Examen de Análisis Inteligente de Datos

Junio 2016

Parte Teórica

I- Analice la veracidad de las siguientes proposiciones justifique en un renglón.

1) El análisis de componentes principales solo es valido si se ha rechazado la hipótesis de independencia.

2) Si se rechaza la hipótesis nula de una prueba de independencia significa que la variable tiene una distribución diferente en cada una de las poblaciones.

3) Dada una matriz de datos de 1800 registros de 7 variables cada uno, la matriz de varianzas y covarianzas tiene 49 valores a estimar.

4) El análisis de correspondencias múltiples se puede aplicar solo en el caso que la inercia sea grande.

5) El nombre “biplot” en el contexto de compontes principales, obedece al hecho de que permite ver al mismo tiempo las observaciones y las componentes.

6) Una componente principal se dice de tamaño cuando sus coeficientes tienen todo el mismo signo.

7) Si en una tabla de contingencia no se ha rechazado la hipótesis nula del test de independencia de Chi-Cuadrado, es válido realizar un Análisis de Correspondencias.

8) En el análisis de componentes principales algunas variables explican el comportamiento de otras.

9) En un biplot, si dos variables están negativamente correlacionadas entonces las flechas que las identifican aparecen ortogonales (perpendiculares) entre si.

10) Se dice que un valor es un outlier cuando se encuentra a más de dos desvíos standard de la media de la variable.

II- Responda las siguientes preguntas (máximo tres renglones por pregunta):

a) Que permite observar el gráfico de mosaicos?. Y el gráfico de caritas de Chernov?.

b) Explique el criterio del bastón roto y ejemplifique.

c) Explique el concepto de perfil fila y perfil medio en el contexto de análisis de correspondencias simples.

d) En qué caso se quedaría con una sola componente en análisis de correspondencias?

e) Defina los conceptos de traza y determinante en función de los autovalores de una matriz. Explique qué relación puede establecerse con el análisis de componentes principales y el análisis de correspondencias.

Parte Prática

Ejercicio 1

Muestra de 41 ciudades de USA donde se midieron diferentes variables relacionadas con la contaminación atmosférica.

Las variables son:
(SO2): Contenido en SO2
(Temp): Temperatura anual en grados F
(Emp): Número de empresas mayores de 20 trabajadores
(Pob): Población (en miles de habitantes) (Viento): Velocidad media del viento
(Precipt): Precipitación anual media
(Días): Días lluviosos al año

Se condujo un análisis de componentes principales y los resultados fueron los siguientes:

