Aux Extra MA34B-02
280605

PROBLEMA 1

Se estudian las diferencias de comportamiento en el consumo en los hogares de 19 países de la OCDE. Para cada país se consideran los consumos promedios por hogar de 8 ítems (Alimentación, vestuario,..., diversiones, otros). Los datos se llevaron a porcentajes, es decir que para cada país los 8 datos suman 100% (Tabla 5). Los resultados del análisis en componentes principales sobre estos datos se encuentran en las tablas 6 y 7 y el gráfico 1.

a) ¿Que representan los valores propios de la matriz de correlaciones R (tabla 6)? Dé las proporciones de la varianza reproducida por el plano principal. Porque se tienen solamente 7 valores propios no nulos.
b) Exprese la primera componente principal en función de las 8 variables y comente.
c) A partir de las correlaciones (tabla 7), haga un gráfico de las variables sobre las 2 primeras componentes principales (círculo de correlaciones). Interprete el gráfico (No olvide que las variables son porcentajes).
d) ¿Puede expresar la variable porcentaje de "Alimentación" a partir de las dos primeras componentes principales?
e) Interprete el gráfico 1. En particular, en que difieren España, Suiza, Estados Unidos y Alemania. ¿Qué pueden decir de Francia?

Tabla 5: Perfiles de consumo

	Alimentación	Vestuario	Alojamiento	Muebles	Educación	Transporte	Diversiones	Otros	Total
Alemania	20.5	7.9	21.1	9.4	3.5	17.8	10.5	9.3	100
Australia	20.7	5.5	20.3	6.6	7.3	14.8	9.8	15	100
Austria	19.4	8.8	17.9	7.7	5.6	16.8	7.5	16.3	100
Belgica	17.8	7.7	16.8	10.6	11.7	13.1	6.2	16.1	100
Canada	15.8	5.3	24.5	8.8	4.7	14.3	11.1	15.5	100
Dinamarka	21.2	5.3	28.2	6.2	2.3	15.3	10.1	11.4	100
España	20.4	8.6	12.5	6.5	4.4	15.5	6.6	25.5	100
Estados Unidos	12	6.1	18.3	5.8	17.5	13.6	10.2	16.5	100
Finlandia	24	4.9	22.7	6.2	5.1	14.3	9.5	13.3	100
Francia	18.6	6.1	20	7.7	9.8	16.1	7.6	14.1	100
Grecia	36.7	8	12.6	7.6	3.9	15	5.6	10.6	100
Holanda	14.9	6.7	18.5	7	13.1	13.4	10.1	16.3	100
Ingleterra	21.6	5.7	19.4	6.5	1.6	16.8	10.2	18.2	100
Irlandia	35	6.9	12.4	7.1	4	12.8	12.1	9.7	100
Islandia	26.1	8	16.2	8.2	2.3	15.2	11.7	12.3	100
Italia	20	9.8	15.9	9.3	6.8	12.1	8.9	17.2	100
Japon	20.1	6.2	20.2	6.1	11	9.8	10.3	16.3	100
Suecia	19.8	6.5	31.4	5.8	3.1	16.1	9.8	7.5	100
Suiza	26.5	4.2	20.5	4.7	11.1	11.7	10.2	11.1	100

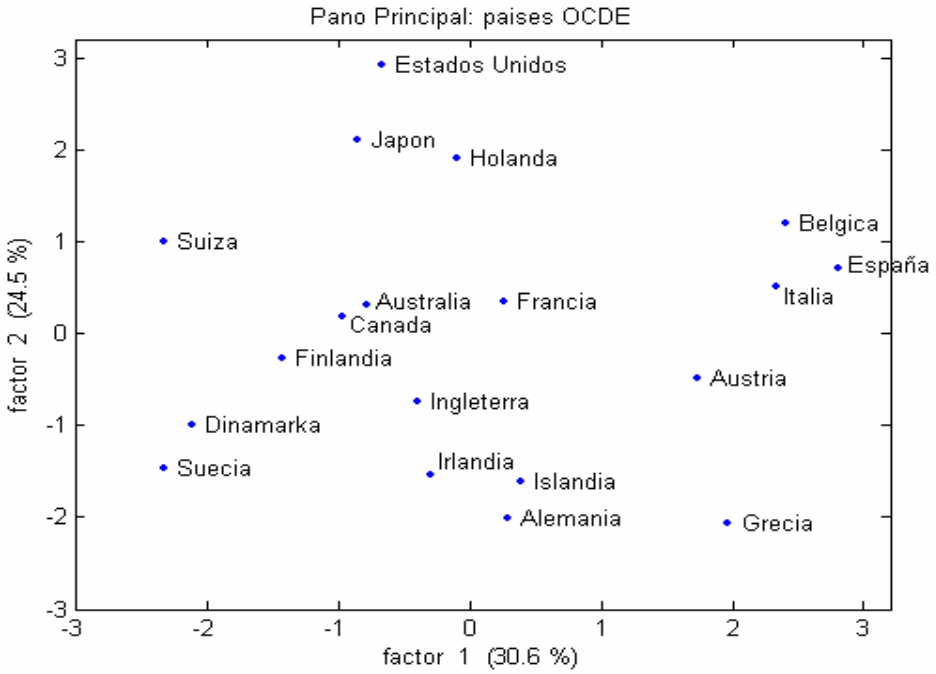
Tabla 6 : Valores y vectores propios de R (de norma 1)

