13, 14 and 15 on the VLE cover these three analytical methods.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Examination of eigenvalues, cumulative percentage variance explained and scree plot to determine the number of factors. Interpreting the factors using the rotated component matrix and associated component plot is important, as well as commenting on how good or poor the results are.
- Model fit is assessed via an examination of residuals, the differences between the observed correlations obtained from the input correlation matrix and the reproduced correlations estimated from the factor matrix. If many large residuals exist, then one can infer that the factor model does not provide a good fit to the data. This analysis is based on the implicit assumption that the observed correlation between the variables is due to the common factors, therefore the correlations between the variables can be reproduced from the estimated correlations between the variables and the factors.
- A discussion of the relative merits of using 'hard' variables versus 'soft' variables is expected.

(b) Discriminant analysis, being a data analysis technique, is related to both regression and ANOVA since they all involve a single criterion or dependent variable and multiple predictor or independent variables. However, the dependent variable is metric in regression and analysis of variance (ANOVA) but it is categorical in the case of discriminant analysis. The independent variables are categorical in the case of ANOVA but metric in the other two procedures.

Two-group discriminant analysis, in which the dependent variable has two categories, is closely related to multiple regression analysis. Here, multiple regression with dummy variables results in partial regression coefficients, which are proportional to the discriminant function coefficients.

Examiners' commentaries 2019

ST3188 Statistical methods for market research

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2018–19. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2018). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone B

This paper contains two sections. **Section A** is compulsory and contains one question. **Section B** contains three questions.

Answer **THE** question in **Section A** and **TWO** questions from **Section B**.

Section A carries 40 marks and **Section B** carries 60 marks. All questions in **Section B** will be given equal weight (30 marks).

Section A: Compulsory

Question 1

(a) **A government is considering the need for additional airport capacity due to projections of increased demand for air travel in the years ahead. The government is deciding how the expected demand can be met in the long term. Short-listed options for increasing airport capacity include expansion of one of two existing airports, A or B. The decision to expand either site involves numerous trade-offs.**

You have been asked to devise an appropriate sampling scheme of airport A and airport B users (passengers as well as non-passengers, such as staff and local residents to the airport) to research their views of expanding one or other site. Explain how each of the following sampling methods could be applied to the overall sampling strategy for this study. Make sure you describe the merits and limitations of each as well as how each would be applied in practice.

i. Convenience sampling.

ii. Quota sampling.

iii. Stratified sampling.

iv. Cluster sampling.

(20 marks)

Reading for this question

Block 9 on the VLE covers sampling – design and procedures.

Approaching the question

Candidates should avoid generic 'textbook' descriptions of the named sampling techniques. Rather, it is necessary to *explain* how each sampling scheme may be used in the specified application (i.e. researching additional airport capacity in this question).

Clearly distinguishing between non-probability and probability methods, for the latter the sampling frame should be identified. As explicitly mentioned in the question, as well as the mechanics of each method the merits and limitations should be stated.

While there is no single 'right' answer to such a question, the examiners rewarded responses which directly related to the airport capacity case – in particular the different types of users.

As with any sampling, for the results to be meaningful the objective is to obtain a *representative* sample, which different sampling techniques achieve to varying degrees. The extent of representativeness for each technique should be addressed.

(b) Suppose we are interested in estimating the proportion of a population using a simple random sample of size n.

i. State a suitable estimator of the population proportion as well as its sampling distribution. State clearly any assumptions which you make.

ii. Explain statistically how to determine the minimum sample size necessary to estimate a population proportion to within e units.

iii. Define the terms 'incidence rate' and 'completion rate'.

iv. Explain how you would adjust the statistically-determined sample size, n, in light of incidence and completion rates.

(20 marks)

Reading for this question

Block 10 on the VLE covers sample size determination.

