INTRODUCCION AL ANALISIS MULTIVARIADO

Lab. No.1 - COMPONENTES PRINCIPALES

HEPTATLON

La pentatlón se realizó por primera vez en Alemania en 1928. Inicialmente incluía lanzamiento de peso, salto largo, 100m, salto alto y javalina. En 1964 llegó a ser el primer evento olímpico combinado para mujeres y consistía en 80m vallas, lanzamiento, salto alto, salto largo y 200m. En 1977 los 200m fueron reemplazados por carrera de 800m, y desde 1981 se pasó a la heptatlón con 100m vallas, lanzamiento, salto alto, 200m, salto largo, jabalina y 800m. Existe un sistema de puntuación que da puntos en cada evento, y la ganadora es la mujer que acumule la mayor cantidad de puntos.
En 1984 se realizó la heptatlón por primera vez. En 1988 en los juegos de Seúl la heptatlón fue ganada por una de las estrellas de EU, Jackie Joyner-Kersee.
Se cuenta con los resultados de las 25 competidoras en cada una de las 7 disciplinas. Se desea explorar estos datos usando un análisis de componentes principales para explorar la estructura de las datos y evaluar de qué forma los puntajes derivados del análisis se pueden relacionar con los puntajes totales asignados con el sistema oficial de puntuación.

1. Antes de llevar a cabo cualquier análisis note que un valor alto en un resultado no siempre significa algo mejor, por ejemplo en las carreras de 200m, 800m y vallas los resultados son tiempos y se sabe que el ganador es el que haga un tiempo menor, mientras que en los saltos o lanzamientos gana el que obtenga la mayor distancia. Por esta razón es recomendable invertir algunas variables para que todas apunten en la misma dirección. En este caso podemos invertir "vallas", "car200" y "car800" restando al máximo valor el respectivo valor obtenido, por lo que el perdedor tendrá un cero y los demás tendrán valores positivos hasta un máximo que será el valor del ganador. Por ejemplo: `base$vallas=max(base$vallas)-base$vallas`.

2. Obtenga y grafique las correlaciones de las 7 variables que contienen los resultados de las disciplinas (sin tomar en cuenta el "puntaje"). Incluya el diagrama de dispersión y colores para las correlaciones. Usa la función `corrgram` de la librería `corrgram`, indicando `lower.panel=panel.pts` para que ponga los diagramas de dispersión y `order=T` para que ordene las variables según la correlación. Pruebe también con `upper.panel=panel.conf` para que ponga los intervalos de confianza para las correlaciones y `diag.panel=panel.density` para que grafique las densidades univariadas en la diagonal.

- Lleve a cabo una discusión e investigue por qué hay deportes que pueden estar más correlacionados y alguno no lo esté tanto.

3. Identifique un valor extremo usando los "leverage" a partir de la matriz $H = X(X^TX)^{-1}X^T$. Del curso de Regresión recuerde que los "leverage" se llaman "hatvalues" en R y el límite para detectar un valor extremo es 2 veces la media de ellos. La función para obtener los "leverage" es `hat`.

4. Ahora se hace una nueva base eliminando la competidora de Papua Nueva Guinea (PNG), llámela "base1". Para esto se puede utilizar la información de la fila donde se encuentra que es la 25. Calcule nuevamente las correlaciones y observe los cambios.

- A partir de ahora use solo base1 puesto que el dato 25 distorsiona mucho las correlaciones

5. Obtenga el determinante de R (matriz de correlaciones de base1). Use la función `det` para obtener el determinante.

- Lleve a cabo la prueba de esfericidad de Bartlett para determinar si vale la pena llevar a cabo un PCA. Use la función `cortest.bartlett` de la librería `psych`. Puede usar los datos o la matriz de

correlaciones, pero en este caso debe indicar el tamaño de muestra en `n`.

- Evalúe el índice KMO con el mismo objetivo del punto anterior. Use la función `KMO` de la librería `psych`.

ANALISIS CON MATRIZ DE COVARIANZAS

6. Obtenga las variancias de las 7 variables y vea las diferencias que existen entre esas variancias. Además note que las variables tienen diferentes escalas por lo que más adelante convendrá estandarizar las variables, en cuyo caso se usará la matriz de correlaciones. Por ahora seguimos con las variables originales.

- Obtenga la variabilidad total de las 7 variables originales.
- Obtenga la matriz de covarianzas (S).
- Obtenga la traza de S. Compárela con el resultado de la variabilidad total obtenido anteriormente.

7. Haga la descomposición espectral de S usando: `espec=eigen(S)`.

- Observe los valores propios con `espec$val` y los vectores con `espec$vec`.

