Act 4 PCA

Andrés Villarreal González

2024-10-08

Act 4 PCA

```
data <- read.csv("corporal.csv")
data <- subset(data, select = -sexo)
head(data)
```

```
##   edad peso altura muneca biceps
## 1   43 87.3  188.0   12.2   35.8
## 2   65 80.0  174.0   12.0   35.0
## 3   45 82.3  176.5   11.2   38.5
## 4   37 73.6  180.3   11.2   32.2
## 5   55 74.1  167.6   11.8   32.9
## 6   33 85.9  188.0   12.4   38.5
```

Parte 1

Calcule las matrices de varianza-covarianza S con cov(X) y la matriz de correlaciones R con cor(X) y realice los siguientes pasos con cada una:

```
S <- cov(data)
R <- cor(data)
```

Calcule los valores y vectores propios de cada matriz.La función en R es: eigen().

```
# Valores y vectores propios de S
eigen_S <- eigen(S)
valores_propios_S <- eigen_S$values
vectores_propios_S <- eigen_S$vectors

# Valores y vectores propios de R
eigen_R <- eigen(R)
valores_propios_R <- eigen_R$values
vectores_propios_R <- eigen_R$vectors
```

Calcule la proporción de varianza explicada por cada componente en ambas matrices. Se sugiere dividir cada lambda entre la varianza total (las lambdas están en eigen(S)$values). La varianza total es la suma de las varianzas de la diagonal de S. Una forma es sum(diag(S)). La varianza total de los componentes es la suma de los valores propios (es decir, la suma de la varianza de cada componente), sin embargo, si sumas la diagonal de S (es decir, la varianza de cada x), te da el mismo valor (¡comprúebalo!). Recuerda que las combinaciones lineales buscan reproducir la varianza de X.

```
# Para la matriz S
# Varianza total (suma de las varianzas de las variables originales)
varianza_total_S <- sum(diag(S))

# Proporción de varianza explicada por cada componente
prop_varianza_S <- valores_propios_S / varianza_total_S

# Varianza acumulada
varianza_acumulada_S <- cumsum(prop_varianza_S)
varianza_acumulada_S
```

```
## [1] 0.7615357 0.9318456 0.9903763 0.9995034 1.0000000
```

```
# Para la matriz R
# Varianza total (en una matriz de correlaciones es igual al número de
variables)
varianza_total_R <- sum(diag(R))

# Proporción de varianza explicada por cada componente
prop_varianza_R <- valores_propios_R / varianza_total_R

# Varianza acumulada
varianza_acumulada_R <- cumsum(prop_varianza_R)
varianza_acumulada_R
```

```
## [1] 0.7514995 0.8966708 0.9607368 0.9856605 1.0000000
```

Verificar que la suma de los valores propios es igual a la varianza total

```
# Suma de los valores propios
suma_valores_propios_S <- round(sum(valores_propios_S),4)

# Comprobación
suma_valores_propios_S
```

```
## [1] 471.9385
```

```
varianza_total_S
```

```
## [1] 471.9385
```

```r
# Suma de los valores propios
suma_valores_propios_R <- round(sum(valores_propios_R),4)

# Comprobación
suma_valores_propios_R
```

```
## [1] 5
```

```r
varianza_total_R
```

```
## [1] 5
```

```r
# Crear un data frame con los resultados
resultados_S <- data.frame(
  Componente = 1:length(valores_propios_S),
  Valor_Propio = valores_propios_S,
  Prop_Varianza_Explicada = round(prop_varianza_S, 4),
  Varianza_Acumulada = round(varianza_acumulada_S, 4)
)

# Mostrar la tabla
print(resultados_S)
```

```
##   Componente Valor_Propio Prop_Varianza_Explicada Varianza_Acumulada
## 1          1  359.3980243                  0.7615             0.7615
## 2          2   80.3757858                  0.1703             0.9318
## 3          3   27.6229011                  0.0585             0.9904
## 4          4    4.3074318                  0.0091             0.9995
## 5          5    0.2343571                  0.0005             1.0000
```

```r
# Crear un data frame con los resultados
resultados_R <- data.frame(
  Componente = 1:length(valores_propios_R),
  Valor_Propio = valores_propios_R,
  Prop_Varianza_Explicada = round(prop_varianza_R, 4),
  Varianza_Acumulada = round(varianza_acumulada_R, 4)
)

# Mostrar la tabla
print(resultados_R)
```

```
##   Componente Valor_Propio Prop_Varianza_Explicada Varianza_Acumulada
## 1          1   3.75749733                  0.7515             0.7515
## 2          2   0.72585665                  0.1452             0.8967
## 3          3   0.32032981                  0.0641             0.9607
## 4          4   0.12461873                  0.0249             0.9857
## 5          5   0.07169749                  0.0143             1.0000
```

```r
varianza_acumulada_R
```

```
## [1] 0.7514995 0.8966708 0.9607368 0.9856605 1.0000000
```

```
varianza_acumulada_S
```

```
## [1] 0.7615357 0.9318456 0.9903763 0.9995034 1.0000000
```

