UNIVERSITY OF LONDON

ST3188 Statistical methods for market research

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. Candidates are strongly advised to divide their time accordingly.

Please see questions on following page.

SECTION A: Compulsory

1. (a) High levels of inflation are thought to be putting pressure on many household budgets in what many are calling a 'cost of living crisis'. A supermarket chain has an established position in the marketplace catering to the middle classes with historically reasonable levels of disposable income. Senior management is concerned that it will lose market share to rival supermarket chains known to be discount retailers. However, supply chain issues have meant the company faces higher costs from its suppliers, affecting the capacity for the company to reduce its prices significantly.

   You work in the marketing department of the supermarket and are tasked with better understanding how your customers are affected by the cost of living crisis. You have access to the company's database of customers, who are members of its loyalty scheme.

   To better understand customers' household finances, the company's senior management has decided to use a survey of all types of customers (such as '10 items or less' shoppers, and those undertaking a weekly shop for the whole family) and has 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. Convenience sampling.

   ii. Quota 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. *In your own words*, answer the following.

   i. State a suitable estimator of the population proportion as well as its sampling distribution. Mention any assumptions which you make and define all terms used.

   ii. Explain statistically how to determine the minimum sample size necessary to estimate a population proportion to within e units, defining all terms used.

   iii. If you had to propose a confidence level when estimating a population proportion, which level would you choose, and why?

   iv. Define the terms *incidence rate* and *completion rate*, explaining how each is calculated in practice.

   (20 marks)

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

2. (a) Following an increase in customer complaints, a manager wanted to research whether any particular customer segments were the main source of complaints, or whether recent customer dissatisfaction was widespread.

   A random sample of 30 customers was surveyed, ensuring a mix of age groups and income levels. Each respondent rated their satisfaction on a 9-point Likert scale (1 = very dissatisfied, 9 = very satisfied). Respondents were classified as being either the 'under 40s' or '40 and older', and income levels were classified as being either 'low', 'middle' or 'high'.

   Analyse the selected SPSS output in Figure 1 (on the next page) 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 and interpret the results.
   - What conclusions, if any, could be drawn about customer satisfaction.

   (20 marks)

   (b) *In your own words*, answer the following. **Write a maximum of 250 words in total.**

   Discuss the reasons for the frequent use of cross-tabulations in market research. What are some of the limitations?

   (10 marks)

Figure 1

Tests for Heteroskedasticity

F Test for Heteroskedasticity[a,b,c]

F	df1	df2	Sig.
.073	1	28	.788

a. Dependent variable: Satisfaction

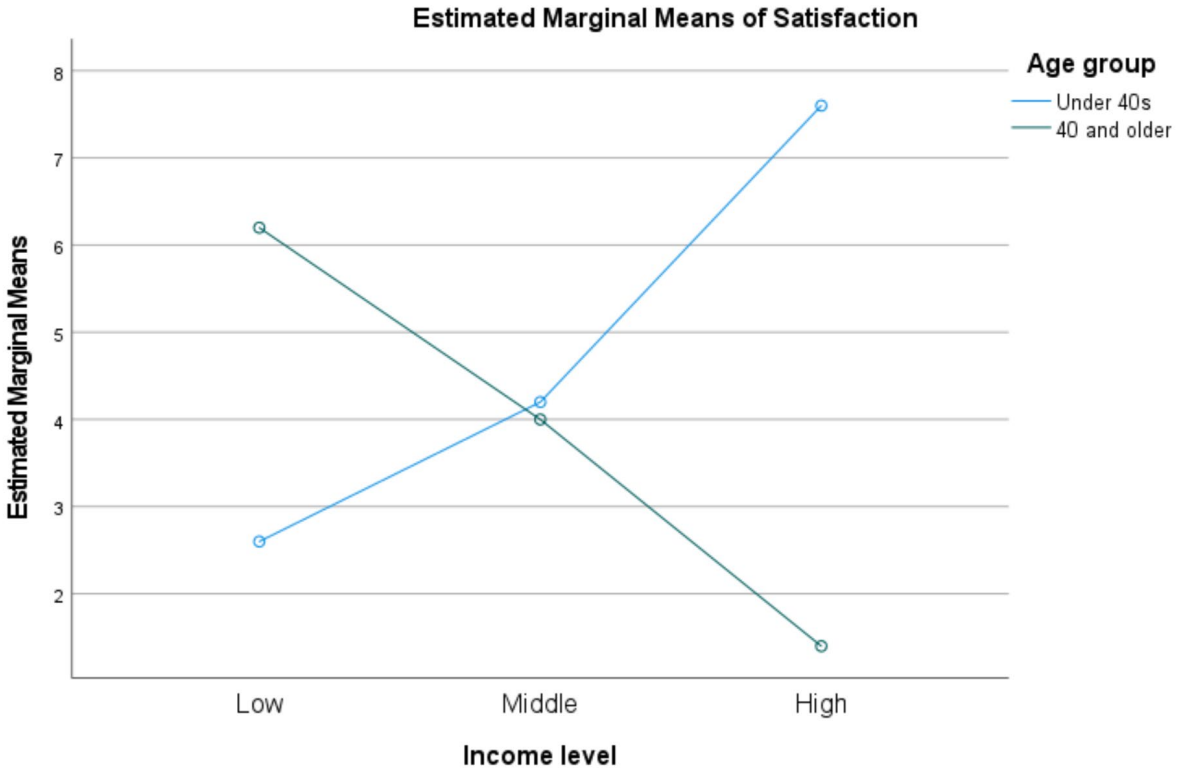
b. Tests the null hypothesis that the variance of the errors does not depend on the values of the independent variables.

