**THIS PAPER IS NOT TO BE REMOVED FROM THE EXAMINATION HALL**

UNIVERSITY OF LONDON

**ST3188 ZB**

**BSc DEGREES AND GRADUATE DIPLOMAS IN ECONOMICS, MANAGEMENT, FINANCE AND THE SOCIAL SCIENCES, THE DIPLOMA IN ECONOMICS AND SOCIAL SCIENCES AND THE CERTIFICATE IN EDUCATION IN SOCIAL SCIENCES**

**Statistical Methods for Market Research**

Tuesday 7 May 2019 : 14:30 – 16:30

Time allowed: 2 hours

**DO NOT TURN OVER UNTIL TOLD TO BEGIN**

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.

**[Calculator]**

**SECTION A: Compulsory**

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)**

   (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)**

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

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 |
|---|---|---|---|---|
| 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 |

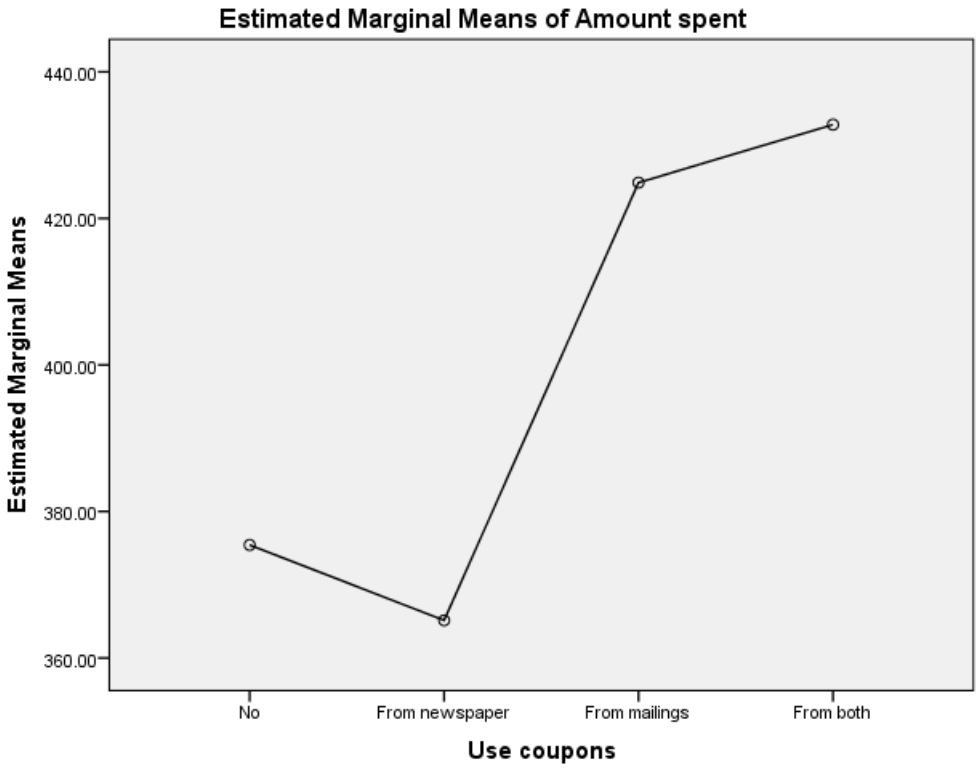
**Tests of Between-Subjects Effects**

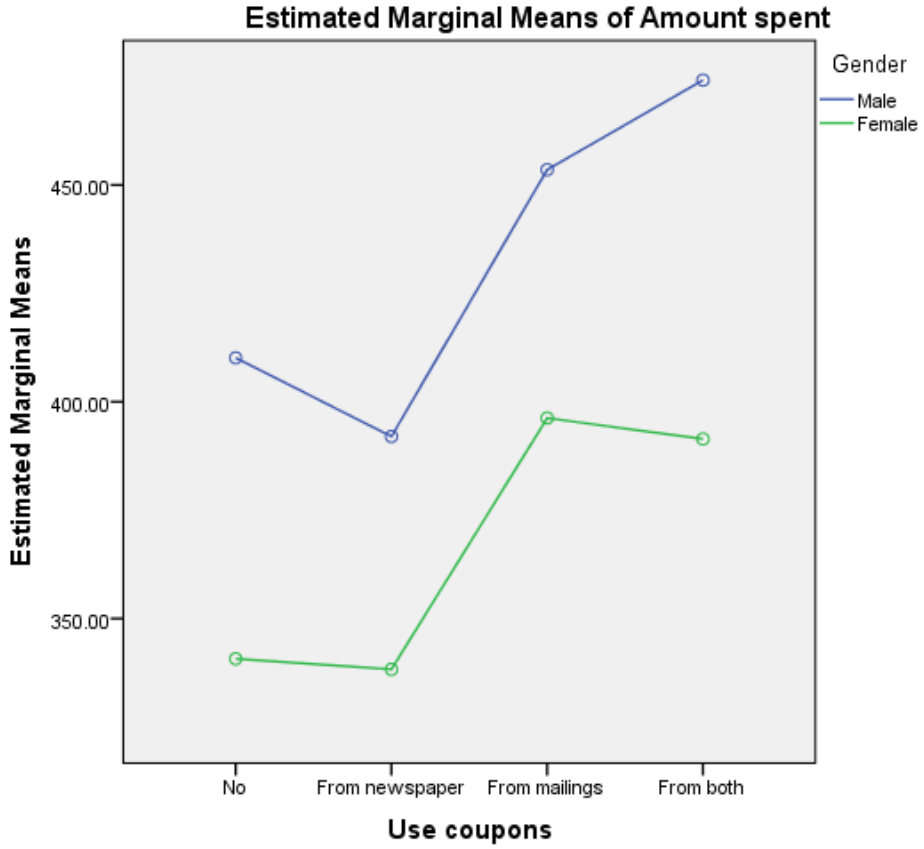
Dependent Variable: Amount spent

| 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)**



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: No | Predicted Group Membership: 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 |

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**

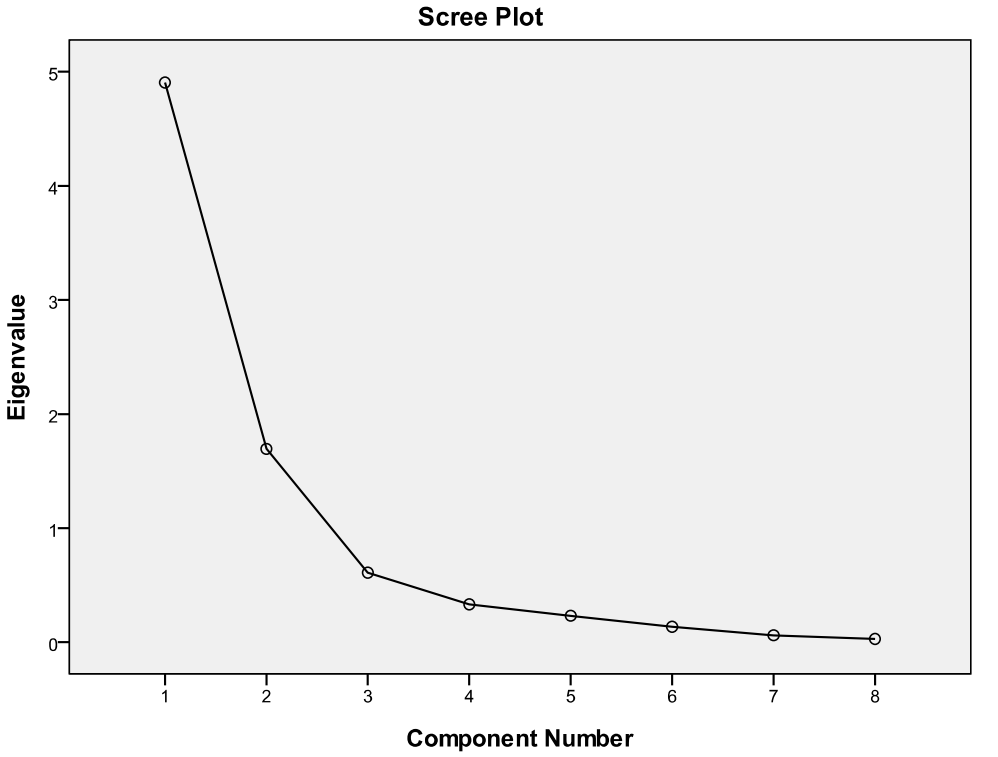
| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .792 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 1051.073 |
| | df | 28 |
| | Sig. | .000 |