Segun los resultados anteriores los componentes mas importantes son los primeros 3 ya que en el caso de R representa el 96% de la varianza y en el caso de S representa el 99% de la varianza.

Escriba la ecuación de la combinación lineal de los Componentes principales CP1 y CP2 (eiX, donde ei está en eigen(S)$vectors[1], e2X para obtener CP2, donde X = c(X1, X2, ...)) ¿qué variables son las que más contribuyen a la primera y segunda componentes principales? (observe los coeficientes en valor absoluto de las combinaciones lineales). Justifique su respuesta.

```
# Vectores propios de S
vectores_propios_S <- eigen(S)$vectors

# Primer vector propio (CP1)
e1 <- vectores_propios_S[,1]

# Segundo vector propio (CP2)
e2 <- vectores_propios_S[,2]


# Vectores propios de S
vectores_propios_R <- eigen(R)$vectors

# Primer vector propio (CP1)
e11 <- vectores_propios_R[,1]

# Segundo vector propio (CP2)
e22 <- vectores_propios_R[,2]

e1
```

```
## [1] -0.34871002 -0.76617586 -0.47632405 -0.05386189 -0.24817367
```

```
e2
```

```
## [1]  0.9075501 -0.1616581 -0.3851755  0.0155423 -0.0402221
```

```
e11
```

```
## [1] -0.3359310 -0.4927066 -0.4222426 -0.4821923 -0.4833139
```

```
e22
```

```
## [1]  0.8575601 -0.1647821 -0.4542223  0.1082775 -0.1392684
```

Para S las variables que mas contribuyen en PC1 son X2, X3 y X1 (peso, altura, edad)

Para S las variables que mas contribuyen en PC2 son X1, X3, X2 (edad, altura, peso)

Para R las variables que mas contribuyen en PC1 son X2, X5, X4 (peso, biceps, muñeca)

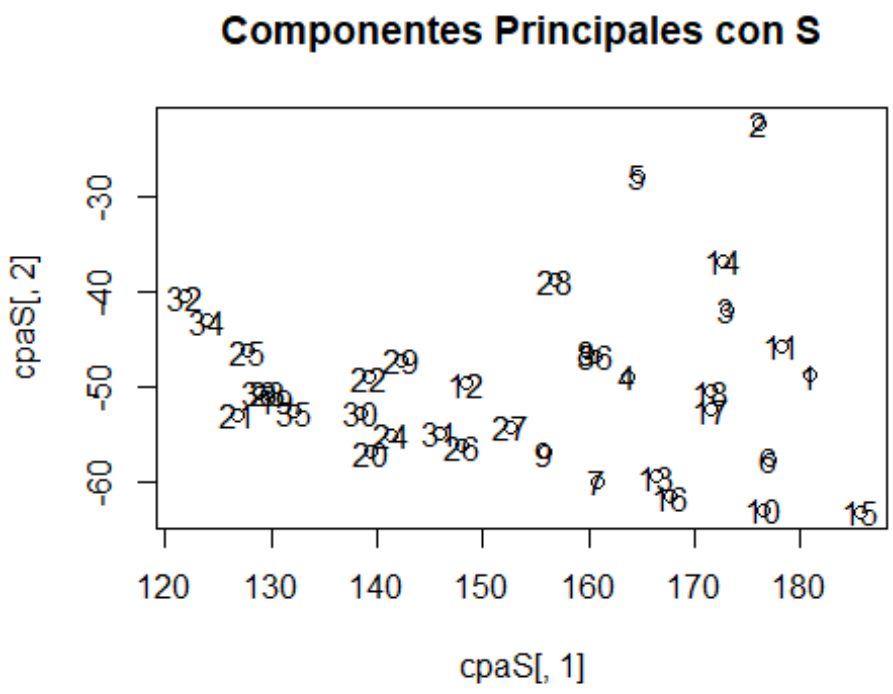
Para R las variables que mas contribuyen en PC2 son X1, X3, X2 (edad, altura, peso)

