Examiners' commentaries 2020

ST3188 Statistical methods for market research

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2019–20. 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 (2019). 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.)

Overall performance

The performance of candidates in the examination was generally pleasing, with some excellent answers. Use of SPSS is not 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 2020

ST3188 Statistical methods for market research

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2019–20. 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 (2019). 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

Candidates should answer the **ONE** question in Section A and **TWO** questions from Section B. Section A carries 40 marks. Questions in Section B carry 30 marks each.

Section A: Compulsory

Question 1

(a) **A travel agency offers customers a range of ways to make holiday bookings – in store, online and through their call centres. Revenues are generated through different commission rates on holidays paid by tour operators (the actual suppliers of the holiday products). The company prides itself on delivering customer satisfaction. Therefore, management is keen to research customer satisfaction levels across the different booking methods and by tour operators (some tour operators offer higher commission margins).**

To determine the level of customer satisfaction, the company's management has decided to use a survey of all types of customers and have asked you to devise an appropriate sampling scheme. Explain in detail 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. Quota sampling

ii. Snowball sampling

iii. Systematic sampling

iv. Stratified 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 customer satisfaction levels 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 customer satisfaction of holiday bookings – in particular the different types of booking methods (in store, online and telesales).

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. Mention 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. Provide a practical marketing example of a 95% confidence interval for a proportion.

iv. Explain the purpose of the finite population correction factor (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 $\{X_1, X_2, \ldots, X_n\}$ 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:

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

where e is the maximum tolerance for the sampling error, 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. 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:

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

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

Question 2

(a) A retail chain has 5 stores: A, B, C, D and E in a city. The average number of sales transactions (sales volume) based on a random sample has been calculated for three periods during the day (morning, afternoon and evening). The retailer conducts a two-way analysis of variance to investigate whether there appears to be a difference in the average number of sales transactions in different stores and at different times of the day.

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:

- **Describe the strength of the joint effect of the factors.**
- **Test the significance of the variables individually and the interaction between them.**
- **Describe the pattern of interaction.**

(20 marks)

(b) Discuss the reasons for the frequent use of cross-tabulations in market research. What are some of the limitations? Give two examples of cross-tabulations you might expect to find in a market research study.

(10 marks)

Figure 1

Tests of Between-Subjects Effects

Dependent Variable:Sales volume

Source	Sum of Squares	df	Mean Square	F	Sig.
Time_of_day	1431.667	2	715.833	421.078	.000
Store	1367.467	4	341.867	201.098	.000
Time_of_day * Store	812.333	8	101.542	59.730	.000
Error	25.500	15	1.700		
Total	3636.967	29			

