SINGAPORE INSTITUTE OF MANAGEMENT

UNIVERSITY OF LONDON
PRELIMINARY EXAM 2019

MODULE CODE : ST3188

MODULE TITLE : STATISTICAL METHODS FOR MARKET RESEARCH

DATE OF EXAM :

TIME OF EXAM :

DURATION : 2 hours

TOTAL NUMBER OF PAGES (INCLUDING THIS PAGE) : 10

INSTRUCTIONS TO CANDIDATES :-

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).

The coursework mark (worth 30% of the final mark) will be combined with the examination mark (worth 70%) to give the overall percentage for the module.

A calculator may be used when answering questions on this paper and it must comply in all respects with the specification given with your Admission Notice. The make and type of machine must be clearly stated on the front cover of the answer book.

DO NOT TURN OVER THIS QUESTION PAPER UNTIL YOU ARE TOLD TO DO SO.

Candidates are strongly advised to divide their time accordingly.

SECTION A: Compulsory

1. (a) Cryptocurrencies, such as Bitcoin and Ethereum, are digital currencies which employ encryption techniques to regulate the generation of units of currency and verify the transfer of funds, operating independently of a central bank. Volatile price movements in the past year have resulted in financial regulators leading calls for cryptocurrencies to be reined in for fear of contagion risks if they enter the financial mainstream.

   Recently, some retail banks have banned customers from trading in cryptocurrencies using credit cards (due to the bank's liability in the event of default), but still permit transactions using debit cards (as a customer's own funds are used), although some banks are considering also banning the use of debit cards to trade as well to safeguard against customers realising large losses.

   One bank is considering imposing a cap on debit card transactions rather than an outright ban (for fear of losing customers to competitors). The bank has invited you to devise an appropriate sampling scheme to research attitudes of its customers regarding the introduction of a cap, including the level of any cap. Explain in detail how each of the following sampling methods could be applied to the overall sampling strategy. 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. Stratified sampling.

   iv. Cluster sampling.

   (20 marks)

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

   i. Explain what a sampling distribution is.

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

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

   iv. When constructing a confidence interval for a proportion, which level of confidence would you propose and why?

   (20 marks)

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

2. (a) A national estate agency is attempting to build a model to explain house prices. By identifying key predictor variables, the company hopes to accurately value properties and emphasise important factors when marketing properties which are put up for sale.

Data from $n = 452$ neighbourhoods were analysed using a multiple linear regression. The dependent variable is the median value of owner-occupied homes, in £000s. Seven predictor variables are initially included in the model:

- per capita crime rate of neighbourhood
- property has parking (1 = yes, 0 = no)
- nitric oxides concentration (parts per 10 million)
- proportion of homes built prior to 1940
- weighted distances to five employment centres
- index of accessibility to motorways
- proportion of non-retail business sites in neighbourhood.

Selected SPSS output is provided in Figure 1 (on the next page). Analyse the regression results, making sure you:

- Write out the full regression model, including any assumptions, and the estimated model.
- Comment on the statistical significance of the model, and the individual variables.
- 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 objective of explaining house prices.

(20 marks)

(b) i. What is the purpose of stepwise regression?

ii. Explain the stepwise regression approaches of forward selection, backward elimination and stepwise selection.

(10 marks)