	1	2	3	4	5	6	7
Valor propio	2.4520	1.9632	1.3991	0.8799	0.6649	0.3479	0.2930
Alimentación	0.0407	-0.4562	-0.6271	0.1628	0.1628	0.0592	-0.1238
Vestuario	0.5475	-0.1404	0.0381	-0.2129	-0.2129	-0.0565	0.7732
Alojamiento	-0.4794	-0.0588	0.4614	-0.0813	-0.0813	0.4471	0.2622
Muebles	0.4104	-0.1386	0.2188	-0.6656	-0.6656	0.1827	-0.5114
Educación	-0.0051	0.6305	-0.1005	-0.1796	-0.1796	-0.4968	0.0126
Transporte	0.0675	-0.4292	0.5441	0.2994	0.2994	-0.6056	-0.1496
Diversiones	-0.4289	-0.0473	-0.0985	-0.4297	-0.4297	-0.2676	0.0270
Otros	0.3344	0.4069	0.1692	0.4146	0.4146	0.2735	-0.1824

Tabla 7 : Correlaciones entre variables antiguas y componentes principales

	Componentes principales						
	1	2	3	4	5	6	7
Valor propio	2.4520	1.9632	1.3991	0.8799	0.6649	0.3479	0.2930
Alimentación	0.0638	-0.6392	-0.7418	0.1527	0.0894	0.0349	-0.0670
Vestuario	0.8573	-0.1967	0.0451	-0.1997	-0.0901	-0.0333	0.4186
Alojamiento	-0.7506	-0.0823	0.5457	-0.0763	0.1908	0.2637	0.1419
Muebles	0.6426	-0.1942	0.2588	-0.6244	0.0649	0.1077	-0.2768
Educación	-0.0081	0.8834	-0.1188	-0.1684	0.3020	-0.2930	0.0068
Transporte	0.1057	-0.6013	0.6436	0.2808	0.0006	-0.3572	-0.0810
Diversiones	-0.6716	-0.0663	-0.1165	-0.4031	-0.5860	-0.1579	0.0146
Otros	0.5237	0.5701	0.2001	0.3889	-0.4166	0.1613	-0.0988

Gráfico 1

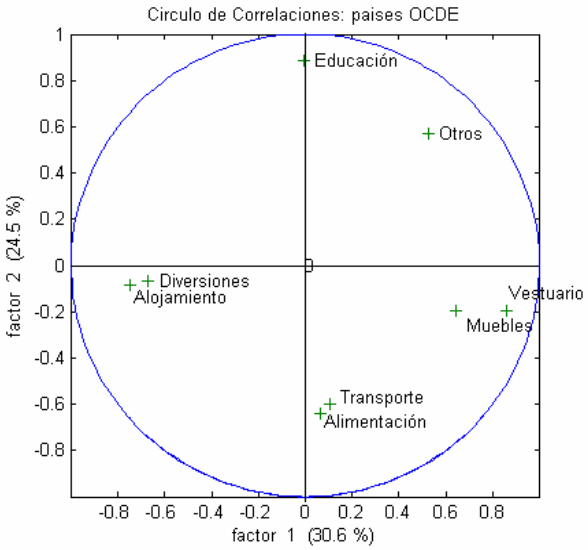


Solucion Problema 1:

a) Los valores propios representan las varianzas de las componentes principales. Como las C.P. son no correlacionadas, el plano principal reproduce 55.1% (30.6+24.5) de la varianza total. Como cada fila de la matriz suma 100%, los países pertenecen a un hiperplano, es decir a un espacio de dimensión 7.

b) C_1=0.0407*Alimentación+ 0.5475*Vestuario -0.4794*Alojamiento+ 0.4104*Muebles -0.0051*Educación+ 0.0675*Transporte -0.4289*Diversiones+ 0.3344*Otros

c) Se observa que si se tiene un alta consumo en Alojamiento y Diversiones será a costa del Vestuario y de los Muebles. Mientras que el consumo en Educación va en contra del consumo en Alimentación y Transporte. Las variables Educ. Alim. Y Transp.. son poco relacionadas con Aloj., Divers., Vestuario y Muebles.



d) No se puede expresar la variable porcentaje de "Alimentación" a partir de las dos primeras componentes principales, ya que no esta sobre la circunferencia del circulo.

e) Estados Unidos usa gran parte de su gasto en Educación en comparación del gasto en Transporte y Alimentación, siendo lo contrario para Alemania. En Suiza, el gasto en Alojamiento y Diversiones es elevado en comparación del Vestuario y Muebles, siendo lo contrario para España. Francia parece tener un perfil equilibrado, ya que esta situado cerca del punto medio.