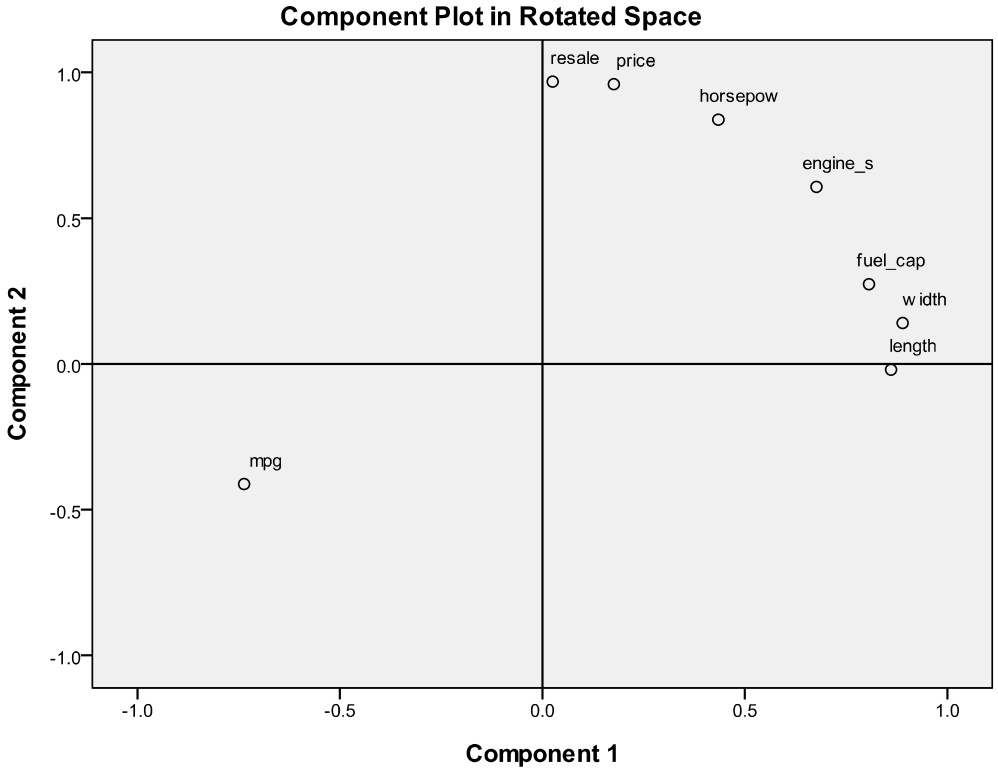
| Component | Initial Eigenvalues | | |
|---|---|---|---|
| | Total | % of Variance | Cumulative % |
| 1 | 4.905 | 61.310 | 61.310 |
| 2 | 1.694 | 21.181 | 82.491 |
| 3 | .610 | 7.625 | 90.117 |
| 4 | .332 | 4.148 | 94.264 |
| 5 | .232 | 2.906 | 97.170 |
| 6 | .136 | 1.699 | 98.870 |
| 7 | .061 | .765 | 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)**




END OF PAPER