Examiners' commentaries 2019

ST3188 Statistical methods for market research: Preliminary examination

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.

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. The *Examiners' commentaries* suggest ways of addressing common problems and improving your performance. 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 may not cover all topics in the syllabus in the same depth, but you need to be aware that the 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 in the section of the VLE dedicated to each course. 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: Preliminary examination

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

Section A

Question 1

(a) **Reading for this question**

Block 9 of the course.

Approaching the question

Standard responses expected here, noting which are probability and non-probability methods. For iii. and iv., a viable sampling frame should be identified (such as customer database). As explicitly mentioned in the question, as well as the mechanics of each method the merits and limitations should be stated.

(b) **Reading for this question**

Block 10 of the course.

Approaching the question

i. A sampling distribution is the probability distribution of an estimator, indicating how likely different point estimates are.

ii. 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 $n \to \infty$.

iii. 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.

iv. A discussion of the trade-off between the coverage probability and the confidence interval width.

Section B

Answer two questions. Each question carries equal weight.

Question 2

(a) **Reading for this question**

Block 14 of the course.

Approaching the question

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.
- Weighted distances to five employment centres and proportion of non-retail business sites in neighbourhood are by far the most significant and important explanatory variables of house prices.
- Sensible discussion of model improvements.

(b) **Reading for this question**

Block 14 of the course.

Approaching the question

i. The purpose of stepwise regression is to select, from a large number of predictor variables, a small subset of variables which account for most of the variation in the dependent or criterion variable. In this procedure, the predictor variables enter or are removed from the regression equation one at a time. There are several approaches to stepwise regression.

ii. Forward selection starts with an empty model. The method computes an F statistic for each independent variable not in the model and examines the largest of these statistics. If it is significant at a specified significance level, the corresponding variable is added to the model. After a variable is entered in the model it is never removed from the model. The process is repeated until none of the remaining variables meets the specified level for entry.

Backward elimination starts off with the full model. Results of the F tests for individual parameter estimates are examined, and the least significant variable which falls above the specified significance level is removed. After a variable is removed from the model it remains excluded. The process is repeated until no other variable in the model meets the specified significance level for removal.

Stepwise selection is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. The backward component of the method removes variables from the model which do not meet the significance criteria specified to stay in the model. The stepwise selection process terminates if no further variable can be added to the model, or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Question 3

(a) **Reading for this question**

Block 15 of the course.

Approaching the question

Output interpretation should include at least 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) **Reading for this question**

Block 14 of the course.

Approaching the question

i. Multicollinearity arises when intercorrelations among the predictors are very high.

ii. Multicollinearity can result in several problems, including the following.

- The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.
- The magnitudes, as well as the signs, of the partial regression coefficients may change from sample to sample.
- It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.
- Predictor variables may be incorrectly included or removed in stepwise regression.

iii. Pearson's correlation coefficients (in the correlation matrix) could be consulted.

Question 4

(a) Reading for this question

Block 17 of the course.

Approaching the question

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) Reading for this question

Block 19 of the course.

Approaching the question

i. The basic conjoint analysis model may be represented by the following formula:

$$U(X) = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \alpha_{ij} x_{ij}$$

where:

- $U(X)$ = the overall utility of an alternative
- α_{ij} = the part-worth contribution or utility associated with the jth level ($j = 1, 2, \ldots, k_i$) of the ith attribute ($i = 1, 2, \ldots, m$)
- $x_{ij} = 1$ if the jth level of the ith attribute is present, and 0 otherwise
- k_i = the number of levels of attribute i
- m = the number of attributes.

ii. The importance of an attribute, I_i, is defined in terms of the range of the part-worths, α_{ij}, across the levels of that attribute:

$$I_i = (\max(\alpha_{ij}) - \min(\alpha_{ij})) \quad \text{for each } i.$$

The attribute's importance is normalised to ascertain its importance relative to other attributes, W_i, where:

$$W_i = \frac{I_i}{\sum\limits_{i=1}^{m} I_i} \quad \text{such that} \quad \sum_{i=1}^{m} W_i = 1.$$