



UNIVERSITY *of* LIMERICK
OLLSCOIL LUIMNIGH

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS & STATISTICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4128

SEMESTER: Spring 2010/11

MODULE TITLE: Advanced Data Modelling

DURATION: 3 hours

LECTURER: Dr. Ailish Hannigan

EXTERNAL

EXAMINER: Prof. B. Murphy

INSTRUCTIONS TO CANDIDATES:

This exam is worth 70% of your final grade.

The data is provided on a disk. Your answers should be submitted in the form of a Word document containing relevant PASW output. The Word document and final PASW data file should be saved on the disk provided and should also be emailed to ailish.hannigan@ul.ie. Put your name and ID number on the disk.

MA4128

Advanced Data Modelling

MA4128 is a module for Business students taking a minor option in data analysis. The module is taught as a computer based module and the end of semester exam is computer based.

MODULE AIMS/OBJECTIVES

To familiarise students with the main techniques of multivariate data analysis and generalised linear models and to enable students to implement these methods on real data via computer packages.

SHORT SYLLABUS

Cluster analysis, principle component analysis, factor analysis, discriminant analysis, the generalised linear model, logistic regression.

PRIME TEXTS

Prime Texts (1) : Afifi, A.A and Clark, V. Sharma, S. (2003) Computer Aided multivariate analysis (4th edition) Chapman and Hall.

The file *nutrition.sav* gives data from a survey on attitudes to food from a sample of 89 adults. Respondents used a 5 point scale (1=strongly disagree to 5=strongly agree) to represent their views on each statement below.

List of statements:

<i>advice:</i>	I often seek out the advice of my friends regarding which brand of food to buy.
<i>influence:</i>	I sometimes influence what foods my friends buy.
<i>sales:</i>	I usually watch the advertisements for announcements of sales.
<i>friends:</i>	I spend a lot of time talking with my friends about products and brands of food.
<i>specials:</i>	I shop a lot for "specials".
<i>newrecipe:</i>	I often try new recipes.
<i>goodcook:</i>	I am a good cook.
<i>lovecook:</i>	I love to cook.
<i>bake:</i>	I love to bake and frequently do.
<i>neighbours:</i>	My friends and neighbours often come to me for advice about food.
<i>pricesmall:</i>	I find myself checking the prices in the supermarket even for small items.
<i>goodadvice:</i>	My friends usually give me good advice on what foods to buy in the supermarket.

The *age* (in years) of the respondent and *gender* (where 1=male and 2= female) are also given in the dataset.

Each respondent was also classified into the variable *weight* where 1=overweight and 0=not overweight. The objective of the analysis of the data is to distinguish between respondents who are overweight or not based on their attitudes to food.

- (a) Carry out a **bivariate exploratory data analysis**, briefly describing the relationship between the data collected on **each** of the variables for the respondents and the variable *weight*.
[7 marks]
- (b) Distinguish between the objectives of an **exploratory factor analysis** and the objectives of a **confirmatory factor analysis**. What is the most commonly used method of extraction for an exploratory factor analysis? What is the most commonly used method of extraction for a confirmatory factor analysis?
[6 marks]
- (c) Obtain the **correlation matrix** for the variables *advice*, *influence*, *sales*, *friends*, *specials*, *newrecipe*, *goodcook*, *lovecook*, *bake*, *neighbours*, *pricesmall* and *good advice* and briefly **comment** on this matrix. Carry out an **exploratory factor analysis** using these variables, decide on your final solution and interpret the output from this analysis. Investigate the relationship between the factors scores for the first factor extracted and the variables *age* and *gender*.
[18 marks]

- (d) Distinguish between **hierarchical** and **non-hierarchical** clustering techniques. Distinguish between **agglomerative** and **divisive** hierarchical clustering techniques.

[4 marks]

- (e) Use the variables *advice*, *influence*, *sales*, *friends*, *specials*, *newrecipe*, *goodcook*, *lovecook*, *bake*, *neighbours*, *pricesmall*, *goodadvice*, *age* and *gender* to cluster the respondents using a **two step cluster** analysis. Why was a two step cluster analysis used here? Interpret the results of the analysis. Cross-tabulate the cluster membership variable with the *weight* variable and comment.

[12 marks]

- (f) The objective of the data analysis is to establish which of the independent variables *age*, *gender* and the factors extracted from part (c) best predict the variable *weight* i.e. whether a respondent is overweight or not. Choose an appropriate model for this problem, justifying your choice. Fit a model between *weight* and each of the independent variables **on its own**. Give an interpretation of the importance of each independent variable on its own.

