GUIA EXAMEN

Problema 1

Se estudian 6 características de juego de 35 jugadores de tenis (derecho, revés, servicio, volea, retorno del servicio y el estado psíquico), que corresponden a notas entre 0 y 10. Los resultados del análisis en componentes principales sobre estos datos se encuentran en la tabla 1 y grafico 1.

a) Interprete los valores propios adjuntos (tabla 1). Dé las proporciones de la varianza reproducida por cada componente principal. Exprese la primera componente principal en función de las 6 variables y comente. ¿Pueden expresar las 6 variables a partir de las componentes principales? ¿Cómo?
b) A partir de las correlaciones adjuntas (tabla 1), haga un gráfico de las variables sobre las 2 primeras componentes principales. Interprete el gráfico. Deduzca los coeficientes de correlación aproximados entre las 6 variables.
c) Interprete el gráfico 1. En particular, en que difieren Connors, Pecci, Solomon y Mc Enroe.
d) Un nuevo jugador VILAS tiene como valores (centradas y reducidas) para las 6 variables: 1.1418 0.8038 -0.6381 -0.9754 0.8507 0.9692. Cuales son sus coordenadas en el plano principal de los 2 primeros factores. Describe su juego y ¿a quien se parece su juego?
e) Se quiere hacer la regresión lineal de una nueva variable el "Smash" sobre las dos primeras componentes principales sabiendo que sus correlaciones con las dos primeras C. P. son 0.1982 y -0.9022. Dé los coeficientes de la regresión sabiendo que la desviación estándar de la variable “Smash” es 1.9462.
f) Dé el coeficiente de correlación múltiple. ¿Este último aproxima bien el coeficiente de correlación múltiple de "Smash" sobre las 6 otras variables?

Tabla 1: Correlaciones entre variables antiguas y componentes principales

	Componentes principales					
	1	2	3	4	5	6
Valor propio	2.935	1.9598	0.4557	0.3547	0.1657	0.1290
Derecho	0.7962	0.2659	-0.4014	0.3596	-0.0687	-0.0134
Revés	0.9162	-0.0525	-0.07	-0.3095	-0.0821	0.2246
Servicio	0.0104	-0.9448	-0.1438	0.1124	0.2533	0.0993
Volea	-0.1033	-0.9422	-0.1379	-0.0799	-0.2490	-0.1189
Retorno servicio	0.9348	-0.0042	-0.0095	-0.2175	0.1569	-0.2328
Estado psíquico	0.7597	-0.3254	0.5	0.2515	-0.0594	0.0120

PAUTA

a) Los 6 valores propios suman 6 (la matriz R es de rango a lo más 6). Las dos primeras componentes principales reproducen 81,58% de la varianza.

	Componentes principales					
	1	2	3	4	5	6
Valor propio	2.935	1.9598	0.4557	0.3547	0.1657	0.1290
% accumulado	48,92	81,58	89,18	95,09	97,85	100,00

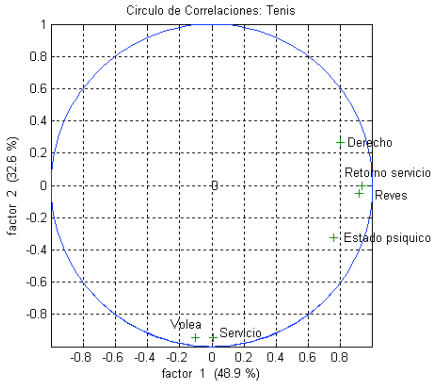
$$c_1 = \frac{0.79}{\sqrt{2.935}} x_1 + \frac{0.92}{\sqrt{2.935}} x_2 + ... + \frac{0.759}{\sqrt{2.935}}$$

La primera C.P. tiene que ver con el derecho, revés, retorneo servicio y estado físico, mientras que la 2da con Servicio y Volea, es decir fuerza.
Cada variable puede expresar a partir de las 6 C. P.:

$$x_1 = \frac{0.79}{\sqrt{2.935}} c_1 + \frac{0.266}{\sqrt{1.96}} c_2 + ... + \frac{-0.0134}{\sqrt{0.129}} c6$$