Análisis de componentes principales

Datos estandarizados

Variables de clasificación

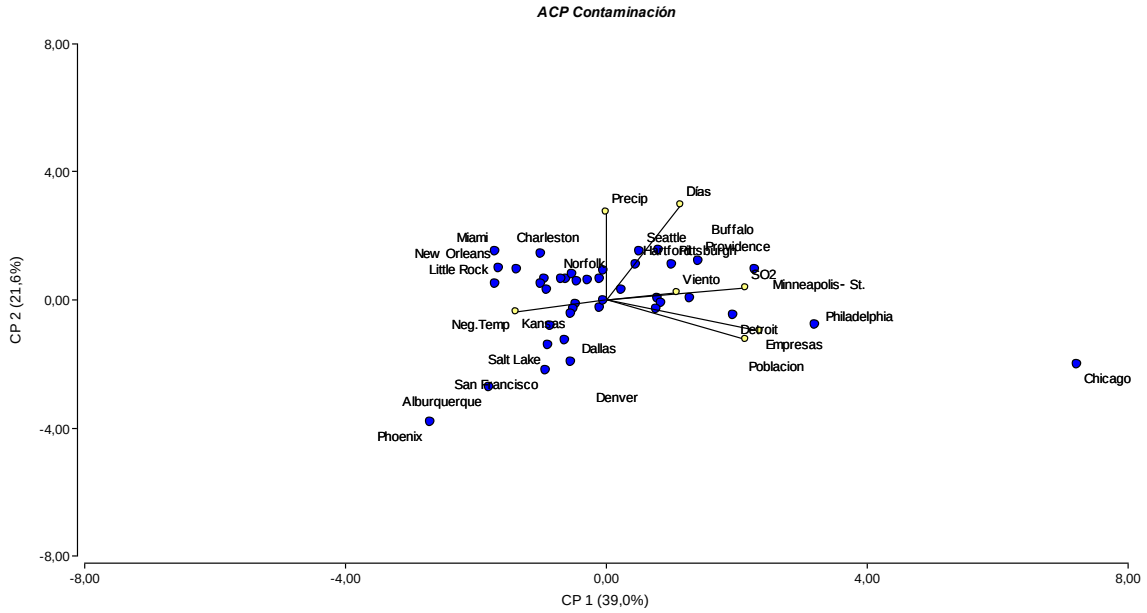
Ciudad

Autovalores

Lambda	Valor	Proporción	Prop Acum
1	2,73	0,39	0,39
2	1,51	0,22	0,61
3	1,39	0,20	0,81
4	0,89	0,13	0,93
5	0,35	0,05	0,98
6	0,10	0,01	1,00
7	0,03	3,6E-03	1,00

Autovectores

Variables	e1	e2
SO2	0,49	0,08
Neg.Temp	-0,32	-0,09
Empresas	0,54	-0,23
Población	0,49	-0,28
Viento	0,25	0,06
Precip	1,9E-04	0,63
Días	0,26	0,68

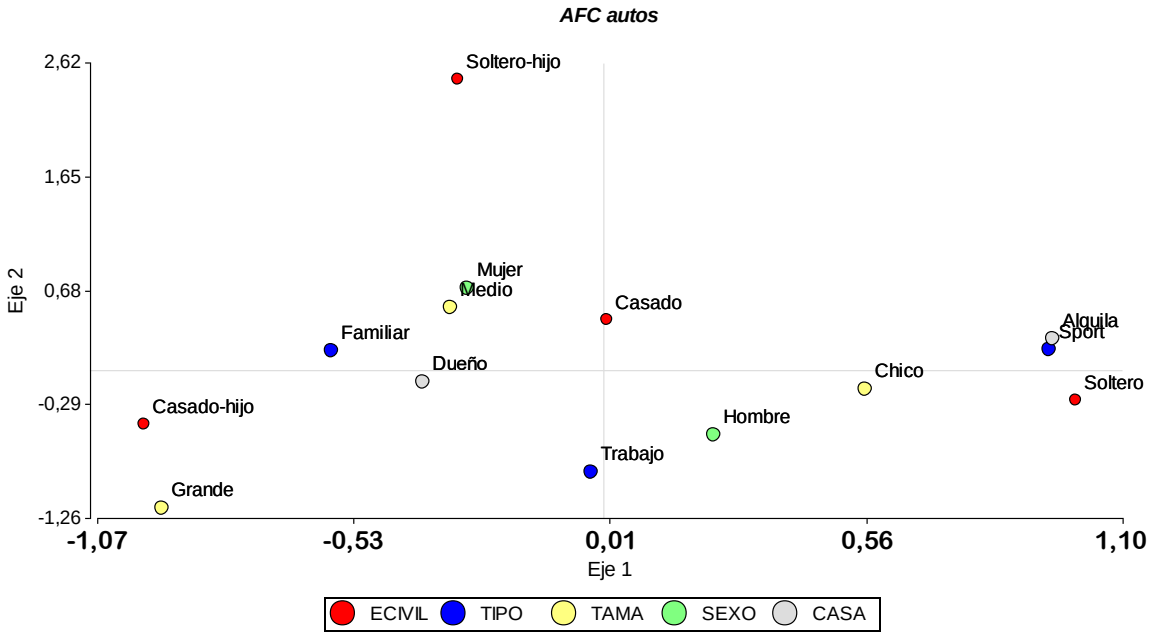


En base a las salidas, se pide responder a las siguientes preguntas:

a- Qué porcentaje de la variabilidad logra explicar cada componente principal?
b- Cuantas componentes elegiría? Indique en qué criterio se basaría?.
c- Indique si cada una de las componentes es de tamaño o de forma?
d- Señale dos ciudades similares y dos ciudades muy distintas.
e- Señale una ciudad promedio y una muy diferente a las demás.
f- Como interpretaría la primer componente principal?