Parte 2

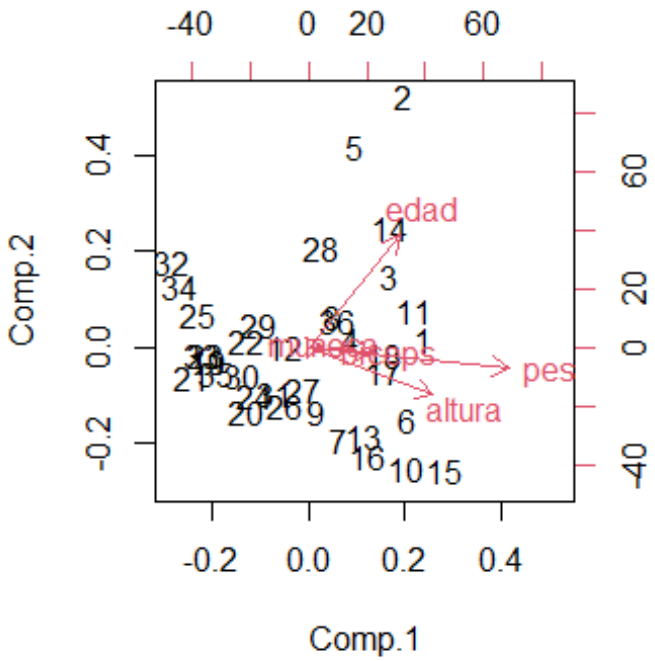
Obtenga las gráficas de respectivas con S (matriz de varianzas-covarianzas) y con R (matriz de correlaciones) de las dos primeras componentes.

Matriz de varianzas-covarianzas

```
# Realizar PCA en la matriz de datos con matriz de varianzas-covarianzas (S)
cpS <- princomp(data, cor = FALSE)
cpaS <- as.matrix(data) %*% cpS$loadings

# Gráfica de las dos primeras componentes con matriz de varianzas-covarianzas (S)
plot(cpaS[, 1], cpaS[, 2], type = "p", main = "Componentes Principales con S")
text(cpaS[, 1], cpaS[, 2], 1:nrow(cpaS))
```