Approaching the question

i. Let $\{X_1, \ldots, X_n\}$ be a simple random sample of size n from a Bernoulli(π) distribution, where the X_is are independent and identically distributed. We have:

$$P = \bar{X} \sim N\left(\pi, \frac{\pi\,(1-\pi)}{n}\right)$$

approximately, by the central limit theorem as $n \to \infty$.

ii. For a $100(1-\alpha)\%$ confidence interval, we have:

$$z_{\alpha/2} \times \sqrt{\frac{\pi(1-\pi)}{n}} \leq e$$

and so:

$$\frac{(z_{\alpha/2})^2 \, \pi \, (1-\pi)}{e^2} \leq n.$$

The value of π should be either an assumed value, an estimate based on a pilot study, or set equal to 0.5 as a conservative estimate which provides the maximum standard error.

iii. The incidence rate refers to the rate of occurrence, or the percentage, of persons eligible to participate in a study. In general, if there are c qualifying factors with an incidence of Q_1, Q_2, Q_3, …, Q_c, each expressed as a proportion, then:

$$\text{incidence rate} = Q_1 \times Q_2 \times Q_3 \times \cdots \times Q_c.$$

The completion rate is the percentage of qualified respondents who complete the interview, enabling researchers to account for anticipated refusals by people who qualify.

iv. In light of incidence and completion rates, the initial sample size is:

$$\text{initial sample size} = \frac{\text{final sample size}}{\text{incidence rate} \times \text{completion rate}}.$$

Section B: Answer two questions. Each question carries equal weight.

Question 2

(a) **A grocery store chain surveyed a set of customers concerning their purchasing habits. Given the survey results and how much each customer spent in the previous month, the store wants to see if gender, denoted as `gender`, (categorised as 'male' and 'female') and usage of coupons, denoted as `usecoup`, (categorised as 'no', 'from newspaper', 'from mailings' and 'from both') are related to the amount they spend in a month.**

Analyse the selected SPSS output in Figure 1 (spread over the next two pages) and discuss what conclusions can be drawn from the data. In your analysis, be sure to address at least the following:

- **Determine the strength of the joint effect of the factors.**
- **Test the significance of the variables individually and the interaction between them and interpret the results.**
- **State any other three potential factors or covariates which you think might affect monthly spending in the grocery store chain, justifying your choice.**

(20 marks)

(b) i. **What are the major decisions involved in constructing an itemised rating scale?**

ii. **How many scale categories should be used in an itemised rating scale? Briefly explain your answer.**

iii. **Construct a simple example of a question with an itemised rating scale.**

(10 marks)

Figure 1

Descriptive Statistics

Dependent Variable: Amount spent

Use coupons	Gender	Mean	Std. Deviation	N
	Male	410.1245	75.71493	49
No	Female	340.7190	84.25149	52
	Total	374.3910	87.10401	101
	Male	391.9883	57.55010	47
From newspaper	Female	338.2806	84.69003	35
	Total	369.0643	74.87278	82
	Male	453.5324	93.20937	54
From mailings	Female	396.2361	94.03784	46
	Total	427.1761	97.43944	100
	Male	474.1714	125.37160	35
From both	Female	391.4142	94.99714	33
	Total	434.0099	118.41767	68
	Male	430.3043	93.47877	185
Total	Female	365.6671	92.64058	166
	Total	399.7352	98.40821	351