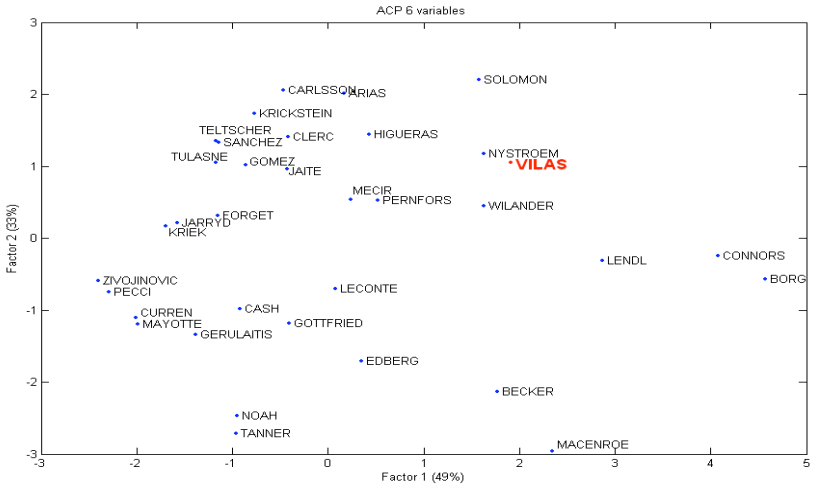
b) El circulo de correlaciones permite: Interpretar las componentes principales, ver la representatividad de las antiguas variables en los planos factoriales y representar la matriz de correlación R de las antiguas variables.

Los valores aproximados son los cosenos de los ángulos entre las variables en el círculo.



c) En el gráfico 1 se puede interpretar las proximidades entre los jugadores. En particular Connors tiene un muy buen derecho, revés, retorno de servicio y estado psíquico pero mediano en servicio y volea. Para Solomon es mediano salvo que es malo para el servicio y volea. McEnroe es bastante bueno en derecho, revés, retorno servicio y estado físico y muy bueno en servicio y volea. Finalmente Pecci es mediano en servicio y volea y malo en el resto.

d) VILAS: 1.9094 1.0460; Se parece a NYSTROEM.



e) Los coeficientes de la regresión de "Smash" sobre las 2 primeras C.P. son los coeficientes de correlación divididos por la raíz del valor propio y multiplica par la desviación estándar de "Smash":

$$0.1982 * \quad 1.9462 / \sqrt{2.935} = 0.2252$$

f) El coeficiente de correlación múltiple es: $\sqrt{0.1982^2 + 0.9022^2} = \sqrt{0.8532} = 0.9237$.

Las dos primeras componentes principales reproducen 81,58% de la varianza, 0.9237 aproxima probablemente bien el coeficiente de correlación múltiple de "Smash" sobre las 6 otras variables. (A titulo indicativo, el valor real es: 0.96).

Problema 2

Sean X una matriz de datos (*n=200* observaciones y *p=4* variables X_1, X_2, X_3, X_4) que se supondrán centrados y reducidos. Sea *R* la matriz de correlación.

a) Dibuje el circulo de correlaciones a partir de la tabla 1.1. ¿Rango de la matriz *R*?
b) El círculo de correlaciones tiene 3 funciones. Cítelas.

c) Se tiene una quinta variable *Z* cuyas correlaciones con los 3 factores son respectivamente: 0.75, 0.2 y 0.5. Dé el coeficiente de correlación múltiple de *Z* sobre las cuatro variables X_1, X_2, X_3, X_4.
d) Deduzca los coeficientes de la regresión lineal de *Z* (centrada y reducida) sobre las cuatro variables X_1, X_2, X_3, X_4 (¡OJO! Con la varianza de los factores).

Tabla 1.1

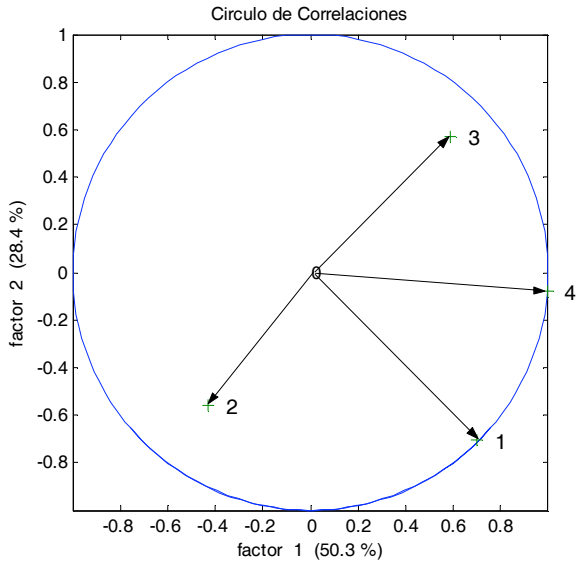
	Factor 1	Factor 2	Factor 3
Valor propio	2.0145	1.1372	0.8483
Vector propio	0.4944 -0.3012	-0.6609 -0.5233	0.1134 -0.7722