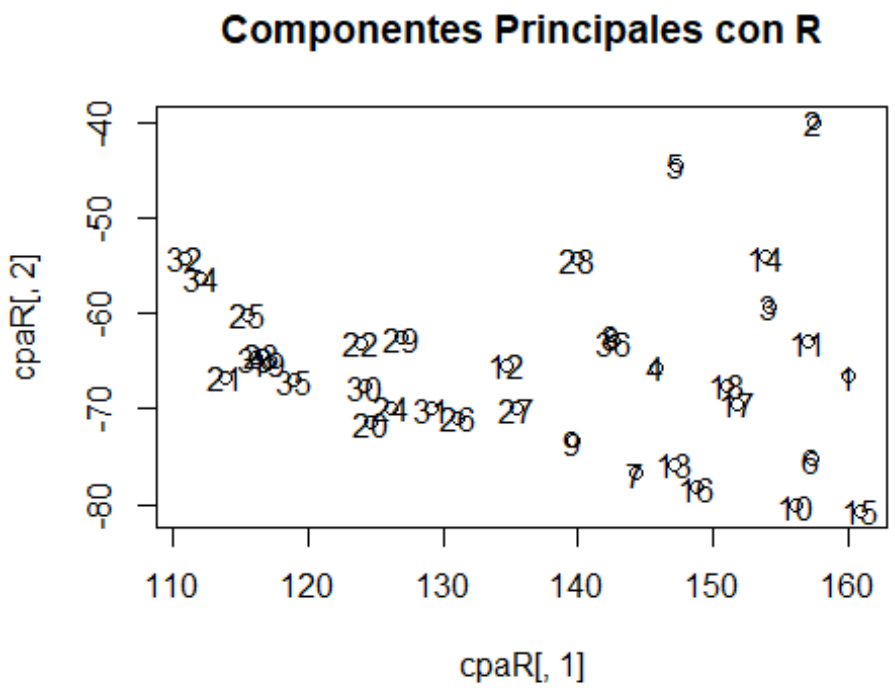

```
biplot(cpS)
```



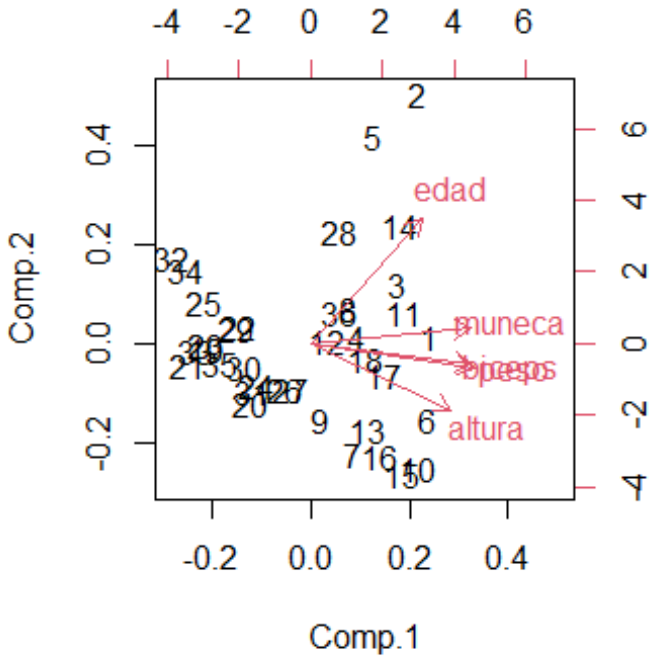
En el grafico se puede observar como se había visto anteriormente que las 3 variables mas significativas para S son altura, peso y edad para el componente 1 y el 2

Matriz de correlaciones

```
# Realizar PCA en la matriz de datos con matriz de correlaciones (R)
cpR <- princomp(data, cor = TRUE)
cpaR <- as.matrix(data) %*% cpR$loadings

# Gráfica de las dos primeras componentes con matriz de correlaciones (R)
plot(cpaR[, 1], cpaR[, 2], type = "p", main = "Componentes Principales
con R")
text(cpaR[, 1], cpaR[, 2], 1:nrow(cpaR))
```




```
biplot(cpR)
```



En el grafico se puede observar como se había visto anteriormente que las 3 variables mas

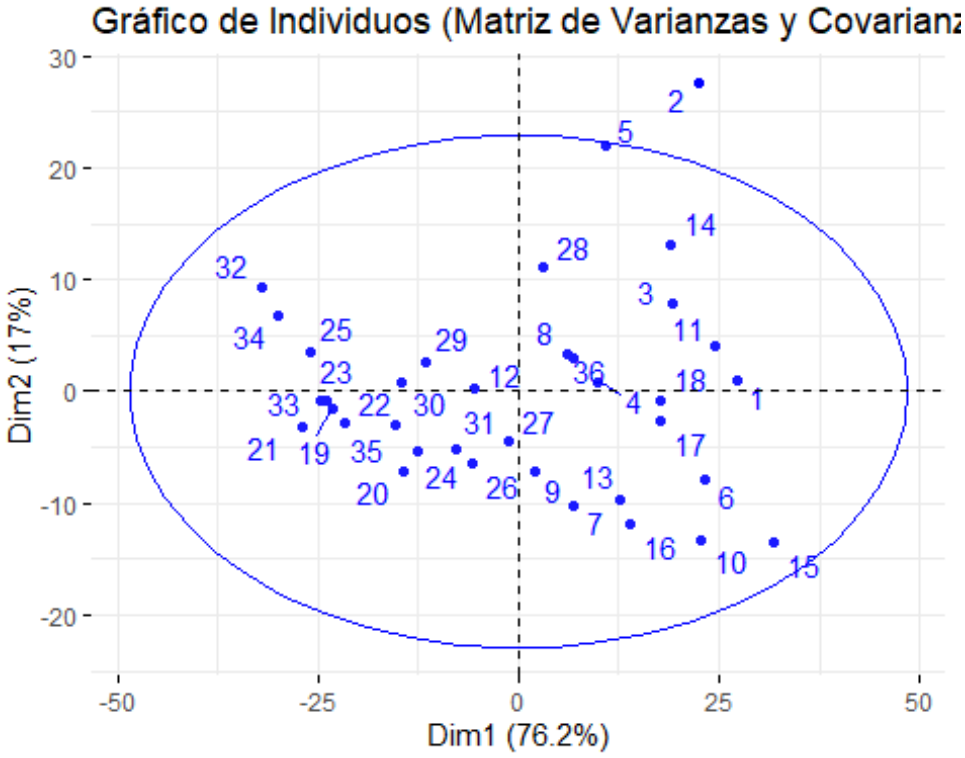
significativas para R son para el componente 1 son peso, biceps, muñeca, aunque todas parecen ser significativas y para el componente 2 serían altura, edad y peso.