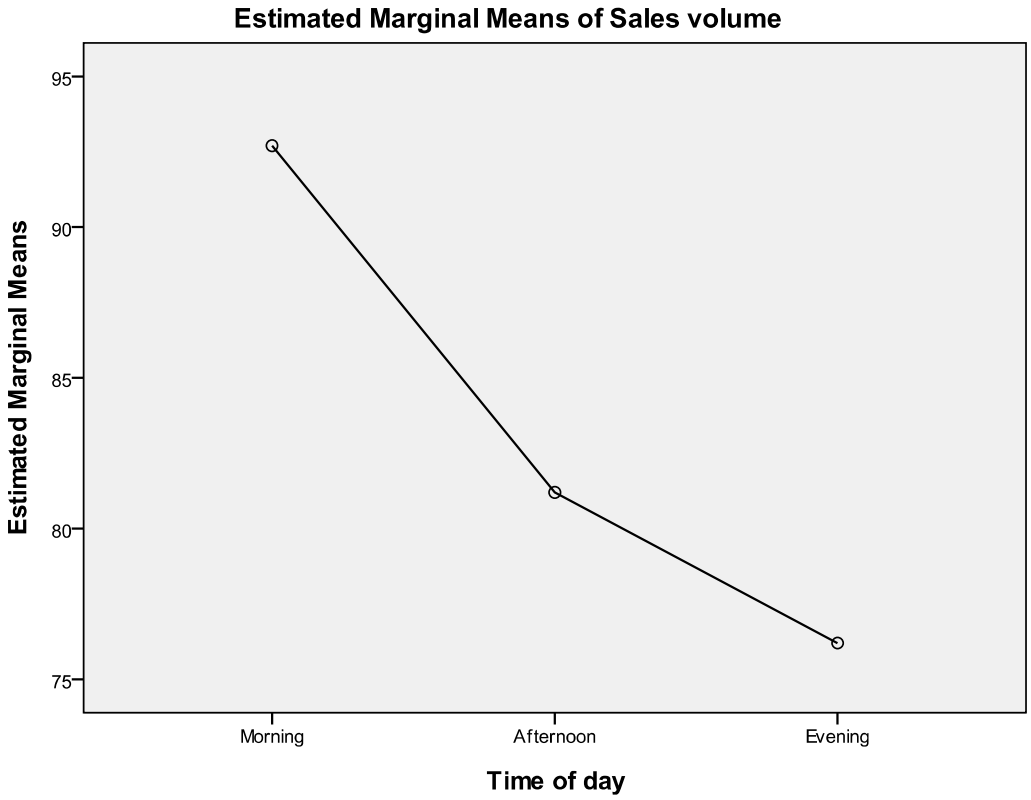
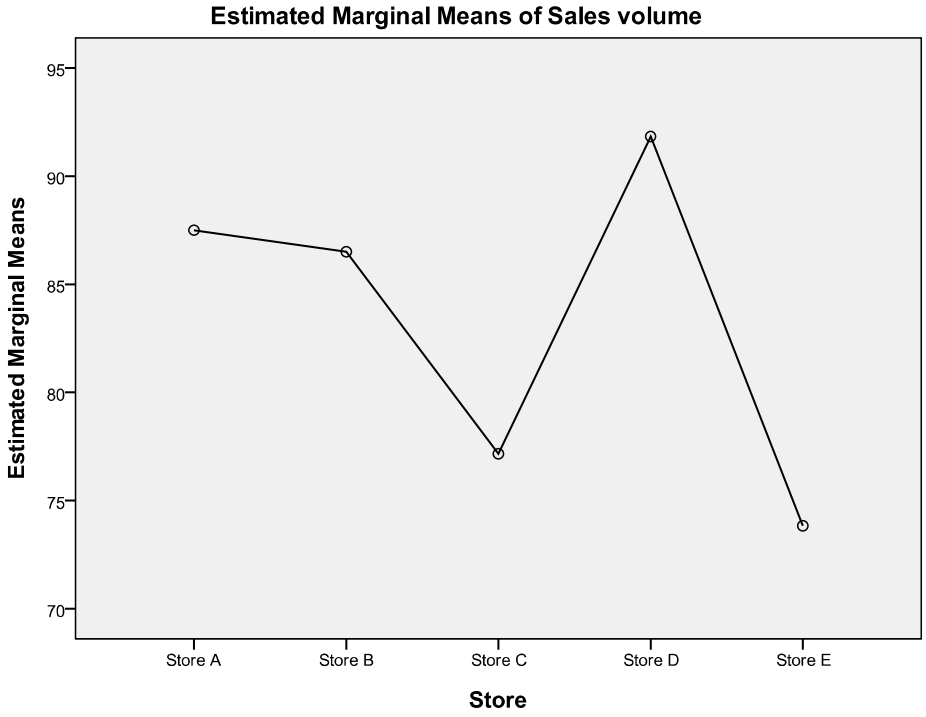
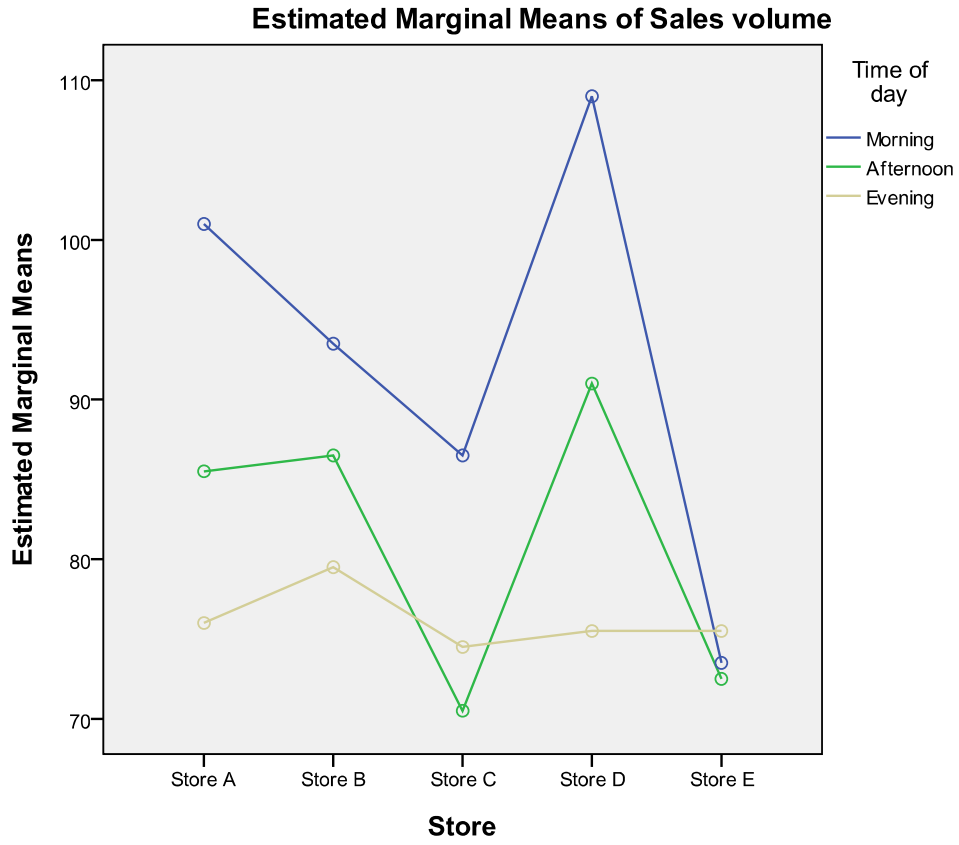


