

## Guía Regresión y ACP

### Problema 1

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:

$$E(y) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 \quad (1)$$

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:

$$E(y) = b_0 + b_1x_1 + b_2x_2 \quad (2)$$

Se propone resolver el test: de hipótesis  $H_0 : E(y) = b_0 + b_1x_1 + b_2x_2$  contra  $H_1 : E(y) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$ .

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 calcule la variación porcentual.

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 *n° 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 |
|-----------|--|-------|-------|-------|

**Coefficiente de correlación múltiple  $R=0.785$**

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

Tabla n°5

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

Table n°6

| Variabl e  | Estimac ión | Desviación típica | t- Stude nt | P- Valor |
|------------|-------------|-------------------|-------------|----------|
| Consta nte | 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    |

**Coefficiente de correlación múltiple  $R=0.772$**

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

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     |           |         |

## Problema 2

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 (n° 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:

$$\text{ozono}_i = \beta_0 + \beta_1 \text{radsolar}_i + \beta_2 \text{viento}_i + \beta_3 \text{temp}_i + \varepsilon_i \quad (1)$$

Se supone que los errores  $\varepsilon_i$  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) Suponga que para comenzar el estudio se tomó una muestra de prueba con solo 4 observaciones (tabla 1). Encuentre el estimador de mínimos cuadrados. Calcule los residuos (errores estimados) y el coeficiente de determinación  $R^2$  utilizando la tabla 2. Generalice lo anterior para un modelo en que el número de observaciones es igual al número de coeficientes (HINT: Para lo anterior examine la estructura de la matriz X). Comente.
- b) En una segunda etapa se amplió la muestra a 15 observaciones, y se registraron las correlaciones entre las variables en la siguiente matriz (Tabla 3) y se procedió a plantear el modelo (1). Los resultados de la regresión se encuentran en las tablas 4 y 5. Complete los resultados de la regresión lineal (1) dados en las tablas n° 2 y 3.
- c) Basándose en los resultados anteriores, uno de los investigadores propone un modelo alternativo:  $ozono_i = \beta_0 + \beta_1 radsolar_i + \beta_2 viento_i + \varepsilon_i$  (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

matriz de correlaciones

| Obs. | Ozono | Radsolar | Viento | Temp |
|------|-------|----------|--------|------|
| 1    | 44    | 286      | 8,6    | 78   |
| 2    | 168   | 238      | 3,4    | 81   |
| 3    | 20    | 223      | 11,5   | 68   |
| 4    | 23    | 148      | 8,0    | 82   |

Table 3:

| Variable | Ozon<br>o | Radsola<br>r | Vient<br>o | Tem<br>p |
|----------|-----------|--------------|------------|----------|
| Ozono    | 1         | 0,117        | -0,803     | 0,551    |
| Radsolar |           | 1            | 0,136      | -0,352   |
| Viento   |           |              | 1          | -0,818   |
| Temp     |           |              |            | 1        |

Tabla 2

$$(X^t X)^{-1} = \begin{pmatrix} 226.121 & -0.0930 & -3.6486 & -2.2825 \\ & 0.000124 & 0.00117 & 0.000725 \\ & & 0.08243 & 0.03536 \\ & & & 0.0238 \end{pmatrix} \quad X^t Y = \begin{pmatrix} 255 \\ 60432 \\ 1363 \\ 20286 \end{pmatrix}$$

Table 4

| Variable  | Estimación | Desviación<br>típica | t-<br>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 5

| Fuente    | Grados libertad | Suma<br>cuadrados | Cuadrados<br>Medio | F    | P-valor |
|-----------|-----------------|-------------------|--------------------|------|---------|
| Regresión | 3               | 21987             | 7329               | 8,78 | 0,00295 |
| Residuos  | ?               | 9182              | ?                  |      |         |

|       |    |   |  |  |  |
|-------|----|---|--|--|--|
| Total | 14 | ? |  |  |  |
|-------|----|---|--|--|--|