Parte 3

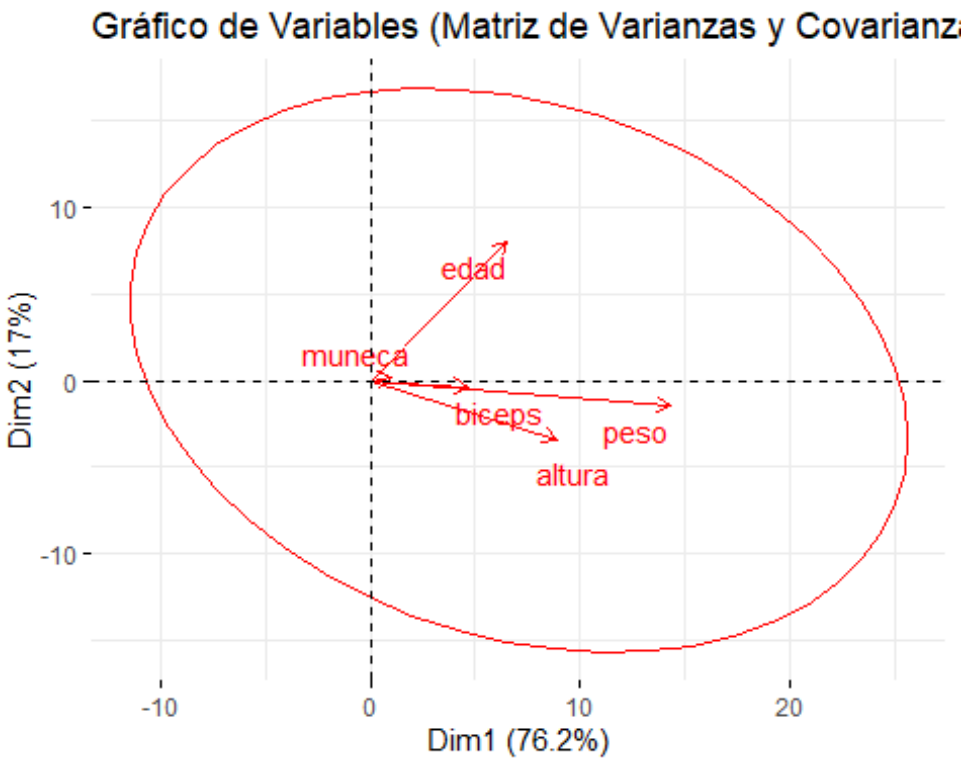
Explore los siguientes gráficos relativos a Componentes Principales.