Figure 1 (continued)





Reading for this question

For part (a), Block 13 on the VLE covers analysis of variance. For part (b), Block 12 on the VLE covers cross-tabulation.

Approaching the question

(a) Output interpretation should include at least the following:

- Calculation of multiple η^2:

$$\frac{1{,}431.667 + 1{,}367.467 + 812.333}{3{,}636.967} = 0.9929$$

  and interpretation.
- Both factors are highly significant, as is the interaction. Hypotheses, test statistics, test statistic values and p-values should all be cited and correctly interpreted.
- Discussion of plots, including disordinal interaction featuring crossover.

Note that Store D has the greatest variation in sales throughout the day, although its evening sales are similar to those in other stores. Hence there is scope to recommend that the retail chain conducts some further market research to understand why Store D is performing so strongly in the mornings. However, Store E's sales are struggling regardless of time of day. If this cannot be remedied, then closure of Store E should be considered.

(b) Cross-tabulations are popular for the following reasons:

- Ease of comprehension, i.e. cross-tabulation analysis and results can be easily interpreted and understood by managers who have little statistical orientation.
- Versatility, i.e. a series of cross-tabulations may provide greater insights into a complex phenomenon than a single multivariate analysis.
- Clarity, i.e. the clarity of interpretations provides a stronger link between the research results and managerial action.
- Simplicity, i.e. cross-tabulation analysis is simple to conduct and appealing to the less sophisticated researcher. Cross-tabulation, though meant for describing the joint distribution of two or more variables, is seldom used in computations involving more than three variables. This is because the interpretation becomes quite complex. Also, since the number of cells increases multiplicatively, maintaining an adequate number of respondents in each cell becomes problematic. Consequently the statistics computed could be unreliable. Besides, since only two or three variables are tabulated at a time, cross-tabulation is not a very efficient way of examining the relationships when there are several variables.