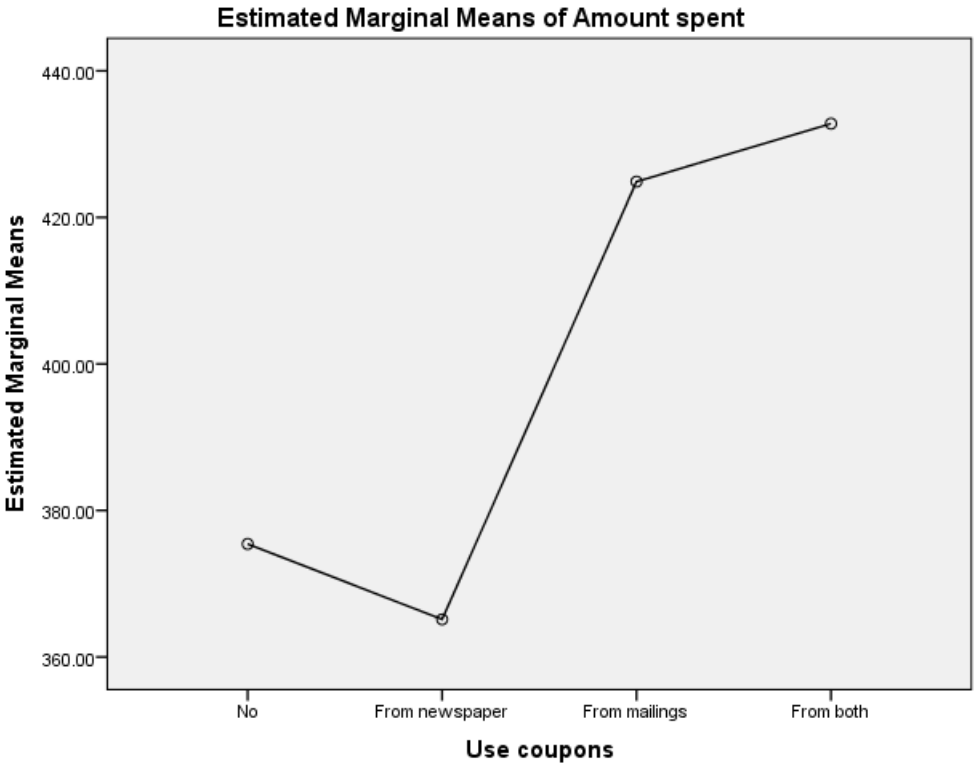
Tests of Between-Subjects Effects

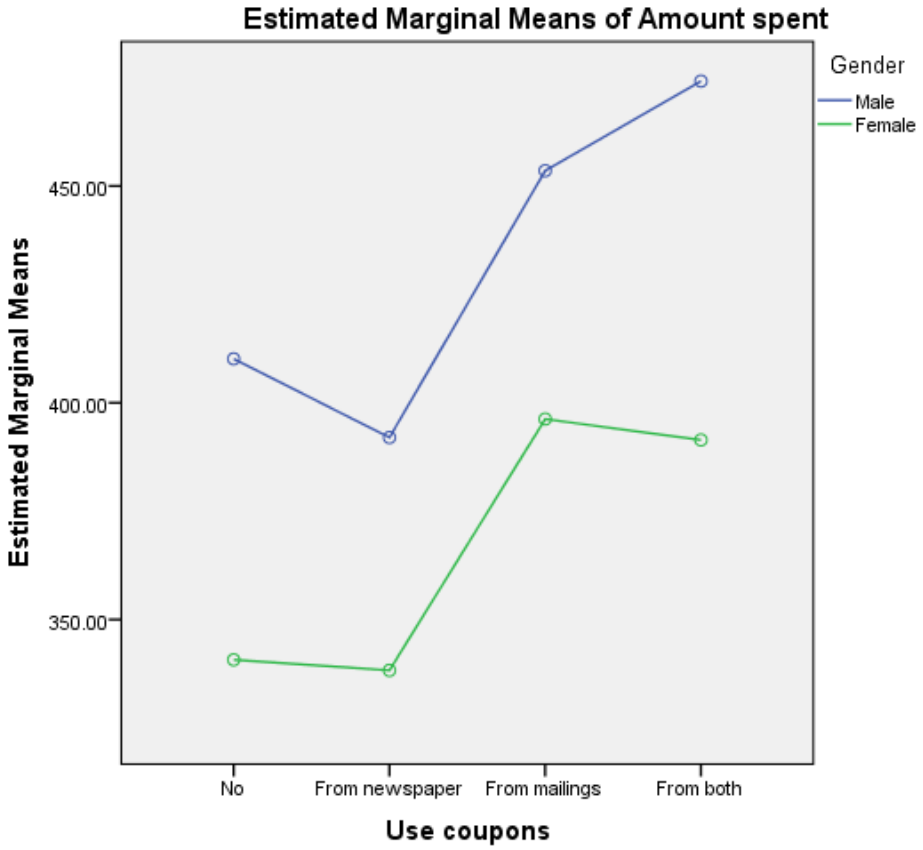
Dependent Variable: Amount spent

Source	Sum of Squares	df	Mean Square	F	Sig.
usecoup	56382853.167	4	14095713.292	1780.785	.000
gender	367388.346	1	367388.346	46.414	.000
usecoup * gender	9877.335	3	3292.445	.416	.742
Error	2714999.592	343	7915.451		
Total	59475118.440	351			

a. R Squared = .954 (Adjusted R Squared = .953)

Figure 1 (continued)





Reading for this question

For part (a), Block 13 on the VLE covers analysis of variance. For part (b), Block 7 on the VLE covers measurement and scaling.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Calculation of multiple η^2 (called 'R Squared' in the output):

$$\frac{56382853.167 + 367388.346 + 9877.335}{59475118.440} = 0.9544$$

  and the subsequent interpretation of this value in terms of the proportion of total variation explained.

- Explaining why both factors are highly significant, and why the interaction is insignificant. Hypotheses, test statistics, test statistic values and p-values should all be cited and correctly interpreted.

- Discussion of the plots, including disordinal interaction with non-crossover (although this is insignificant, as noted above, candidates could comment on any *practical* significance despite the lack of *statistical* significance).

- Any sensibly justified factors and covariates accepted.

(b) i. The major decisions involved in constructing an itemised rating scale are the following.

- The number of scale categories to use – i.e. the means to finely discriminate.
- Balanced versus unbalanced scale – i.e. should the scale be skewed or not?
- Odd or even number of categories – i.e. what role does the middle scale position have?
- Forced versus non-forced nature of the scale – i.e. do respondents really think through the issues or opt for a simple middle path?
- The nature and degree of verbal description to employ – i.e. how the use of words in scale items clarifies the issues for respondents.