Matriz de varianzas y covarianzas

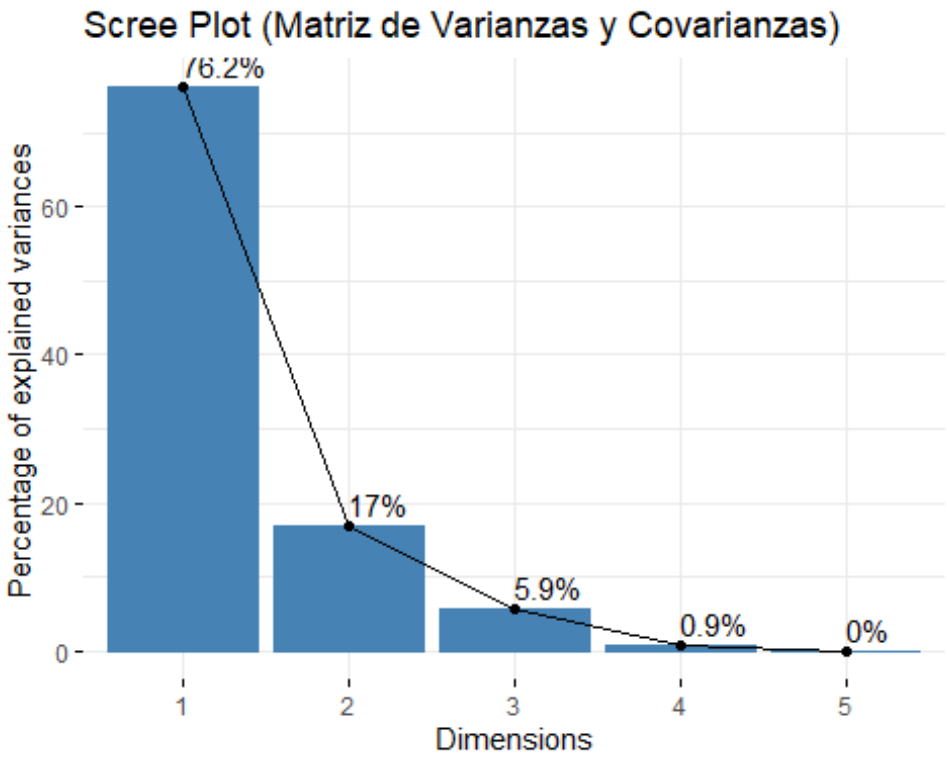
```
library(FactoMineR)
library(factoextra)

## Loading required package: ggplot2

## Welcome! Want to learn more? See two factoextra-related books at
https://goo.gl/ve3WBa

library(ggplot2)
cpS <- PCA(data, scale.unit = FALSE, graph = FALSE)
fviz_pca_ind(cpS, col.ind = "blue", addEllipses = TRUE, repel = TRUE,
title = "Gráfico de Individuos (Matriz de Varianzas y Covarianzas)")
```



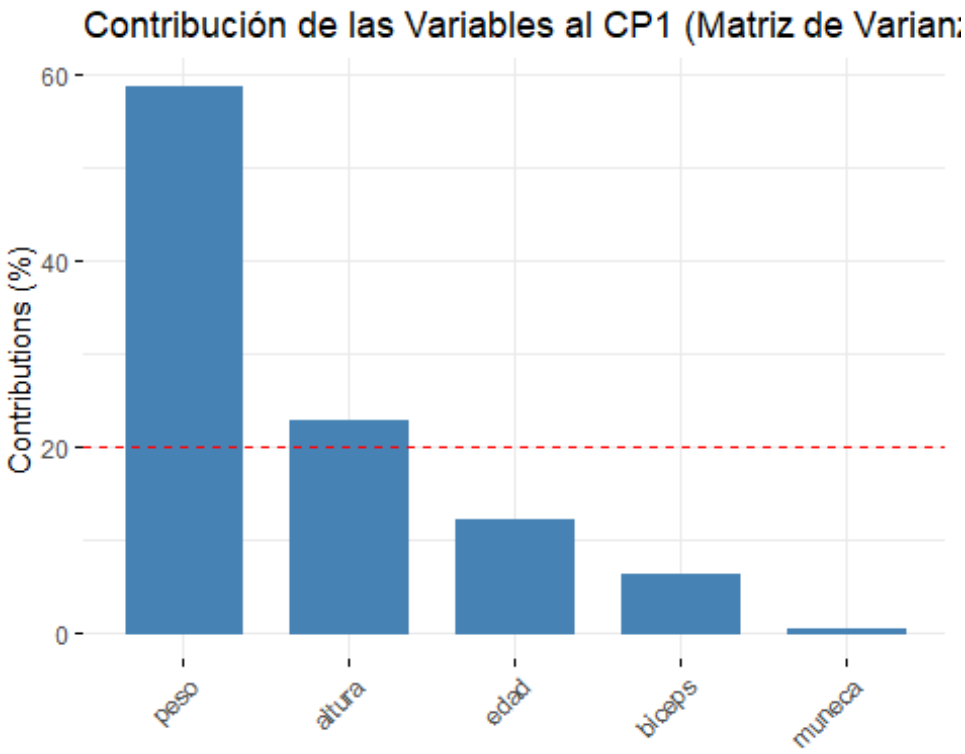

```
fviz_pca_var(cpS, col.var = "red", addEllipses = TRUE, repel = TRUE,
title = "Gráfico de Variables (Matriz de Varianzas y Covarianzas)")
```




```
fviz_screeplot(cpS, addlabels = TRUE, title = "Scree Plot (Matriz de
Varianzas y Covarianzas)")
```



```r
fviz_contrib(cpS, choice = "var", axes = 1, top = 10, title =
"Contribución de las Variables al CP1 (Matriz de Varianzas y
Covarianzas)")
```




```r
fviz_pca_biplot(cpS, repel = TRUE, col.var = "red", col.ind = "blue",
title = "Biplot de PCA (Matriz de Varianzas y Covarianzas)")
```