	0.4152 0.7018	0.5327 -0.0742	-0.6238 -0.0423

PAUTA

a) Se obtienen las correlaciones con $\sqrt{\lambda_j}u_j$:

Correlación	C1	C2	C3
Valor propio	2.0145	1.1372	0.8483
X1	0.7017	-0.7048	0.1045
X2	-0.4275	-0.5581	-0.7112
X3	0.5893	0.5681	-0.5745
X4	0.9961	-0.0791	-0.0389



Los 3 valores propios suman 4. La matriz R es de rango 3.

b) El circulo de correlaciones permite:
- i) Interpretar las componentes principales.
- ii) Ver la representatividad de las antiguas variables en los planos factoriales.
- iii) Representar la matriz de correlación R de las antiguas variables.

c) El coeficiente de correlación múltiple es: $\sqrt{0.75^2+0.2^2+0.5^2}=0.8732$.

d) Los coeficientes de la regresión de *Z* sobre los 3 factores son los coeficientes de correlación divididos por la norma de los factores si estos se obtienen como: $C_j = Xu_j = \sum_k u_{jk}X_k$

$Z=\frac{0.75}{\sqrt{\lambda_1}}C_1+\frac{0.2}{\sqrt{\lambda_2}}C_2+\frac{0.5}{\sqrt{\lambda_3}}C_3$. Luego $Z=\frac{0.75}{\sqrt{\lambda_1}}Xu_1+\frac{0.2}{\sqrt{\lambda_2}}Xu_2+\frac{0.5}{\sqrt{\lambda_3}}Xu_3$

o sea $Z=\frac{0.75}{\sqrt{\lambda_1}}(\sum_k u_{1k}X_k)+\frac{0.2}{\sqrt{\lambda_2}}(\sum_k u_{2k}X_k)+\frac{0.5}{\sqrt{\lambda_3}}(\sum_k u_{2k}X_k)$

$$Z = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \text{ con } \beta_k = \sum_j \frac{cor(C_j, Z)}{\sqrt{\lambda_j}} u_{jk}$$

$$\boxed{Z = 0.1347 X_1 - 0.6593 X_2 - 0.1194 X_3 + 0.2233 X_4}$$

Problema 3 (propuesto)

Un médico que realiza delicadas y costosas operaciones quirúrgicas ha oído hablar de un método estadístico que le permitiría estimar el tiempo de sobrevivencia de un paciente después de operarse. El cuenta con los resultados de los exámenes de los tests de hígado, enzima, coagulación y presión arterial, realizados a pacientes antes de ser operados. Además del tiempo de vida después de la operación de los mismos pacientes.
Se plantea el modelo lineal: $tiempo\ \ sobrevivencia = \beta_0 + \beta_1 higado$ (1)

a) Encuentre la expresión de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 respectivamente en función de los datos con los que cuenta el médico. (Hint: llame X e Y a las variables involucradas).

b) Al realizar la matriz de correlaciones de las variables se obtiene la tabla 2.
   i) ¿Dé la formula de calculo de los coeficientes de la matriz?
   ii) ¿Qué rango de valores puede tomar cada coeficiente?¿Cómo se interpreta?
   iii) ¿Qué cuidados se deben tener al interpretar cada coeficiente de la matriz?
   iv) Si usted pudiera observar solo un examen de un paciente y con este resultado predecir su tiempo de sobrevivencia ¿Cuál examen de los 4 exámenes utilizaría?