Problema 2

El ministerio de educación quiere estudiar de qué depende el gasto anual en educación de un hogar, para ello, recolecta información en 100 hogares y plantea el modelo lineal:

$$E(y) = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 \quad (1)$$

Donde x_1 es el ingreso del hogar (en miles de pesos), x_2 el número de hijos, x_3 la talla del jefe de hogar y x_4 el número de perros en la casa.

2.1 Complete los resultados de la regresión lineal (1) dados en las tablas n° 4 y 5.

2.2 Interprete los resultados.

2.3 Se plantea un modelo con el ingreso y el n° de hijos solamente:

$$E(y) = b_o + b_1 x_1 + b_2 x_2 \quad (2)$$

Se propone resolver el test: de hipótesis $H_o : E(y) = b_o + b_1 x_1 + b_2 x_2$ contra

$H_1 : E(y) = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$.

Para esto, se resuelve el modelo (2) obteniéndose el conjunto de resultados presentados en las tablas 6 a 7.

Comente el cambio en la suma de los cuadrados de los residuos *SSR* del modelo (1) al modelo (2) y cómo afecta ésta al coef de correlación multiple.

Se propone como estadístico para medir la significación del cambio en la suma residual a:

$$\frac{(SSR_2 - SSR_1)/(k_2 - k_1)}{SSR_1/(n - k_1)}$$

donde SSR_2 y SSR_1 corresponden a la suma de los cuadrados de los residuos de los modelos (2) y (1) respectivamente, y donde k_1 y k_2 son la cantidad de coeficientes de cada modelo. Encuentre la distribución que sigue este estadístico y concluye con un error de tipo I de 5% si las variables *nº de perros* y *talla del jefe* son significativas en el modelo (1) utilizando los resultados de las tablas 4 a 7.

2.4 Dé intervalos de confianza de nivel 95% para los tres parámetros del modelo (2).

2.5 Se tiene un nuevo hogar con un ingreso de 400 y 3 hijos. Dé una estimación de su gasto en educación.

Table n°4

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	20.387	20.384	1.000	0.319
Ingreso	0.189		9.242	0.000
N° hijos	17.379	2.978	5.836	0.000
Talla jefe	8.869	6.176		0.154
N° perros		0.107	1.749	0.083

Coeficiente de correlación múltiple R=0.785

Estimación insesgada de la varianza del error $\hat{s} = 29.12$

Tabla n°5

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión		129489.083	38.185	0.0000
Residuos	95	80539.635		
Total	99			

Tabla n°6

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	54.03477	8.575475	6.301	0.000
Ingreso	0.197514	0.019715	10.019	0.000
N° hijos	17.804395	2.969696	5.995	0.000

Coeficiente de correlación múltiple R=0.772

Estimación insesgada de la varianza del error $\hat{s} = 29.56$

Tabla n°7

Fuente	Grados libertad	Suma cuadrados	F	p-valor
Regresión	2	125292.851	71.713473	0.0000
Residuos	97	84735.8665		
Total	99	210028.718		

Solución:

2.1 Tabla n°3

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	20.387	20.384	1.000	0.319
X_1	0.189	0.020	9.242	0.000
X_2	17.379	2.978	5.836	0.000
X_3	8.869	6.176	1.436	0.154
X_4	0.188	0.107	1.749	0.083

Tabla n°4

Fuente	Grados libertad	Suma cuadrados	Cuadrados medio	F	p-valor
Regresión	4	129489.083	32372.271	38.185	0.0000
Residuos	95	80539.635	847.786		
Total	99	210028.718			

2.2 El modelo (1) es globalmente significativo. Sin embargo se observa que las tercera y cuarta variables no parecen significativas.
2.3 El modelo (2) es tan bueno como el modelo (1). El coeficiente de correlación múltiple disminuye muy poco y un test F lo permite confirmar:

$$\frac{(\sum (\hat{e}_i^{H_0})^2 - \sum (\hat{e}_i^{H_1})^2)/2}{\sum (\hat{e}_i^{H_1})^2/(100-5)}$$

Como $\sum (\hat{e}_i^{H_0})^2 = 97 * \hat{s}^2 = 97 * 29.56^2 = 84758$; en la tabla es 84735.8665 exactamente.
$\sum (\hat{e}_i^{H_1})^2 = 95 * \hat{s}^2 = 95 * 29.12^2 = 80558$ en la tabla es exactamente 80539.635

F=(84735-80539)*95/2*80539=2.4748

Luego el estadístico es el p-valor del test es: $Pr(F_{2,95} > 2.4748) \approx 0.09$
No se rechaza la hipótesis nula. El modelo (2) es tan significativo como el modelo (1).

2.4 De la tabla nº6 se deduce el intervalo: estimación ±(valor_tabla)*Desv. típica:

Variable	Estimación	Desviación típica	t-Student	Intervalo
Constante	54.03477	8.575475	6.301	[37.23, 70.84]
Ingreso	0.197514	0.019715	10.019	[0.159, 0.236]
N° hijos	17.804395	2.969696	5.995	[11.98, 23.62]

2.5 La estimación del gasto es: y^*=(1 400 3).(54.0348 0.19751 17.8044)'=186.45.