c. Predicted values from design: Age_group + Income_level + Age_group * Income_level

Tests of Between-Subjects Effects

Dependent Variable: Satisfaction

Source	Type I Sum of Squares	df	Mean Square	F	Sig.
Model	692.800[a]	6	115.467	161.116	<.001
Age_group	569.867	2	284.933	397.581	<.001
Income_level	.867	2	.433	.605	.554
Age_group * Income_level	122.067	2	61.033	85.163	<.001
Error	17.200	24	.717		
Total	710.000	30			



3. (a) A toy company wishes to understand the factors influencing the sales of its most popular doll, in 000s of units. Annual data for the past 23 years have been collected and the following factors are thought to influence doll sales:

- a general upward trend over time (modelled by year number, i.e. 1 to 23)
- whether dolls were on sale at Christmas (modelled by a dummy variable)
- whether there was a recession (modelled by a dummy variable).

Selected SPSS output is provided in Figure 2 (on the next page). Analyse the regression results. In your analysis, be sure to address at least the following:

- Write out the full regression model, including any assumptions, and the estimated model.
- Comment on the performance of the model, and the significance of the independent variables.
- Discuss, with reasons, any changes you would propose making to the regression model.

(20 marks)

(b) *In your own words*, answer the following. **Write a maximum of 250 words in total.**

Explain the purpose of a *semantic differential scale* and a *Stapel scale* in market research, and provide an example of each.

(10 marks)