8. Obtenga el primer componente principal usando el primer vector característico. Primero centre todas las variables sin estandarizarlas. Para esto se puede usar scale indicando dentro de la función que use `scale=F` para que no divida entre la desviación estándar, de la siguiente forma: `scale(base2,scale=F)`. Lo anterior da todas las variables centradas y con ellas se pueden obtener los valores del primer componente llamado Z1.

- Obtenga la variancia de Z1. ¿Le parece familiar este número?
- ¿Qué porcentaje de la variabilidad total explica Z1?
- Observe los coeficientes que usó para formar Z1 y explique cuáles variables tienen mayor peso en la conformación de esta nueva variable. ¿Por qué sucede esto? ¿Es adecuado esto?

9. Lleve a cabo el análisis de componentes principales de forma automática con R usando la función `prcomp`. Guarde los resultados en un objeto llamado pca1: `pca1=prcomp(base2)`.

- Para ver cuáles son los componentes de "pca1" use `names(pca1)` y luego llame una a uno esos componentes (por ejemplo, `pca1$sdev`, `pca1$rotation`, etc).
- Compare estos resultados con los obtenidos anteriormente, en particular vea que esta función devuelve las desviaciones estándar de los componentes y no las variancias, entonces eleve al cuadrado estas desviaciones y compárelas con los valores propios.
- Use `summary(pca1)` para ver los porcentajes de variabilidad explicada por cada componente.

ANALISIS CON MATRIZ DE CORRELACIONES

10. Tome las primera variable y estandarícela. Para estandarizar una variable X puede hacerlo de dos formas:`(X-mean(X))/sd(X)` o `scale(X)`

- Obtenga la media y la varianza de la variable estandarizada y verifique que la media es 0 y la varianza es 1.
- Estandarice todas las variables. Para hacerlo de una sola vez, puede usar: `scale(base2)`.
- Verifique que todas las medias son 0 y las variancias son 1.

11. Lleve a cabo el análisis de componentes principales con las variables estandarizadas.

- Haga la descomposición espectral de la matriz de correlaciones y obtenga los valores propios.
- Hágalo de forma automática con la función `prcomp`. Para indicar que se quieren estandarizar los datos use `scale=T` en la función `prcomp`. Obtenga los valores propios.

12. Determine qué porcentaje de variabilidad es explicada por los primeros dos componentes.
13. Obtenga el primer vector propio y observe cuáles variables están contribuyendo a obtener este componente.
14. Obtenga el primer componente manualmente usando las fórmulas adecuadas y también obténgalo extrayéndolo con `pca$x`. Compare los resultados. Llame a este componente Z1a.
15. Obtenga la variancia Z1a y compárela con el primer valor propio.
16. Puesto que los coeficientes para el primer componente son negativos, puede obtener una variable equivalente al multiplicar todo por -1. Obtenga Z1b y verifique que tiene la misma variancia que Z1a.
17. Haga un biplot e interprételo. Use la función `biplot`. Puede controlar el tamaño de las etiquetas con `cex`.
    - Trate de visualizar si hay algún agrupamiento de las atletas.
    - Haga un biplot usando una variable categórica para los colores, por ejemplo, el continente al que pertenece la atleta. Para hacer un biplot más elegante se puede usar la función `ggplot_pca` de la librería `AMR`. Lo mejor es hacer un nuevo data.frame llamado `data`, donde se pone la variable categórica en la primera columna, luego los nombres de las atletas en la segunda columna y después las variables que entran en el PCA. Luego cargue la librería `dplyr` y corra el PCA con la función `pca` de la librería `AMR`, de la siguiente forma:

```
pca2=data %>%
  pca(vallas,saltoalto,tiro,car200,saltolargo,javalina,car800)
```

18. Haga un gráfico de sedimentación y sugiera un número de componentes. Puede usar `plot(pca$sdev^2)` o `plot(pca)`.
    - Decida si el porcentaje de variabilidad explicada por esa cantidad de componentes es adecuado.
    - Verifique si estos componentes cumplen los criterios 2 o 3.
19. Reproduzca la matriz de correlaciones:
    - Usando todos los componentes.
    - Ahora con sólo dos componentes obtenga una predicción de R y vea las discrepancias. Para facilidad vea las discrepancias con solo un decimal.
    - Haga lo mismo con tres componentes y decida si vale la pena usar 3 componentes.
20. Obtenga la correlación entre cada variable original y Z1a o Z1b (use los datos).
    - Obtenga las correlaciones a partir del primer valor propio y el primer vector propio. Compare los resultados.
    - Observe los valores de Z1a y Z1b para comprender el significado del valor que obtuvo la atleta que tuvo el mejor rendimiento.
21. Realice una regresión usando los dos primeros componentes como predictores y el puntaje como respuesta.
    - ¿Cuál de los dos componentes está influyendo más en el puntaje final del atleta?
    - ¿Tiene alguna importancia el signo del coeficiente del Z1?
    - Escriba la ecuación de regresión resultante en términos de las variables originales.