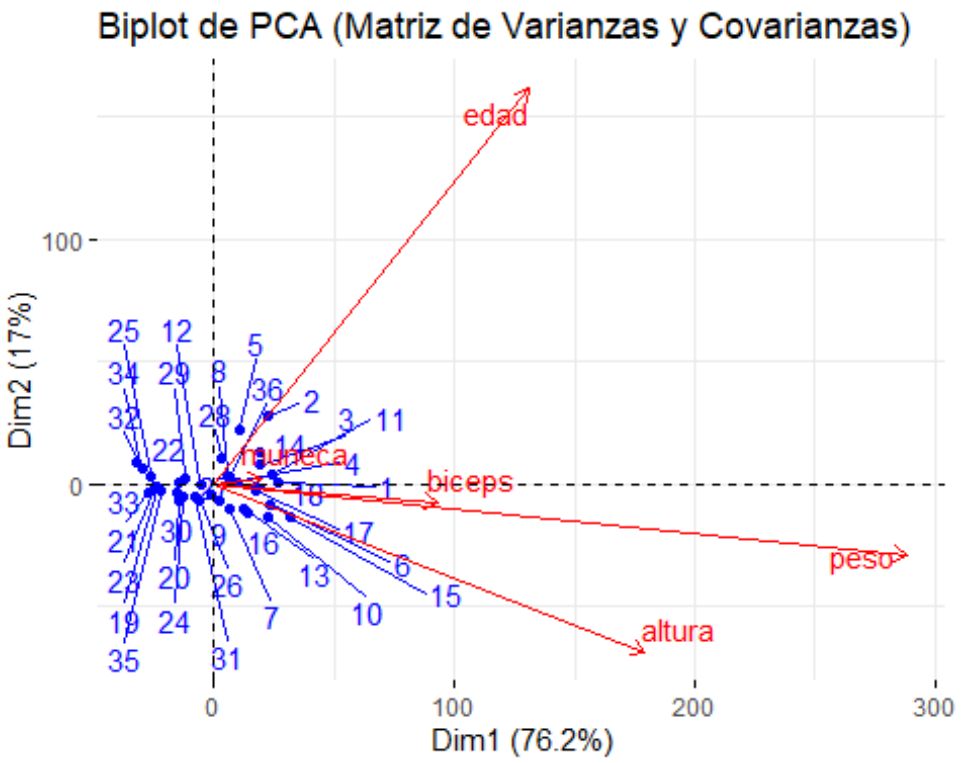


Grafico de individuos: La primera dimensión explica el 76.2% de la varianza, mientras que la segunda dimensión explica el 17%.

Grafico de variables: El gráfico muestra cómo las variables están alineadas con las dos primeras dimensiones. Las variables peso y altura están fuertemente correlacionadas con la primera dimensión (Dim1), lo que indica que estas variables son las que más contribuyen a la variabilidad explicada por Dim1. La variable edad parece estar alineada más hacia Dim2, lo que sugiere que aporta mayor información en esta dimensión secundaria.