Any sensible examples of cross-tabulations accepted.

Question 3

(a) **A hotel chain decides to use data on existing hotels to determine desirable locations for new hotels. Data on 90 hotels in the chain have been collected in an attempt to understand operating margin. The variables are:**

- `Operating margin` – **the dependent variable, in %**
- `Indoor pool` – **whether the hotel has a heated indoor pool, Yes = 1 and No = 0**
- `Competitor rooms` – **total number of competitor rooms in hotels within 3 miles**
- `Distance to competitor` – **miles to nearest competitor hotel**
- `Office space` – **amount of office space available within 3 miles, in thousands of square feet**
- `Distance to downtown` – **miles to downtown (the central part of a city or town).**

Selected SPSS output is provided in Figure 2 (on the next page). Analyse the regression results, making sure you first write out the full regression model, including any assumptions, and the estimated model. In light of the regression results, propose any changes you would make to the model, including any other explanatory variables which you would consider using, or any you would remove from the existing model. Keep in mind the hotel's objective of understanding operating margin to inform its decisions about new hotel sites.

(20 marks)

(b) i. **Explain the stepwise regression approach. What is its purpose?**

ii. **What is multicollinearity? Also, what problems could arise because of the presence of multicollinearity?**

(10 marks)

Figure 2

Model Summary

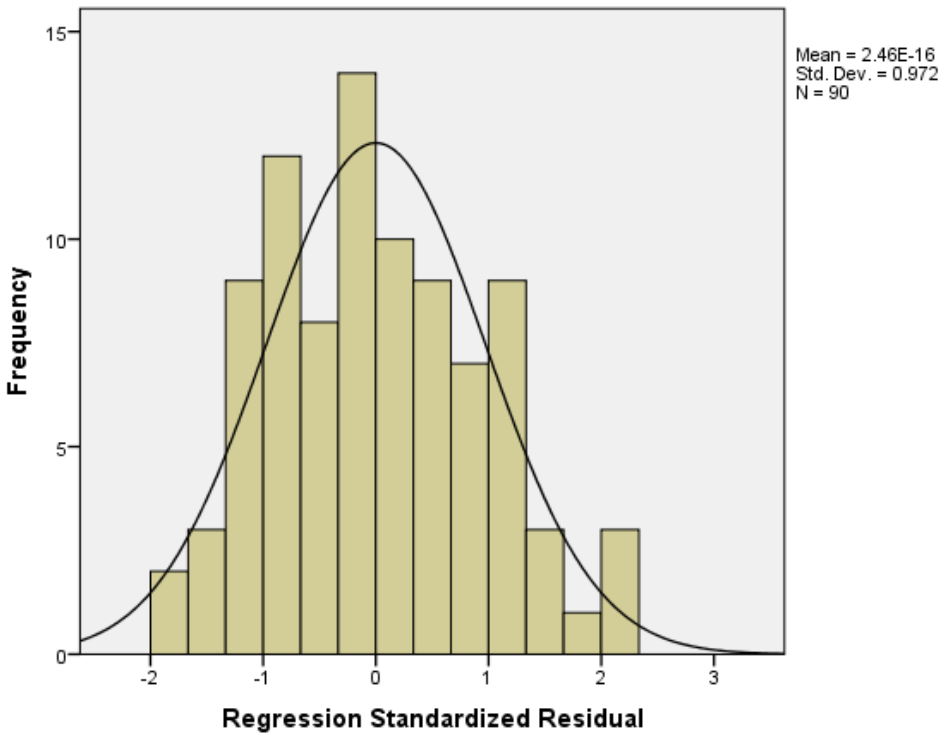
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.884[a]	.782	.769	2.9158%

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2557.772	5	511.554	60.169	.000[b]
	Residual	714.164	84	8.502		
	Total	3271.936	89			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	58.081	2.659		21.840	.000
	Indoor pool	2.316	.671	.184	3.453	.001
	Competitor rooms	-.007	.001	-.536	-10.263	.000
	Distance to competitor	-1.864	.360	-.272	-5.172	.000
	Office space	.019	.002	.512	9.717	.000
	Distance to downtown	.195	.099	.103	1.971	.052



Reading for this question

Block 14 on the VLE covers linear regression, which includes a discussion of multicollinearity.