[12 marks]

- (g) Decide on a final model (justifying your choice) and interpret the output from your model. Give a measure of goodness of fit for your final model, identify and describe any outliers and comment on the overall classification ability of the model.

[11 marks]

Solutions to MA4128 2010/11

- (a) Table 1 gives the median (IQR) for agreement with each of the attitude statements for those who were overweight and those who weren't.

Table 1

	Overweight	Not overweight
advice	2 (1)	3 (2)
influence	3 (2)	3.5 (2)
sales	4 (1)	4 (0)
friends	2 (1)	3 (2)
specials	3 (2)	4 (0)
newrecipe	3 (1)	3 (1)
goodcook	4 (1)	4 (1)
lovecook	3 (2)	4 (1)
bake	3 (2)	4 (1)
neighbours	2 (1)	3 (1)
pricesmall	4 (1)	4 (0)
goodadvice	2 (1)	3 (2)

Respondents who were not overweight tended to agree more with the positively worded statements about food. They also tended to be younger with a mean age of 26 years (SD=5.04) compared to 30 years (SD=5.74) for those who were overweight. 21 (54%) of the male respondents were overweight compared to 20 (40%) of the 50 female respondents.

(b)

In **exploratory factor analysis** the structure of the factor model or the underlying theory is not known before the analysis is carried out. The data are used to help identify the underlying structure and build the theory and the a priori assumption is that any indicator variable may be associated with any factor. In **confirmatory factor analysis** the structure of the factor model, based on some underlying theory, is known beforehand. We want to establish if the data fit this model and test the underlying theory. Indicator variables are selected on the basis of prior theory and the factor analysis is used to see if they load as predicted on the expected number of factors. A minimum requirement of confirmatory factor analysis is that the number of factors in the model is known beforehand, but usually expectations about which variables will load on which factors are also made.

Exploratory factor analysis typically uses the **correlation matrix** whereas the theory for confirmatory factor analysis is derived for **covariance matrices**. Principal components analysis is the most commonly used method of extraction for exploratory factor analysis whereas principal axis factoring is used for confirmatory factor analysis. Confirmatory factor analysis forms part of covariance structure models i.e. **Structural Equation Modelling (SEM)**.

(c) Correlation matrix shows no evidence of very strong correlations. Variables show weak to moderate relationship with each other. No variable in the dataset which is uncorrelated to any of the other variables.

Correlations

[illegible]

goodcook	Pearson	-.156	.020	.084	.177	-.010	.400**	1	.485**	.383**	.290**	.405**	-.075
	Correlation												
	Sig. (2-tailed)	.145	.853	.431	.097	.928	.000		.000	.000	.006	.000	.485
N		89	89	89	89	89	89	89	89	89	89	89	89
lovecook	Pearson	-.144	.053	.211*	.343**	.184	.451**	.485**	1	.606**	.378**	.258*	.091
	Correlation												
	Sig. (2-tailed)	.179	.623	.047	.001	.084	.000	.000		.000	.000	.015	.395
N		89	89	89	89	89	89	89	89	89	89	89	89
bake	Pearson	-.022	.093	.166	.265*	.122	.439**	.383**	.606**	1	.385**	.151	.211*
	Correlation												
	Sig. (2-tailed)	.839	.388	.121	.012	.256	.000	.000	.000		.000	.157	.047
N		89	89	89	89	89	89	89	89	89	89	89	89
neighbours	Pearson	.086	.203	.259*	.262*	.255*	.387**	.290**	.378**	.385**	1	.188	.253*
	Correlation												
	Sig. (2-tailed)	.425	.057	.014	.013	.016	.000	.006	.000	.000		.077	.017
N		89	89	89	89	89	89	89	89	89	89	89	89
pricesmall	Pearson	-.042	.015	.469**	.279**	.400**	.219*	.405**	.258*	.151	.188	1	.008
	Correlation												
	Sig. (2-tailed)	.694	.885	.000	.008	.000	.039	.000	.015	.157	.077		.944
N		89	89	89	89	89	89	89	89	89	89	89	89
goodadvice	Pearson	.475**	.506**	.211*	.426**	.098	.252*	-.075	.091	.211*	.253*	.008	1
	Correlation												
	Sig. (2-tailed)	.000	.000	.047	.000	.359	.017	.485	.395	.047	.017	.944	
N		89	89	89	89	89	89	89	89	89	89	89	89

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Exploratory factor analysis used principal components analysis as the method of extraction with a varimax rotation. Three components were extracted which explained 61% of the total variance.