Table n°6

| 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 n°7

| 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          |                 |      |          |

### Problema 3

Se estudia la relación entre el Coeficiente Intelectual (CI) y el rendimiento escalar de los 30 alumnos de un curso de sexto básico: la variable  $y$  representa el CI,  $x_1$  la nota de castellano,  $x_2$  la nota de matemática,  $x_3$  la nota de biología y  $x_4$  la nota de Inglés. Se presentan la matriz de las correlaciones de todas las variables en la tabla 2 y los resultados de 2 modelos de regresión lineal distintos en las tablas 3 y 4.

- Analice los resultados de la regresión dados en la tabla 3. Considerando la tabla 2 ¿qué opina del efecto de la nota de matemática en el modelo?
- Dé un intervalo de confianza de nivel 95% para el coeficiente  $\beta_1$  de la nota de castellano.
- Realice el test  $H_0 : \beta_1 = 4$  contra  $H_1 : \beta_1 < 4$  con un error de tipo I de 5%.
- En el modelo de la tabla 4, se eliminó la variable "Matemática". Examine y compare las dos regresiones.
- ¿Cuáles son los supuestos usuales sobre los errores de un modelo lineal? ¿Cómo se relacionan estos supuestos con las propiedades de los  $\hat{\beta}$ ?

Tabla 2: Matriz de correlaciones

|            | CI   | Castellano | Matemática | Biología | Inglés |
|------------|------|------------|------------|----------|--------|
| CI         | 1.00 | 0.67       | 0.82       | 0.75     | 0.83   |
| Castellano | 0.67 | 1.00       | 0.45       | 0.14     | 0.47   |
| Matemática | 0.82 | 0.45       | 1.00       | 0.76     | 0.65   |
| Biología   | 0.75 | 0.14       | 0.76       | 1.00     | 0.61   |
| Inglés     | 0.83 | 0.47       | 0.65       | 0.61     | 1.00   |

Tabla 3: Modelo  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$

| Variable | Coeficiente | Des. Típica | T-student | P( X >T) |
|----------|-------------|-------------|-----------|----------|
|----------|-------------|-------------|-----------|----------|

|                                                             |          | Coeficiente |        |       |
|-------------------------------------------------------------|----------|-------------|--------|-------|
| Constante                                                   | -6130.72 | 584.06      | -10.50 | 0.000 |
| Castellano                                                  | 3.54     | 0.54        | 6.51   | 0.000 |
| Matemática                                                  | 1.33     | 1.05        | 1.28   | 0.214 |
| Biología                                                    | 7.58     | 1.62        | 4.67   | 0.000 |
| Ingles                                                      | 7.43     | 1.70        | 4.36   | 0.000 |
| Coeficiente de correlación múltiple: 0.94                   |          |             |        |       |
| F-Fisher para las tres variables: 95.24 con $P(X>F)=0.0001$ |          |             |        |       |

Tabla 4: Modelo  $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 x_4$

| Variable                                                     | Coeficiente | Des. Típica Coeficiente | T-student | $P( X >T)$ |
|--------------------------------------------------------------|-------------|-------------------------|-----------|------------|
| Constante                                                    | -6034.00    | 586.03                  | -10.30    | 0.000      |
| Castellano                                                   | 3.84        | 0.50                    | 7.74      | 0.000      |
| Biología                                                     | 8.98        | 1.20                    | 7.47      | 0.000      |
| Ingles                                                       | 7.74        | 1.70                    | 4.54      | 0.000      |
| Coeficiente de correlación múltiple: 0.93                    |             |                         |           |            |
| F-Fisher para las tres variables: 123.47 con $P(X>F)=0.0000$ |             |                         |           |            |