Ejercicio 2

Se desea analizar la preferencia de autos de usuarios de EEUU. Se dispone de una base de datos con estado civil, sexo, ingresos y tipo de auto comprado.



Contribución a la Chi cuadrado

	Autovalor	Inercias	Chi-Cuadrado	(%)	% acumulado
1	0,60	0,36	701,37	20,16	20,16
2	0,51	0,26	493,76	14,19	34,35

Coordenadas fila

	Eje 1	Eje 2
Casado-hijo	-0,97	-0,46
Soltero	1,00	-0,26
Casado	0,01	0,42
Soltero-hijo	-0,31	2,47
Familiar	-0,57	0,17
Sport	0,94	0,17
Trabajo	-0,02	-0,87
Grande	-0,93	-1,18
Chico	0,55	-0,17
Medio	-0,32	0,53
Hombre	0,23	-0,56
Mujer	-0,29	0,69
Dueño	-0,38	-0,11
Alquila	0,95	0,27

	Casado-hijo	Soltero	Casado	Soltero-hijo	Familiar	Sport	Trabajo	Grande	Chico	Medio	Hombre	Mujer	Dueño	Alquila
Casado-hijo	111	0	0	0	81	12	18	22	37	52	61	50	106	5

Soltero	0	112	0	0	35	60	17	11	60	40	76	36	52	60
Casado	0	0	101	0	50	35	16	9	50	42	48	53	76	25
Soltero-hijo	0	0	0	15	10	2	3	1	6	8	2	13	8	7
Familiar	81	35	50	10	176	0	0	31	55	90	91	85	130	46
Sport	12	60	35	2	0	109	0	1	68	39	64	45	71	38
Trabajo	18	17	16	3	0	0	54	11	30	13	32	22	41	13
Grande	22	11	9	1	31	1	11	43	0	0	25	18	35	8
Chico	37	60	50	6	55	68	30	0	153	0	91	62	101	52
Medio	52	40	42	8	90	39	13	0	0	142	71	71	106	36
Hombre	61	76	48	2	91	64	32	25	91	71	187	0	128	59
Mujer	50	36	53	13	85	45	22	18	62	71	0	152	114	38
Dueño	106	52	76	8	130	71	41	35	101	106	128	114	242	0
Alquila	5	60	25	7	46	38	13	8	52	36	59	38	0	97

A partir de las salidas siguientes, confeccione un breve informe en el que incluya la respuesta a las siguientes preguntas, respecto de los datos disponibles:

a- Qué autos son los preferidos por:
   i- Solteros que alquilan
   ii- Casados con hijos
b- Es más usual ser dueño o alquilar?
c- Cuál es el estado civil más usual de la base?
d- Que explican los ejes?
e- Que indican los números resaltados en la matriz de Burt?

Ejercicio 3

Se estudiado la relación entre edad y preferencia musical, interesa testear si en ambos rangos etarios la elección de la música es similar.
Plantee las hipótesis correspondientes y concluya a partir de la salida. Es correcto aplicar la prueba a estos datos?.

Frecuencias absolutas

música	jov	mayor	Total
A	70	0	70
B	45	45	90
C	30	60	90
D	0	100	100
E	35	15	50
Total	180	220	400

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	185,86	4	<0,0001

Ejercicio 4

En el archivo ratas.xls se han registrado los conteos de glóbulos rojos de un conjunto de ratas que fueron asignadas aleatoriamente a cuatro tratamientos distintos de los cuales se sospecha que provocan anemia.

a- Plantee los supuestos del modelo para comparar los valores medios de glóbulos rojos de las ratas de los cuatro grupos.
b- Realice la prueba y el análisis diagnostico(supuestos)
c- Si es válida, concluya, si no lo es, utilice otra prueba y concluya.
d- Explique en qué casos realizaría transformaciones de las variables.