SMM634 Coursework 1 2018/2019

Deadline 12 March 2019

NAME:

1. Dataset `cars.dat` available on Moodle consists of 388 cars from the 2006 model year, with 18 features. Seven features are binary indicators; the other 11 features are numerical (see Table 1)

Variable	Meaning
Sports	Binary indicator for being a sports car
SUV	Indicator for sports utility vehicle
Wagon	Indicator
Minivan	Indicator
Pickup	Indicator
AWD	Indicator for all-wheel drive
RWD	Indicator for rear-wheel drive
Retail	Suggested retail price (US$)
Dealer	Price to dealer (US$)
Engine	Engine size (liters)
Cylinders	Number of engine cylinders
Horsepower	Engine horsepower
CityMPG	City gas mileage
HighwayMPG	Highway gas mileage
Weight	Weight (pounds)
Wheelbase	Wheelbase (inches)
Length	Length (inches)
Width	Width (inches)

Table 1: Features for the `cars` dataframe.

Perform a principal component analysis and report your conclusions. (Not more than 250 words, excluding R codes, plots, tables etc.).

[Total marks: 10]

2. On the same dataset (`cars.dat`) use classical multidimensional scaling of dimension 2 based on the Euclidean distance matrix, comment on your results and compare with those obtained using the PCA above. (Not more than 250 words, excluding R codes, plots, tables etc.)

[Total marks: 10]

3. A factor analysis was carried out of a data matrix of variables relating to the occupational and educational status of three generations of family members. A description of the ten variables used in the analysis is given in Table 2.

Variable	Generation	Code	Description
x_1	1	HF/O	Husband's father's occupational status
x_2	1	WF/O	Wife's father's occupational status
x_3	2	H/FE	Husband's further education
x_4	2	H/Q	Husband's qualifications
x_5	2	H/O	Husband's occupational status
x_6	2	W/FE	Wife's further education
x_7	2	W/Q	Wife's qualifications
x_8	3	FB/FE	Firstborn's further education
x_9	3	FB/Q	Firstborn's qualifications
x_{10}	3	FB/O	Firstborn's occupational status

Table 2: Descriptions of social mobility variables.

(a) The unrotated factor loadings obtained from the three-factor model are given in Table 3. Interpret them. [Marks: 5]

		$\hat{\alpha}_{i1}$	$\hat{\alpha}_{i2}$	$\hat{\alpha}_{i3}$
x_1	HF/O	0.426	0.403	0.053
x_2	WF/O	0.404	0.343	0.008
x_3	H/FE	0.592	-0.026	0.116
x_4	H/Q	0.558	-0.240	0.118
x_5	H/O	0.575	0.481	0.031
x_6	W/FE	0.451	-0.126	0.369
x_7	W/Q	0.477	-0.296	0.462
x_8	FB/FE	0.615	-0.191	-0.289
x_9	FB/Q	0.519	-0.358	-0.381
x_{10}	FB/O	0.602	0.168	-0.219

Table 3: Loading matrix giving the unrotated loadings from a three-factor model of the social mobility data.

(b) Rotations can be carried out to determine whether simple structure can be achieved. The factor loadings obtained from an orthogonal (varimax) rotation and an oblique (oblimin) rotation of the three-factor solution are shown in Tables 4 and 5. Comment on the results. [Marks: 5]

(Not more than 250 words.)

[Total marks: 10]

		$\hat{\alpha}_{i1}$	$\hat{\alpha}_{i2}$	$\hat{\alpha}_{i3}$
x_1	HF/O	0.576	0.042	0.111
x_2	WF/O	0.516	0.086	0.090
x_3	H/FE	0.329	0.288	0.416
x_4	H/Q	0.135	0.360	0.485
x_5	H/O	0.728	0.113	0.144
x_6	W/FE	0.163	0.078	0.568
x_7	W/Q	0.042	0.106	0.718
x_8	FB/FE	0.209	0.645	0.194
x_9	FB/Q	0.018	0.723	0.140
x_{10}	FB/O	0.491	0.434	0.098

Table 4: Loading matrices giving the varimax rotated loadings from a three-factor model of the social mobility data

		$\hat{\alpha}_{i1}$	$\hat{\alpha}_{i2}$	$\hat{\alpha}_{i3}$
x_1	HF/O	0-0.064	0.599	0.025
x_2	WF/O	-0.003	0.530	0.002
x_3	H/FE	0.183	0.246	0.353
x_4	H/Q	0.279	0.015	0.445
x_5	H/O	-0.016	0.747	0.025
x_6	W/FE	-0.051	0.074	0.585
x_7	W/Q	-0.032	-0.085	0.765
x_8	FB/FE	0.637	0.101	0.058
x_9	FB/Q	0.762	-0.109	0.014
x_{10}	FB/O	0.381	0.452	-0.052

Table 5: Loading matrices giving the oblimin rotated loadings from a three-factor model of the social mobility data.

4. Let's consider the `USArrests` data (see Lecture 1). The dataset contains 4 variables that provide information on the numbers of arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

   (a) Calculate the Euclidean distance matrix on the four variables. [Marks: 2]

   (b) Apply a hierarchical clustering analysis, with average linkage and plot the results. [Marks: 2]

   (c) Use function `cutree()` to specifying the desired number of clusters, for instance 4 and plot the results. What does the above function return? [Marks: 2]

   (d) Which states are grouped in cluster 1? [Marks: 2]

   (e) Perform the same analysis on standardized data. Do results change? [Marks: 2]

[Total marks: 10]