Figure 2

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.738[a]	.545	.473	261.479

a. Predictors: (Constant), Recession dummy, Christmas dummy, Year

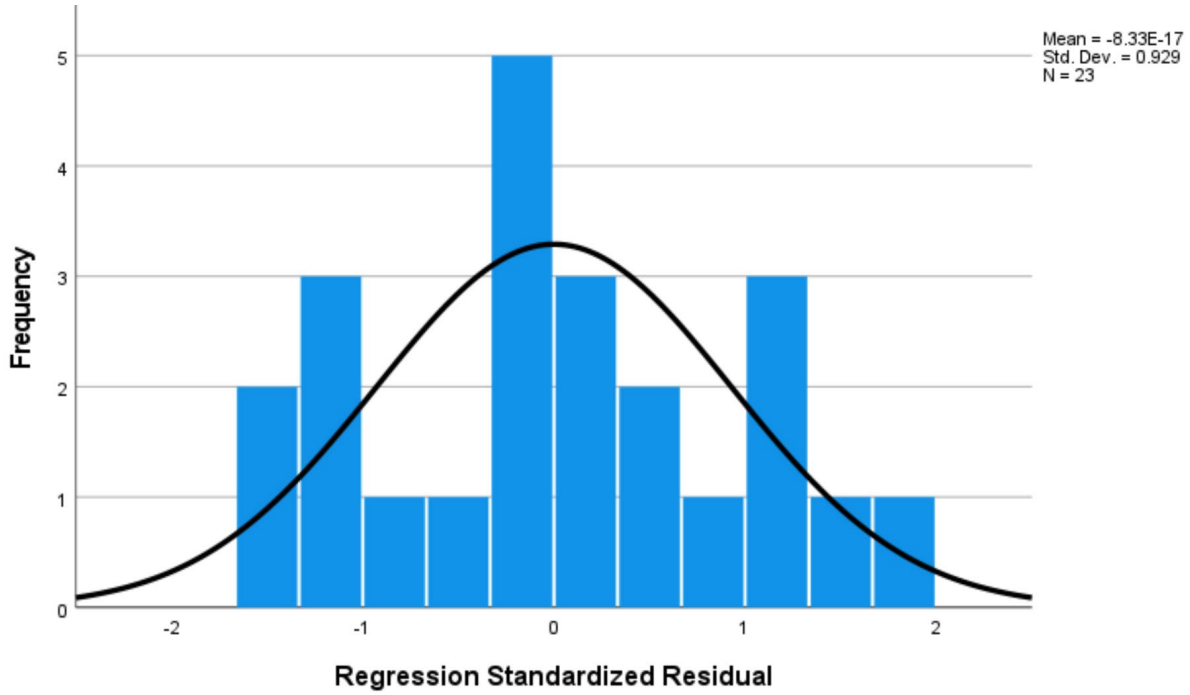
b. Dependent Variable: Dolls sold

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1555982.975	3	518660.992	7.586	.002[b]
	Residual	1299052.504	19	68371.184		
	Total	2855035.478	22			

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	120.978	143.047		.846	.408
	Year	28.100	8.365	.529	3.359	.003
	Christmas dummy	358.295	114.937	.484	3.117	.006
	Recession dummy	-90.948	112.370	-.128	-.809	.428



4. (a) A retailer is concerned about a high level of *churn* recently (i.e. losing its customers to its competitors). It has decided to use discriminant analysis to construct a model capable of predicting which existing customers may churn. It decides to use the following predictor variables (all with respect to the customer):

- Age in years.
- Months of being a customer to date.
- Household income, in thousands.
- Household size.
- Customer category (one of four, based on company engagement, classified from 'very low' to 'very high').

Analyse the selected SPSS output in Figure 3 (spread over the next two pages) and discuss what conclusions can be drawn from the data. Keep in mind the retailer's desire to predict which customers are likely to churn. 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 customer category variable.
- Determine the predictive accuracy of the model.

(20 marks)

(b) *In your own words*, answer the following. **Write a maximum of 250 words in total.**

Explain the purpose of *projective techniques* in market research, and describe their main characteristics.

(10 marks)

Figure 3

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
Age in years	.905	26.001	1	248	<.001
Months of being a customer to date	.926	19.676	1	248	<.001
Household income, in thousands	.981	4.766	1	248	.030
Household size	.993	1.770	1	248	.185
Customer category	.999	.272	1	248	.603

Pooled Within-Groups Matrices

		Age in years	Months of being a customer to date	Household income, in thousands	Household size	Customer category
Correlation	Age in years	1.000	.362	.369	-.289	.043
	Months of being a customer to date	.362	1.000	.263	.036	.234
	Household income, in thousands	.369	.263	1.000	-.120	.174
	Household size	-.289	.036	-.120	1.000	.190
	Customer category	.043	.234	.174	.190	1.000

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	.137[a]	100.0	100.0	.347

a. First 1 canonical discriminant functions were used in the analysis.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.880	31.459	5	<.001

Standardized Canonical Discriminant Function Coefficients

	Function 1
Age in years	.676
Months of being a customer to date	.534
Household income, in thousands	-.011
Household size	-.043
Customer category	-.054

Figure 3 (continued)

Structure Matrix

	Function 1
Age in years	.876
Months of being a customer to date	.762
Household income, in thousands	.375
Household size	-.228
Customer category	.090

Canonical Discriminant Function Coefficients

	Function 1
Age in years	.054
Months of being a customer to date	.026
Household income, in thousands	.000
Household size	-.032
Customer category	-.048
(Constant)	-3.077

Unstandardized coefficients

Functions at Group Centroids

Churn within last month	Function 1
No	.193
Yes	-.702

Unstandardized canonical discriminant functions evaluated at group means

Classification Results[a,c]

		Churn within last month	Predicted Group Membership: No	Yes	Total
Original	Count	No	119	77	196
		Yes	15	39	54
	%	No	60.7	39.3	100.0
		Yes	27.8	72.2	100.0
Cross-validated[b]	Count	No	117	79	196
		Yes	15	39	54
	%	No	59.7	40.3	100.0
		Yes	27.8	72.2	100.0