Tabla 2

Correlación	Coagulación	Presión	Enzima	Hígado	Tiempo
Coagulación	1.00000	0.09012	-0.14963	0.50242	0.37252
Presión	0.09012	1.00000	-0.02361	0.36903	0.55398
Enzima	-0.14963	-0.02361	1.00000	0.41642	0.58024
Hígado	0.50242	0.36903	0.41642	1.00000	0.72233
Tiempo	0.37252	0.55398	0.58024	0.72233	1.00000

c) Se plantea el siguiente modelo lineal (2):

$$tiempo = \beta_0 + \beta_1 coagulación + \beta_2 presion + \beta_3 enzima + \beta_4 higado \quad (2)$$

Obteniéndose al efectuar la regresión un $R^2 = 0.8367$ y los resultados en las tablas 3 y 4. Escriba el modelo (2) en función de los estimadores $\hat{\beta}_i$. Interprete los valores obtenidos y justifique bien sus interpretaciones.

Tabla 3

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrado medio	F	p_valor
Regresión	4	936264.538	234066.134	.787712	0.00000
Residuos	49	182666.962	3727.89719		
Total	53	1118931.5			

Tabla 4

Variable	Estimación parámetro	Error estándar	t Student	P_valor
Coagulación	33.164	7.017	4.726	0.000
Presión	4.272	0.563	7.582	0.000
Enzima	4.126	0.511	8.071	0.000
Hígado	14.092	12.525	1.125	0.266
Constante	-621.598	64.800	-9.592	0.000

d) ¿Su interpretación anterior contradice su repuesta de la parte 2.2(d)? ¿Cómo se explica que exista (o no exista) esta contradicción?

e) Al realizar el gráfico de los residuos y estimaciones de la regresión se obtienen los gráficos 1 y 2. Interprete los resultados obtenidos en función de los supuestos usuales del modelo lineal.

Gráfico1

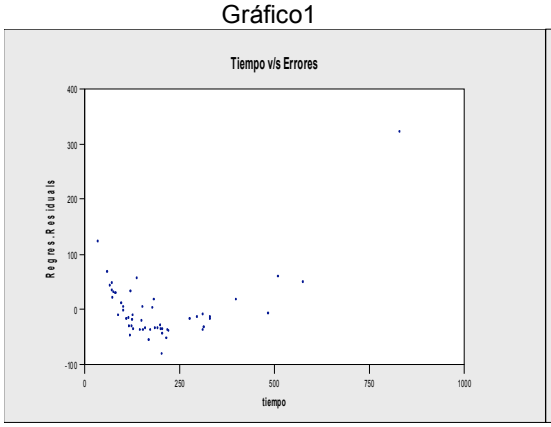
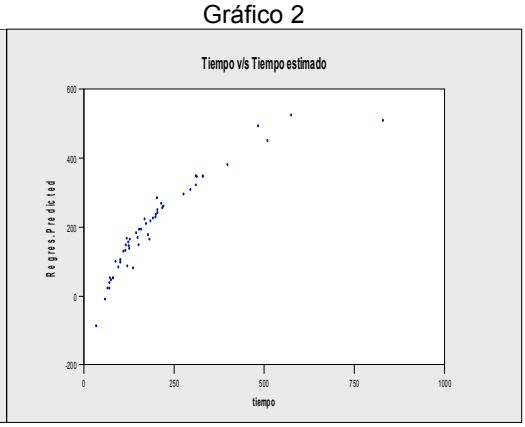


Gráfico 2



Problema 4

Un instituto de estudios ambientales quiere analizar la relación entre diferentes factores atmosféricos y la calidad del ozono en el aire. Las variables consideradas son: ozono (nº de partículas por billón), el nivel de radiación solar (radsolar), la velocidad en millas por hora (viento) y la temperatura del ambiente en grados Fahrenheit (temp). Se propone ajustar un modelo de la forma:

$$ozono_i = \beta_o + \beta_1 radsolar_i + \beta_2 viento_i + \beta_3 temp_i + \varepsilon_i \quad (1)$$

Se supone que los errores ε_i del modelo (1) cumplen los supuestos usuales de normalidad y correlación nula, esto es $\varepsilon_i \sim N(0,\sigma^2)$ y $cov(\varepsilon_i,\varepsilon_j) = 0$ para $i \neq j$.

a) Se tomó una muestra de 15 observaciones, y se registraron las correlaciones entre las variables en la siguiente matriz (Tabla 1) y se procedió a plantear el modelo (1). Los resultados de la regresión se encuentran en las tablas 2 y 3. Complete los resultados de la regresión lineal (1) dados en las tablas n° 2 y 3.

b) Basándose en los resultados anteriores, uno de los investigadores propone un modelo alternativo: $ozono_i = \beta_o + \beta_1 radsolar_i + \beta_2 viento_i + \varepsilon_i$ (2) cuyos resultados se encuentran en las tablas 6 y 7. Interprete el modelo y compárelo con el anterior.