Figure 1

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.541	.293	.282	74.64358

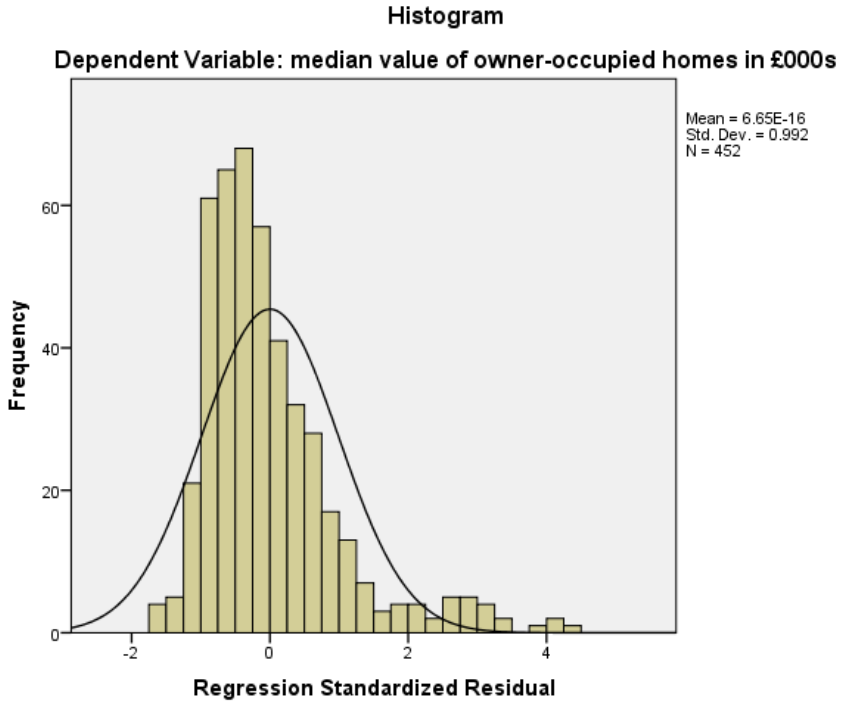
ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1025555.853	7	146507.979	26.295	.000
	Residual	2473819.138	444	5571.665		
	Total	3499374.991	451			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	483.401	36.929		13.090	.000
	per capita crime rate of neighbourhood	-6.538	3.519	-.185	-1.858	.064
	property has parking	62.074	13.308	.189	4.664	.000
	nitric oxides concentration (parts per 10 million)	-136.612	59.668	-.177	-2.290	.023
	proportion of homes built prior to 1940	-.726	.195	-.232	-3.726	.000
	weighted distances to five employment centres	-18.904	2.843	-.449	-6.650	.000
	index of accessibility to motorways	1.973	1.069	.169	1.845	.066
	proportion of non-retail business sites in neighbourhood	-5.717	.818	-.441	-6.987	.000

a. Dependent Variable: median value of owner-occupied homes in £000s



3. (a) With UK interest rates expected to rise in the years ahead from their recent historic lows, a credit card issuer is keen to develop a new model for predicting default on credit card debt, anticipating an increase in defaults as the cost of borrowing rises.

A large sample of existing customers (some of whom had previously defaulted) were modelled using discriminant analysis in order to identify useful predictor variables which could be used to predict the default status of future customers. In total, five predictors were considered:

- age in years
- total years in employment
- income in £000s
- outstanding student loans in £000s
- highest level of education (1 = school, 2 = undergraduate, 3 = postgraduate, 4 = postgraduate research).

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.
- Comment on the suitability of including the 'highest level of education' variable.
- Determine the predictive accuracy of the model.

(20 marks)

(b) i. Explain what multicollinearity is.

ii. What problems could arise as a consequence of multicollinearity?

iii. How might you detect multicollinearity?

(10 marks)

Figure 2

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.194	100.0	100.0	.403

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.838	123.228	5	.000

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age in years	.981	13.482	1	698	.000
Total years in employment	.920	60.759	1	698	.000
Income in £000s	.995	3.533	1	698	.061
Outstanding student loans in £000s	.979	15.142	1	698	.000
Highest level of education	.986	9.945	1	698	.002

Standardized Canonical Discriminant Function Coefficients

	Function
	1
Age in years	.090
Total years in employment	1.000
Income in £000s	.039
Outstanding student loans in £000s	-.860
Highest level of education	-.026

Structure Matrix

	Function
	1
Total years in employment	.670
Outstanding student loans in £000s	-.335
Age in years	.316
Highest level of education	-.271
Income in £000s	.162

Figure 2 (continued)

Pooled Within-Groups Matrices

		Age in years	Total years in employment	Income in £000s	Outstanding student loans in £000s	Highest level of education
Correlation	Age in years	1.000	.524	.475	.368	.032
	Total years in employment	.524	1.000	.627	.471	-.130
	Income in £000s	.475	.627	1.000	.629	.236
	Outstanding student loans in £000s	.368	.471	.629	1.000	.147
	Highest level of education	.032	-.130	.236	.147	1.000

Canonical Discriminant Function Coefficients

	Function
	1
Age in years	.011
Total years in employment	.157
Income in £000s	.001
Outstanding student loans in £000s	-.264
Highest level of education	-.029
(Constant)	-.901