- The physical form of the scale – i.e. how attractive, interesting or simple the scale may appear, drawing in the respondent to want to engage with it.

ii. The number of scale categories that should be used in an itemised rating scale should be between three and ten. However, there is no single, optimal number of categories, which would be applicable for all scaling situations.

iii. Any sensible example accepted.

Question 3

(a) **A loan officer at a bank wants to be able to identify characteristics which are indicative of people who are likely to default on loans, and the officer wants to use these characteristics to identify good and bad credit risks.**

Records of 700 existing loan customers were analysed using discriminant analysis in an attempt to develop a predictive model for the credit risk of new customers. Each of these 700 customers had either previously defaulted on at least one loan or they had not.

The following information is also available about each of the 700 customers:

- **Age in years.**
- **Years with current employer.**
- **Years at current address.**
- **Household income in thousands.**
- **Debt to income ratio ($\times$100).**
- **Credit card debt in thousands.**
- **Other debt in thousands.**

Analyse the selected SPSS output in Figure 2 (spread over the next two pages) and discuss what conclusions can be drawn from the data.

In your analysis, be sure to address at least the following:

- **State the theoretical and estimated discriminant analysis models.**
- **Comment on the relative importance of the predictor variables.**
- **Discuss how the loan officer should assess the default risk of a new customer. Comment on how accurate any such assessment might be.**

(20 marks)

(b) **i. What is non-response bias and why is it problematic?**

ii. Identify three ways to improve response rates, explaining how these could be effective.

iii. Discuss three techniques to adjust for non-response.