Decida cual de los dos es mejor, indicando los supuestos y criterios en los cuales se basa su decisión. Justifique su respuesta.

Tabla 1

Variable	Ozono	Radsolar	Viento	Temp
Ozono	1	0,117	-0,803	0,551
Radsolar		1	0,136	-0,352
Viento			1	-0,818
Temp				1

Tabla 2

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	219,393	169,384	?	0,2218
Radsolar	?	0,106	1,02	0,3307
Viento	-14,965	4,537	-3,30	0,0071
Temp	-0,959	?	-0,58	0,5710

Coeficiente de determinación R^2 =0.705

Tabla 3

Fuente	Grados libertad	Suma cuadrados	Cuadrados Medios	F	P-valor
Regresión	3	21987	7329	8,78	0,00295
Residuos	?	9182	?		
Total	14	?			

Tabla 4

Variable	Estimación	Desviación típica	t-Student	P-Valor
Constante	121,7745	26,7456	4,55	0,0007
Radsolar	0,1344	0,0937	1,43	0,1770
Viento	-12,7650	2,4588	-5,19	0,0002

Coeficiente de determinación R^2 =0.696

Tabla 5

Fuente	Grados libertad	Suma cuadrados	Cuadrados Medio	F	P-valor
Regresión	2	21616	10808	13,7	0,000787
Residuos	12	9467	788,92		
Total	14	31083			

PAUTA

a) La ecuación que relaciona los distintos de la tabla 2 es:

$$\frac{\hat{\beta}_i}{\widetilde{\sigma}_{ii}} = t_{\hat{\beta}_i} \quad \text{es decir} \quad \frac{\text{Estimación}}{\text{Desviación Típica}} = \text{t} - \text{Student}$$

Usando esta ecuación se encuentra que:

t-Student de Constante = 1,295
Estimación de Radsolar = 0,10812
Desviación típica de Temp = 1,653

Para el cuadro ANOVA:

gl(Regresión)+gl(residuos)=gl(Total)
Suma cuadrados(regresión)+Suma cuadrados(residuos)=Suma cuadrados(total)
Cuadrados Medios = Suma cuadrados/gl
F = Cuadrados Medios(Regresión)/Cuadrados Medios(Residuos)

De donde:

gl(Residuos) = 14 - 3= 11
Suma cuadrados(Total) = 21987 + 9182 = 31169
Cuadrados Medios(Residuos) = 9182/11 = 834,73

b) Al analizar el primer modelo podemos observar que si bien el coeficiente de determinación R^2 no es bajo, y el modelo es globalmente significativo, ya que F=8,78 con P-valor=0,00295<5%, la mayoría de las variables son no significativas (P-valor mayor al 5%). En el caso del segundo modelo, el R^2 es un poco más bajo, y el modelo es significativo (F=13,7 con P-valor=0,000787<5 %), pero salvo Radsolar, el resto de las variables son significativas. Si examinamos los residuos del primer y segundo modelo: 9182 y 9467 respectivamente (o los coeficientes de determinación), veríamos que no existe una diferencia apreciable, lo que hace sospechar que la variable Temperatura no es importante en el modelo.

NOTA: Este último hecho se puede probar formalmente haciendo un test de Fisher:

$$F = \frac{\frac{SSR_r - SSR_c}{k_c - k_r}}{\frac{SSR_c}{n - k_c}} = \frac{\frac{9467 - 9182}{4 - 3}}{\frac{9182}{15 - 4}} = 0{,}341$$

Donde SSR_c, k_c y SSR_r, k_r representan el modelo completo y el modelo reducido con la cantidad de sus coeficientes respectivamente. Como 0,341 < $F_{1,\ 11}$(5 %) = 4,844; no se rechaza que el coeficiente de la variable Temp sea 0, y por ende ambos modelos hacen un ajuste similar. Además se puede observar que al remover la variable Temp, las desviaciones estándar de los coeficientes disminuyen, lo que aumenta su precisión. Finalmente notando que existe una alta correlación entre esta

variable y el resto, es posible que Temp no esté haciendo un aporte signidicativo al modelo en términos de información. Es por esto que al parecer, desde el punto de vista de los datos es mejor quedarse con el segundo modelo.