Rotated Component Matrix^a

	Component		
	1	2	3
advice	-.191	.800	.005
influence	.085	.791	-.052
sales	.117	.148	.829
friends	.375	.529	.284
specials	-.031	.035	.856
newrecipe	.742	.147	.008
goodcook	.727	-.179	.069
lovecook	.798	-.023	.170
bake	.752	.108	.056
neighbours	.540	.256	.246
pricesmall	.293	-.079	.688
goodadvice	.138	.808	.075

Extraction Method: Principal Component Analysis.

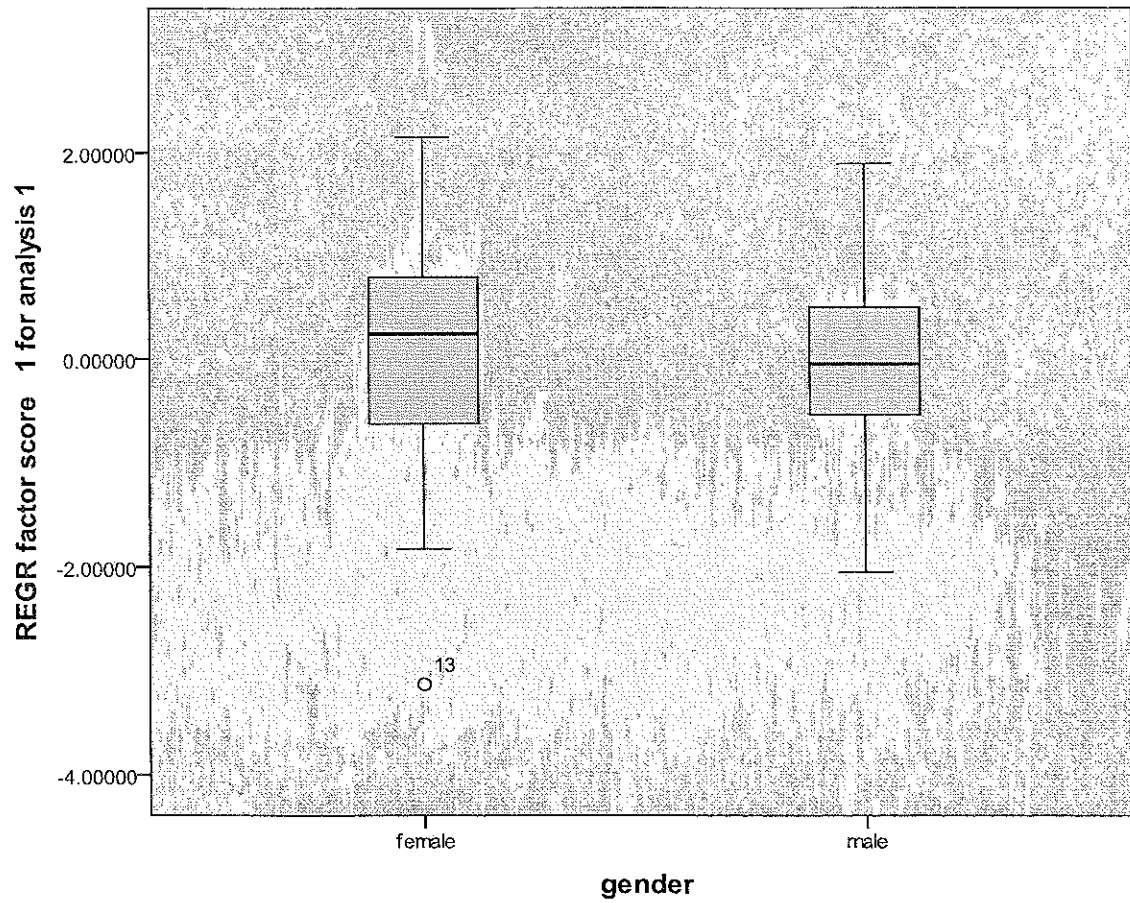
Rotation Method: Varimax with Kaiser

Normalization.

a. Rotation converged in 4 iterations.

The first factor may represent attitudes towards cooking, the second advice about food and the third shopping for food.

Factor 1 scores were slightly more positive for females than males with slightly more variation for females and one outlier (see boxplot below). There was a weak negative correlation (-0.28) between age and factor 1 scores.