Unstandardized coefficients

Functions at Group Centroids

	Function
Default status	1
0	.262
1	-.739

Unstandardized canonical discriminant functions evaluated at group means

Classification Results

			Predicted Group Membership		
		Default status	0	1	Total
Original	Count	0	345	172	517
		1	41	142	183
	%	0	66.7	33.3	100.0
		1	22.4	77.6	100.0
Cross-validated	Count	0	345	172	517
		1	42	141	183
	%	0	66.7	33.3	100.0
		1	23.0	77.0	100.0

4. (a) An employer conducted a survey of $n = 100$ employees to determine their satisfaction with the company's IT provision. Respondents were asked to rate the following:

- `start_time`: satisfaction with the start-time of the IT system
- `system_availability`: satisfaction with the availability of the IT system
- `performance`: satisfaction with the performance of the IT system
- `training_quality`: satisfaction with training offers and their quality
- `user_orientation`: satisfaction with the user orientation of the IT system
- `data_quality`: satisfaction with the data quality in the IT system.

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) extracted factors could be used.

(20 marks)

(b) The basic conjoint analysis model may be represented by the following formula:

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

i. Define each of the five terms $U(X)$, m, k_i, α_{ij} and x_{ij}.

ii. Explain how the relative importance of attributes can be determined from this model.

(10 marks)

Figure 3

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.686
Bartlett's Test of Sphericity	Approx. Chi-Square	98.035
	df	15
	Sig.	.000

Communalities

	Initial	Extraction
satisfaction with the start-time of the IT system	1.000	.656
satisfaction with the availability of the IT system	1.000	.356
satisfaction with the performance of the IT system	1.000	.500
satisfaction with training offers and their quality	1.000	.711
satisfaction with the user orientation of the IT system	1.000	.631
satisfaction with the data quality in the IT system	1.000	.656

Component	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	2.308	38.465	38.465
2	1.202	20.027	58.492
3	.861	14.349	72.841
4	.701	11.679	84.520
5	.469	7.810	92.330
6	.460	7.670	100.000

Rotated Component Matrix

	Component	
	1	2
satisfaction with the start-time of the IT system	.803	.102
satisfaction with the availability of the IT system	.585	.118
satisfaction with the performance of the IT system	.707	-.001
satisfaction with training offers and their quality	-.238	.809
satisfaction with the user orientation of the IT system	.342	.717
satisfaction with the data quality in the IT system	.517	.624

Component Score Coefficient Matrix

	Component	
	1	2
satisfaction with the start-time of the IT system	.439	-.081
satisfaction with the availability of the IT system	.311	-.029
satisfaction with the performance of the IT system	.403	-.135
satisfaction with training offers and their quality	-.289	.607
satisfaction with the user orientation of the IT system	.059	.434
satisfaction with the data quality in the IT system	.177	.336

Figure 3 (continued)

Reproduced Correlations

		satisfaction with the start-time of the IT system	satisfaction with the availability of the IT system	satisfaction with the performance of the IT system	satisfaction with training offers and their quality	satisfaction with the user orientation of the IT system	satisfaction with the data quality in the IT system
Reproduced Correlation	satisfaction with the start-time of the IT system	.656[a]	.482	.568	-.109	.348	.479
	satisfaction with the availability of the IT system	.482	.356[a]	.413	-.044	.285	.376
	satisfaction with the performance of the IT system	.568	.413	.500[a]	-.169	.241	.365
	satisfaction with training offers and their quality	-.109	-.044	-.169	.711[a]	.499	.381
	satisfaction with the user orientation of the IT system	.348	.285	.241	.499	.631[a]	.624
	satisfaction with the data quality in the IT system	.479	.376	.365	.381	.624	.656[a]
Residual[b]	satisfaction with the start-time of the IT system		-.152	-.211	.027	-.006	-.056
	satisfaction with the availability of the IT system	-.152		-.146	.150	-.142	-.129
	satisfaction with the performance of the IT system	-.211	-.146		.159	-.055	-.081
	satisfaction with training offers and their quality	.027	.150	.159		-.226	-.148
	satisfaction with the user orientation of the IT system	-.006	-.142	-.055	-.226		-.103
	satisfaction with the data quality in the IT system	-.056	-.129	-.081	-.148	-.103	

Extraction Method: Principal Component Analysis.

a. Reproduced communalities