(10 marks)

Figure 2

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.404[a]	100.0	100.0	.536

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.712	235.447	7	.000

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age in years	.122
Years with current employer	-.829
Years at current address	-.310
Household income in thousands	.215
Debt to income ratio (x100)	.603
Credit card debt in thousands	.564
Other debt in thousands	-.178

Structure Matrix

	Function 1
Debt to income ratio (x100)	.666
Years with current employer	-.464
Credit card debt in thousands	.397
Years at current address	-.262
Other debt in thousands	.232
Age in years	-.219
Household income in thousands	-.112

Figure 2 (continued)

Canonical Discriminant Function Coefficients

	Function 1
Age in years	.015
Years with current employer	-.130
Years at current address	-.046
Household income in thousands	.006
Debt to income ratio (x100)	.096
Credit card debt in thousands	.275
Other debt in thousands	-.055
(Constant)	-.576

Unstandardized coefficients

Functions at Group Centroids

Previously defaulted	Function 1
No	-.377
Yes	1.066

Unstandardized canonical discriminant functions evaluated at group means

Classification Results

		Previously defaulted	Predicted Group Membership		Total
			No	Yes	
Original	Count	No	393	124	517
		Yes	44	139	183
	%	No	76.0	24.0	100.0
		Yes	24.0	76.0	100.0
Cross-validated[a]	Count	No	391	126	517
		Yes	47	136	183
	%	No	75.6	24.4	100.0
		Yes	25.7	74.3	100.0

Reading for this question

For part (a), Block 15 on the VLE covers discriminant analysis. For part (b), Block 10 on the VLE covers non-response and how to adjust for it.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Theoretical and estimated discriminant analysis models with all terms defined.
- Comment on the relative importance of the predictor variables using the standardised coefficients and the structure matrix.
- Comment on the hit ratio and the development of a classification rule based on the group centroids.

(b) i. Non-response bias leads to non-representative samples due to selected respondents refusing to participate in the survey.

ii. Response rates could be improved through callbacks, prior notification, incentives, follow-ups, different facilitators etc.

iii. Possible strategies for adjusting for non-response include the following.

- Subsampling of non-respondents – a concerted effort is made to contact a subsample of the respondents, usually by means of telephone or personal interviews.
- Replacement – the non-respondents in the current survey are replaced with non-respondents from an earlier, similar survey.
- Substitution – the non-respondents are substituted with other elements from the sampling frame who are expected to respond.
- Subjective estimates – this involves making a subjective evaluation of the likely effects of non-response based on experience and available information.
- Trend analysis – the researcher tries to discern a trend between early and late respondents. This trend is projected to non-respondents to estimate their characteristic of interest.
- Simple weighting – depending on the response rates, differential weights are assigned to the data to account for non-response.
- Imputation – imputing the characteristic of interest of the non-respondents based on the similarity of the variables available for both non-respondents and respondents.

Question 4

(a) **An industry analyst would like to predict car sales from a set of predictors. However, many of the predictors are correlated, and the analyst fears that this might adversely affect her results. The analyst uses factor analysis to focus the analysis on a manageable subset of the predictors. Observed variables used in the factor analysis were the following car characteristics:**

- **Price in thousands (`price`).**
- **4-year resale value (`resale`).**
- **Engine size (`engine_s`).**
- **Horsepower (`horsepow`).**
- **Width (`width`).**
- **Length (`length`).**
- **Fuel capacity (`fuel_cap`).**
- **Fuel efficiency (`mpg`).**

Figure 3 (spread over the next two pages) presents selected SPSS output from a factor analysis with principal components extraction, using the varimax rotation procedure.

In your analysis, be sure to address at least the following:

- **Explain how you determine the number of factors and interpret the extracted factors.**
- **Explain qualitatively how the fit of the factor analysis model should be examined.**
- **Briefly discuss the relative merits of using 'hard' variables (such as these) versus 'soft' variables (such as attitudinal variables) in factor analysis.**

(20 marks)

(b) i. **Discuss the similarity and difference between cluster analysis and discriminant analysis.**

ii. **What are some of the uses of cluster analysis in marketing?**

iii. **What is involved in the interpretation of clusters?**

(10 marks)