Approaching the question

(a) Good answers would include the following:

- Regression model in full with assumptions on the random error term, plus the estimated model.
- Discussion of output results such as R^2, the F statistic, also regression coefficients (including standardised coefficients) and their significance, the construction of confidence intervals, and the residual histogram.
- Competitor rooms and office space are the most important predictor variables, with distance to downtown the least influential, yet still significant if an upper-tailed test is performed.
- Sensible discussion of model improvements.

(b) i. In stepwise regression, predictor variables are entered or removed from the regression model one at a time based on their contribution towards explaining the variation inherent in the data. In forward inclusion, predictor variables are added to the model (initially starting at 0) one at a time if they have significant F-ratios. In backward elimination, the least significant variables are removed from the model (initially all variables are included) until all remaining variables are significant. The stepwise procedure combines forward inclusion with backward elimination to arrive at the model. The purpose of stepwise regression is to select a smaller subset of a larger number of predictor variables that will account for most of the variation in the data.

ii. Multicollinearity refers to very high inter-correlations among the predictor variables. Multicollinearity can result in the following problems:

- Precise estimation of the partial regression coefficients may vary with the sample.
- The magnitudes and signs of the partial regression coefficients may vary with the sample.
- Unambiguous measurements of the relative importance of the independent variables in explaining the dependent variable become difficult.
- Incorrect removal or inclusion of predictor variables may occur in the stepwise procedure.

Question 4

(a) **Applicants for a position in a firm were asked to score themselves, from 0 to 10, through a questionnaire on the following ten characteristics:**

- **Ambition**
- **Appearance**
- **Drive**
- **Experience**
- **Honesty**
- **Likeability**
- **Potential**
- **Salesmanship**
- **Self-confidence**
- **Suitability.**

When reviewing the questionnaire, because many correlations between the variables are high, it was felt that some of the variables might be confusing, and/or some variables might be redundant. Therefore, a factor analysis was conducted to determine if any underlying factors could be extracted.

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. Interpret the output. 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 and quantitatively how the fit of the factor analysis model should be examined.**
- **Briefly discuss for what modelling purpose(s) any extracted factors could be used.**

(20 marks)

(b) **How may 'operational data' held by organisations help them to build up an understanding of customer behaviour? Support your answer with examples.**

(10 marks)

Figure 3

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.782
Bartlett's Test of Sphericity	Approx. Chi-Square	377.094
	df	45
	Sig.	.000

Communalities

	Initial	Extraction
Appearance	1.000	.432
Likeability	1.000	.841
Self-confidence	1.000	.889
Honesty	1.000	.851
Salesmanship	1.000	.888
Experience	1.000	.848
Drive	1.000	.810
Ambition	1.000	.910
Potential	1.000	.840
Suitability	1.000	.864

Extraction Method: Principal Component Analysis.

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	5.336	53.362	53.362
2	1.656	16.562	69.924
3	1.181	11.811	81.735
4	.689	6.891	88.626
5	.336	3.359	91.985
6	.287	2.873	94.858
7	.192	1.925	96.783
8	.132	1.319	98.102
9	.117	1.169	99.271
10	.073	.729	100.000

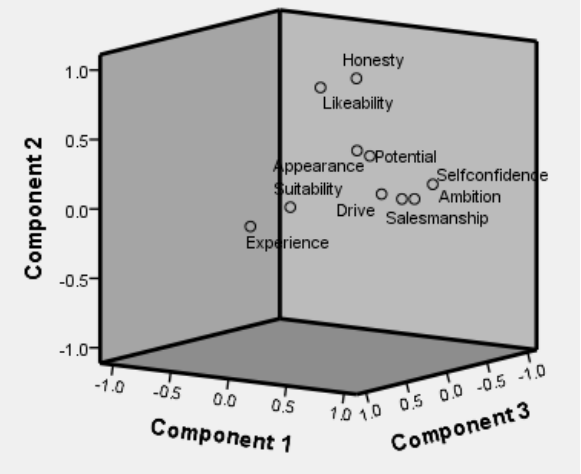


Figure 3 (continued)