## Problema 4

Se desea explicar la **esperanza de vida** de 20 países de América Latina con un modelo lineal a partir de las variables explicativas **gasto en salud**, **calorías consumidas** y **tasa de alfabetización**. Se efectúa la regresión de la esperanza de vida (de media 67.11 años y desviación típica de 5.52) sobre las tres otras variables; se obtiene los resultados siguientes:

| Variable       | Media  | Des. Típica | Coeficiente | Des. Típica Coeficiente | T-student | $P( X >T)$ |
|----------------|--------|-------------|-------------|-------------------------|-----------|------------|
| Constante      |        |             | 29.603      | 5.282                   | 5.600     | 0.0001     |
| Gasto salud    | 2.09   | 1.567       | 0.959       |                         | 2.302     | 0.0336     |
| Calorías       | 108.80 | 14.420      | 0.161       | 0.054                   | 2.980     |            |
| Alfabetización | 83.00  | 11.947      | 0.217       | 0.064                   |           | 0.0041     |

**F-Fisher para las tres variables: 19.714 con  $P(|X|>F)=0.0001$**

- 2.1 Complete la tabla.
- 2.2 Concluya si la variable Alfabetización tiene influencia en el modelo.
- 2.3 Dé los grados de libertad de la F-Fisher. Interprete el resultado del test F-Fisher.
- 2.4 Construya un intervalo de confianza de nivel 95% para el parámetro  $\beta_{alfabet}$  asociado a la variable **tasa de alfabetización**. Interprete para discutir la significación de la tasa de alfabetización para explicar la esperanza de vida.
- 2.1 Estime la esperanza de vida de un país no considerado en los 20 anteriores a partir de 2,09 su gasto en salud, 108 sus calorías y 80 su alfabetización. Precise los supuestos que implica su estimación.

## Problema 5

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 gráfico 1.

- 1 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?
- 2 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.
- 3 Interprete el gráfico 1. En particular, en que difieren Connors, Pecci, Solomon y Mc Enroe.
- 4 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?
- 5 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.
- 6 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  |

### Repuesta

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 |
| % acumulado  | 48,92 | 81,58  | 89,18  | 95,09  | 97,85  | 100,00 |

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

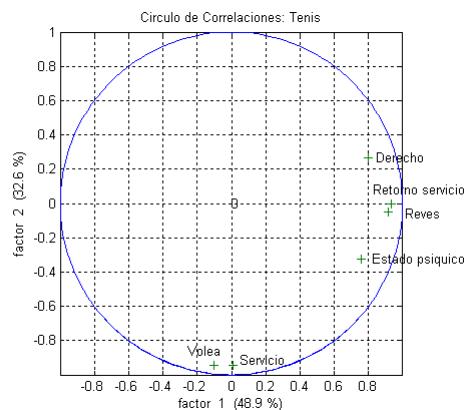
La primera C.P. tiene que ver con el derecho, revés, retorneo servicio y estado físico, mientras que la 2da con Servicio y Volea, es decir fuerza.

Cada variable puede expresarse a partir de las 6 C. P.:

$$x_1 = \frac{0.79}{\sqrt{2.935}} c_1 + \frac{0.266}{\sqrt{1.96}} c_2 + \dots + \frac{-0.0134}{\sqrt{0.129}} c_6$$