(d) Clustering techniques fall into two categories: **hierarchical** and **non-hierarchical**. Hierarchical methods can be either agglomerative (“bottom up”) or divisive (“top down”). In agglomerative methods we begin with n clusters where n is the number of observations, i.e. each observation is in a cluster by itself. We combine the two closest clusters, thus reducing the number of clusters by 1 and we repeat this step. In the last step all observations are grouped into one cluster. In divisive methods we begin with one cluster containing all observations. We split off the observations that are most dissimilar to the remaining ones and repeat this step until all observations are in a cluster of their own. Agglomerative methods are the most commonly used. K-means clustering is the most commonly used non-hierarchical clustering technique. It is used when the number of clusters is known or suggested before the analysis is carried out.

(e) Two step cluster analysis was used because of the mixture of categorical and quantitative variables. Two clusters were identified with cluster quality rated as poor. Clusters were similar sizes (41 and 48 observations) and the most important variable in identifying the clusters was gender. All of the females are in cluster 2 and 95% of the males are in cluster 1.

Clusters

Feature Importance

■ 1.0 ■ 0.8 ■ 0.6 ■ 0.4 ■ 0.2

Cluster	2	1
Label		
Description		
Size	 53.9% (46)	 46.1% (41)
Features	normal (mean: 0.0000)	normal (mean: 0.0000)
	specials 3.83	specials 3.27
	neighbours 2.90	neighbours 2.44
	friends 2.95	friends 2.46
	sales 4.04	sales 3.73
	bake 3.40	bake 3.05
	lovecook 3.65	lovecook 3.34
	pricesmall 4.17	pricesmall 3.95
	newrecipe 3.08	newrecipe 2.88
	influence 2.88	influence 3.10
	advice 2.65	advice 2.41
	age 27.29	age 28.41
	goodadvice 2.79	goodadvice 2.68
	goodcook 3.83	goodcook 3.80

numeric variav

weight * TwoStep Cluster Number Crosstabulation

			TwoStep Cluster Number		Total
			1	2	
weight	normal weight	Count	18	30	48
		% within weight	37.5%	62.5%	100.0%
		% within TwoStep Cluster	43.9%	62.5%	53.9%
		Number			
	overweight	Count	23	18	41
		% within weight	56.1%	43.9%	100.0%
		% within TwoStep Cluster	56.1%	37.5%	46.1%
		Number			
Total	Count	41	48	89	
	% within weight	46.1%	53.9%	100.0%	
	% within TwoStep Cluster	100.0%	100.0%	100.0%	
	Number				

More overweight respondents were in cluster 1.

(f) 46.1% of the respondents were overweight. Logistic regression was used to model the relationship between the binary outcome, weight, and each of the independent variables (gender – categorical, age, factor 1-3 – quantitative). All were statistically significant predictors ($p < 0.001$) of weight except for gender.

A one unit increase in factor scores for factor 1,2 and 3 decreases the odds of being overweight. A one unit increase in age increases the odds of being overweight.

The highest percentage correctly classified (74%) is for the model with factor 1 as the independent variable.

(g) Output for the final model is given below. The Hosmer and Lemeshow test is a goodness-of-fit test. The null hypothesis is that there is no significant difference between the observed and predicted values of the dependent variable implying the model is a “good” fit for the data. We do not reject the null hypothesis ($p = 0.336$) so the model is a reasonable fit for the data.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.077	8	.336

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	60.054 ^a	.506	.676

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Pseudo R squared is 67.6%.

Classification Table^a

Observed			Predicted		
			weight		Percentage Correct
			normal weight	overweight	
Step 1	weight	normal weight	42	6	87.5
		overweight	7	34	82.9
Overall Percentage					85.4

a. The cut value is .500

85.4% of the respondents were correctly classified. The percentage correctly classified was similar for both overweight and normal weight categories.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
								Lower	Upper
Step 1 ^a	FAC1_1	-1.982	.521	14.477	1	.000	.138	.050	.382
	FAC2_1	-1.374	.420	10.722	1	.001	.253	.111	.576
	FAC3_1	-1.995	.554	12.987	1	.000	.136	.046	.402
	age	.225	.070	10.435	1	.001	1.252	1.092	1.435
	Constant	-6.222	1.936	10.332	1	.001	.002		

a. Variable(s) entered on step 1: FAC1_1, FAC2_1, FAC3_1, age.

All of the independent variables are statistically significant predictors of the binary outcome, weight. Increases in factor 1,2,3 scores decrease the odds of being overweight. Increase in age increases the odds of being overweight.

Histogram of the standardized residuals is positively skewed with two outliers .

Histogram