Figure 3

KMO and Bartlett's Test

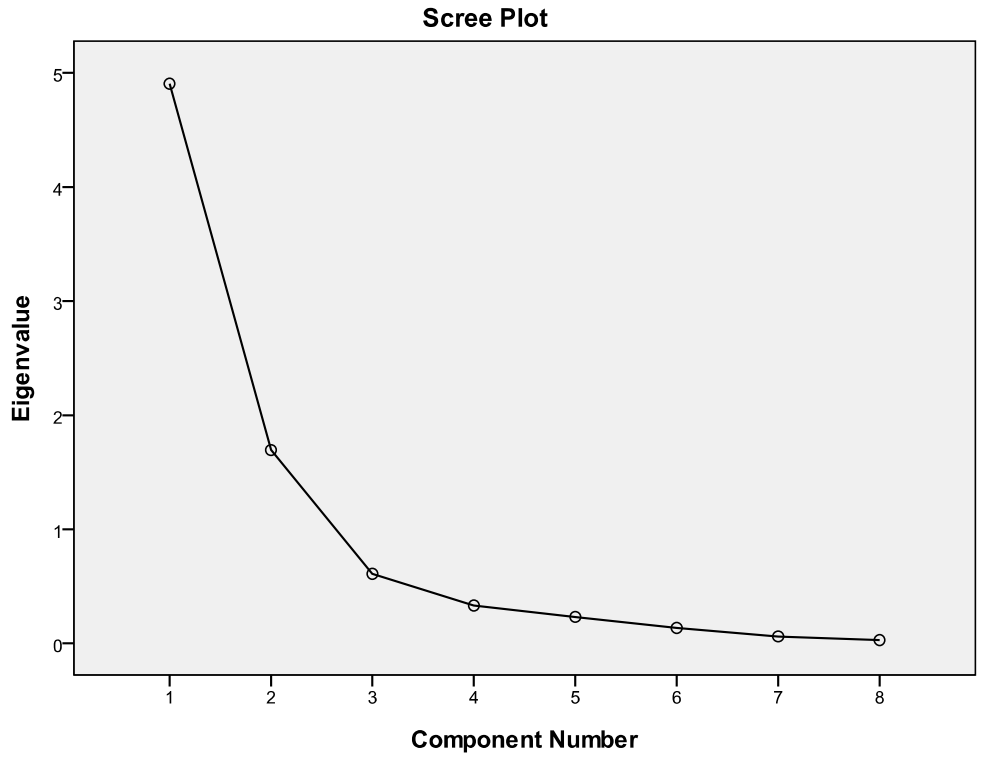
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.792
Bartlett's Test of Sphericity	Approx. Chi-Square	1051.073
	df	28
	Sig.	.000