Rotated Component Matrix[a]

	Component		
	1	2	3
Appearance	.454	.443	.171
Likeability	.170	.875	.217
Self-confidence	.914	.206	-.109
Honesty	.178	.879	-.217
Salesmanship	.897	.150	.248
Experience	.055	-.036	.918
Drive	.801	.192	.363
Ambition	.933	.138	.144
Potential	.703	.458	.368
Suitability	.353	.123	.851

Component Score Coefficient Matrix

	Component		
	1	2	3
Appearance	.040	.180	.023
Likeability	-.184	.529	.108
Self-confidence	.329	-.087	-.233
Honesty	-.101	.525	-.152
Salesmanship	.270	-.115	-.018
Experience	-.141	-.026	.538
Drive	.200	-.064	.069
Ambition	.306	-.133	-.087
Potential	.097	.132	.094
Suitability	-.050	.005	.444

Reproduced Correlations

		Appearance	Likeability	Self-confidence	Honesty	Salesmanship	Experience	Drive	Ambition	Potential	Suitability
Reproduced Correlation	Appearance	.432[a]	.502	.488	.433	.516	.166	.511	.510	.585	.361
	Likeability	.502	.841[a]	.312	.752	.337	.177	.383	.311	.600	.352
	Self-confidence	.488	.312	.889[a]	.367	.823	-.057	.732	.865	.697	.256
	Honesty	.433	.752	.367	.851[a]	.237	-.221	.233	.256	.448	-.013
	Salesmanship	.516	.337	.823	.237	.888[a]	.271	.837	.893	.790	.546
	Experience	.166	.177	-.057	-.221	.271	.848[a]	.370	.178	.360	.797
	Drive	.511	.383	.732	.233	.837	.370	.810[a]	.826	.785	.615
	Ambition	.510	.311	.865	.256	.893	.178	.826	.910[a]	.772	.469
	Potential	.585	.600	.697	.448	.790	.360	.785	.772	.840[a]	.618
	Suitability	.361	.352	.256	-.013	.546	.797	.615	.469	.618	.864[a]

Reading for this question

For part (a), Block 17 on the VLE covers factor analysis. For part (b), Block 3 covers operational data as part of internal secondary data.

Approaching the question

(a) Output interpretation should include at least the following:

- Examination of eigenvalues and cumulative percentage variance explained to determine the number of factors. Interpretation of factors using rotated component matrix and associated component plot. Comment on how good or poor the results are using the reproduced correlations. Discussion of how factor scores can be calculated.
- 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.
- Discussion of how factor scores can be used in multiple regression and discriminant analysis.

(b) Operational data are data which represent the daily activities and transactions of a business. Transactions may be held in different departments such as sales, accounts or human resources and stored in different ways. The use of operational data has presented opportunities to researchers for long as businesses have been recording their daily transactions. Even in the days of transactions being recorded manually, it was the task of market researchers to track down different sources of data and analyse them. Locating and analysing internal sources of secondary data can be the starting point in many market research projects. The main reasons are that, as these data have already been collected, there are no additional data collection costs, there should be no access problems (individual managers may make access difficult for personal or political reasons) and the quality of the data should be easier to establish (in comparison to externally-generated data).

Most organisations have a wealth of in-house information even if they are not marketing- or customer-focused, so some data may be readily available. In building up an understanding of customer behaviour, operational data from invoices, for example, could answer the following questions.

- What products do customers buy?
- Which types of customer buy the most products?
- Which types of customers repeat purchases?
- Which types of customers appear only when there are special offers?
- Where are these customers located?
- How do these customers pay – by cash or credit?
- Which types of customers are the most profitable?
- Seasonal patterns of purchasing behaviour by product types and customer types.

Answers should be supported with real-world examples.