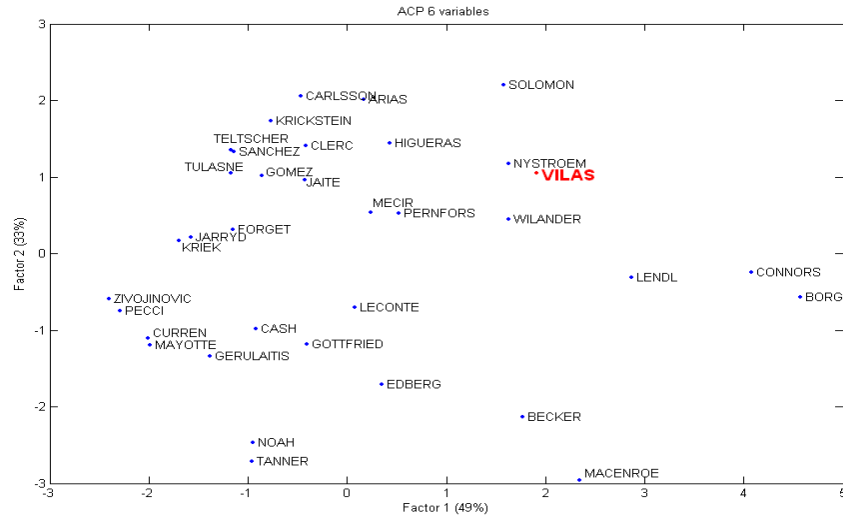
b) El círculo de correlaciones permite: Interpretar las componentes principales, ver la representatividad de las antiguas variables en los planos factoriales y representar la matriz de correlación R de las antiguas variables.

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.



- 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”:

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

y

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

- 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 6

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.

- Dibuje el círculo de correlaciones a partir de la tabla 1.1. ¿Rango de la matriz  $R$ ?
- El círculo de correlaciones tiene 3 funciones. Cítelas.
- 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  $X_1, X_2, X_3, X_4$ .
- 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

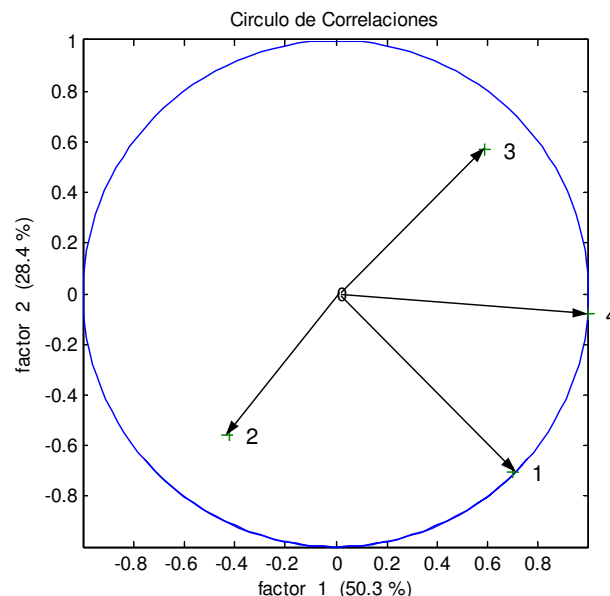
|  | Factor 1 | Factor 2 | Factor 3 |
|--|----------|----------|----------|
|--|----------|----------|----------|

|               |         |         |         |
|---------------|---------|---------|---------|
| Valor propio  | 2.0145  | 1.1372  | 0.8483  |
| Vector propio | 0.4944  | -0.6609 | 0.1134  |
|               | -0.3012 | -0.5233 | -0.7722 |
|               | 0.4152  | 0.5327  | -0.6238 |
|               | 0.7018  | -0.0742 | -0.0423 |

### Repuesta

1.1 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.

- 1.2 El circulo de correlaciones permiten:
- Interpretar las componentes principales.
  - Ver la representatividad de las antiguas variables en los planos factoriales.
  - Representar la matriz de correlación R de las antiguas variables.
- 1.3 El coeficiente de correlación múltiple es:  $\sqrt{0.75^2 + 0.2^2 + 0.5^2} = 0.8732$ .
- 3 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:

$$C_j = Xu_j = \sum_k u_{jk} X_k$$

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

o sea 
$$Z = \frac{0.75}{\sqrt{\lambda_1}}(\sum_k u_{1k} X_k) + \frac{0.2}{\sqrt{\lambda_2}}(\sum_k u_{2k} X_k) + \frac{0.5}{\sqrt{\lambda_3}}(\sum_k u_{3k} X_k)$$

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

$$Z = 0.1347 X_1 - 0.6593 X_2 - 0.1194 X_3 + 0.2233 X_4$$