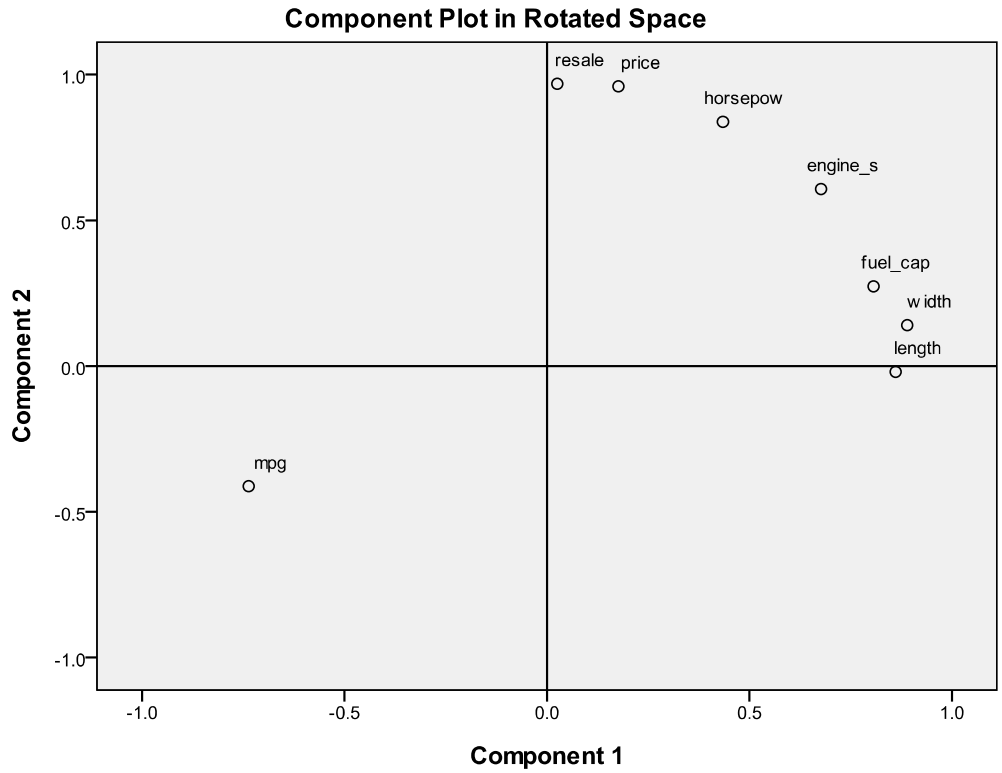
Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	4.905	61.310	61.310
2	1.694	21.181	82.491
3	.610	7.625	90.117
4	.332	4.148	94.264
5	.232	2.906	97.170
6	.136	1.699	98.870
7	.061	.765	99.635
8	.029	.365	100.000

Rotated Component Matrix[a]

	Component	
	1	2
Price in thousands	.177	.959
4-year resale value	.026	.968
Engine size	.677	.607
Horsepower	.435	.838
Width	.890	.140
Length	.861	-.020
Fuel capacity	.806	.274
Fuel efficiency	-.737	-.412

Figure 3 (continued)





Reading for this question

For part (a), Block 17 on the VLE covers factor analysis. For part (b), Block 18 on the VLE covers cluster analysis.

Approaching the question

(a) An excellent answer would consider the *statistical* and *commercial* interpretation of the SPSS output. Key 'ingredients' would include the following.

- Examination of eigenvalues, cumulative percentage variance explained and scree plot to determine the number of factors. Interpreting the factors using the rotated component matrix and associated component plot is important, as well as commenting on how good or poor the results are.
- Model fit is assessed via an examination of residuals, the differences between the observed correlations obtained from the input correlation matrix and the reproduced correlations estimated from the factor matrix. If many large residuals exist, then one can infer that the factor model does not provide a good fit to the data. This analysis is based on the implicit assumption that the observed correlation between the variables is due to the common factors, therefore the correlations between the variables can be reproduced from the estimated correlations between the variables and the factors.
- A discussion of the relative merits of using 'hard' variables versus 'soft' variables is expected.

(b) i. Both cluster analysis and discriminant analysis are concerned with the classification of objects, cases or variables into relatively homogeneous groups. However, in cluster analysis the groups or clusters are suggested by the data, whereas in discriminant analysis they are defined *a priori*.

ii. The technique has several applications in marketing. Some of them are as follows.

- Segmenting the market. Using this technique, consumers are clustered on the basis of benefits sought from the purchase of a particular product.
- Understanding buyer behaviour. Cluster analysis is used to classify homogeneous groups of buyers to facilitate the study of buyer behaviour in each group.
- Identifying new product opportunities. By using this technique, a firm can cluster the brands/products and examine its current offerings compared to those of its competitors to identify potential new product opportunities.
- Selecting test markets. It can be used to group cities into homogeneous clusters and then select comparable cities to test various marketing strategies.
- Reducing data. It may be utilised to develop clusters or subgroups of data which are more manageable than individual observations, to be used in subsequent multivariate analyses.

iii. The clusters should be interpreted in terms of cluster centroids which represent the average values of the objects contained in the cluster on each of the variables. This enables the researcher to describe each cluster by assigning it a name or label, which may be obtained from the cluster programme or through discriminant analysis.