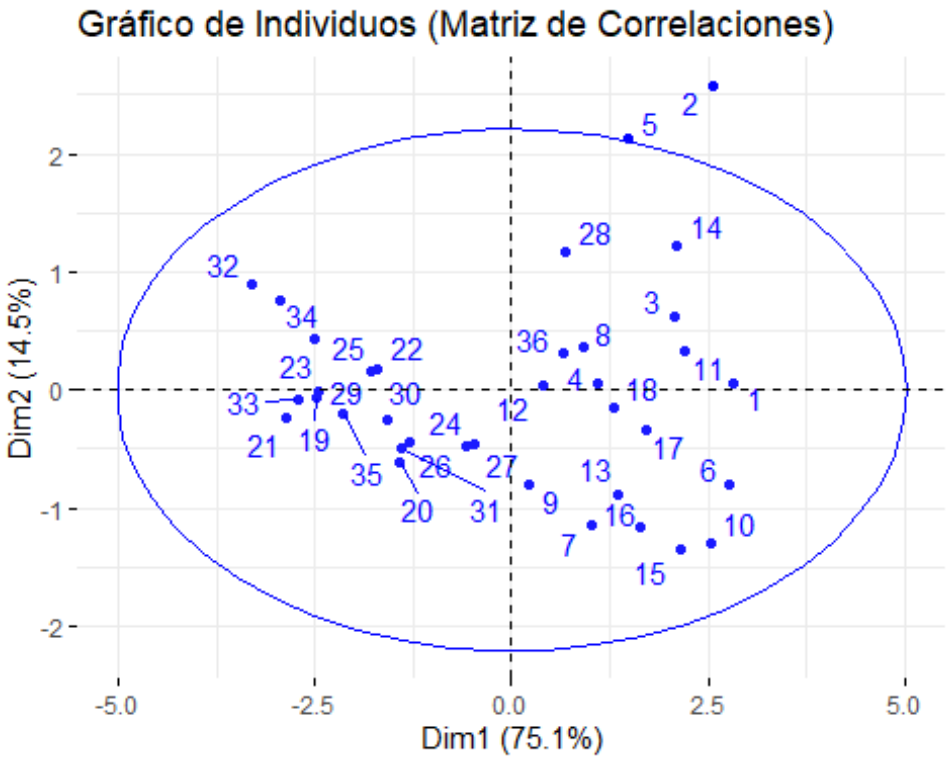
Scree plot: La primera componente principal (CP1) explica la mayor parte de la varianza (76.2%), mientras que la segunda componente principal explica el 17%. Las demás dimensiones explican mucho menos (por debajo del 6%). Este gráfico sugiere que con solo las dos primeras componentes principales, ya se puede explicar más del 90% de la variabilidad en los datos.

Contribución de las Variables al CP1: Peso es la variable que más contribuye a la primera componente principal, con cerca del 60%. Le sigue altura, que también tiene una contribución significativa. Las variables edad, biceps, y muñeca tienen una contribución menor en comparación con peso y altura, lo que indica que estas últimas son las que más influyen en la variabilidad captada por CP1.

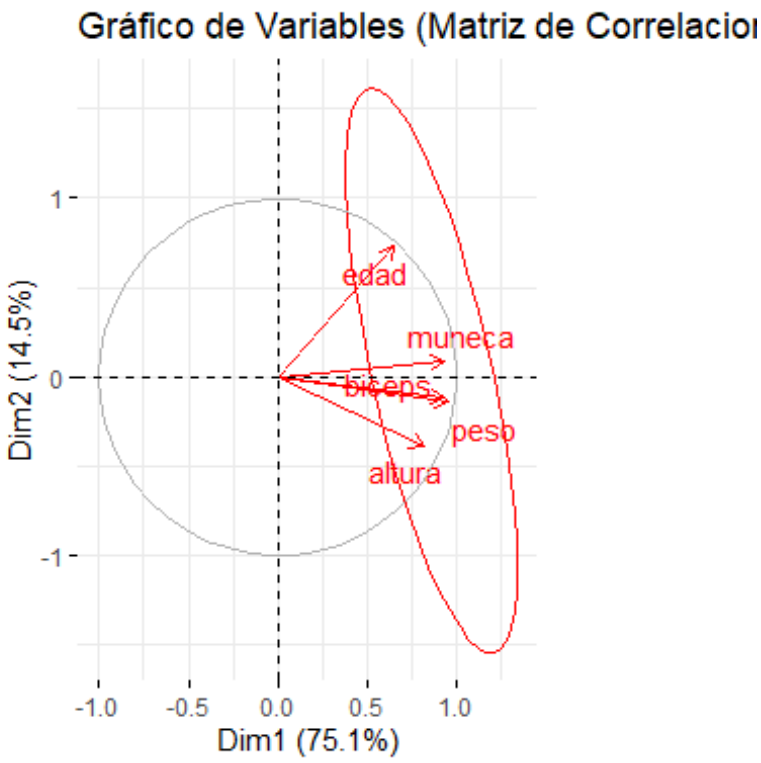
Biplot de PCA: Aquí se puede observar que los individuos con valores altos en Dim1 tienden a estar más relacionados con valores altos de peso y altura, mientras que la variable edad parece ser más relevante para la segunda dimensión (Dim2).

Matriz de correlaciones

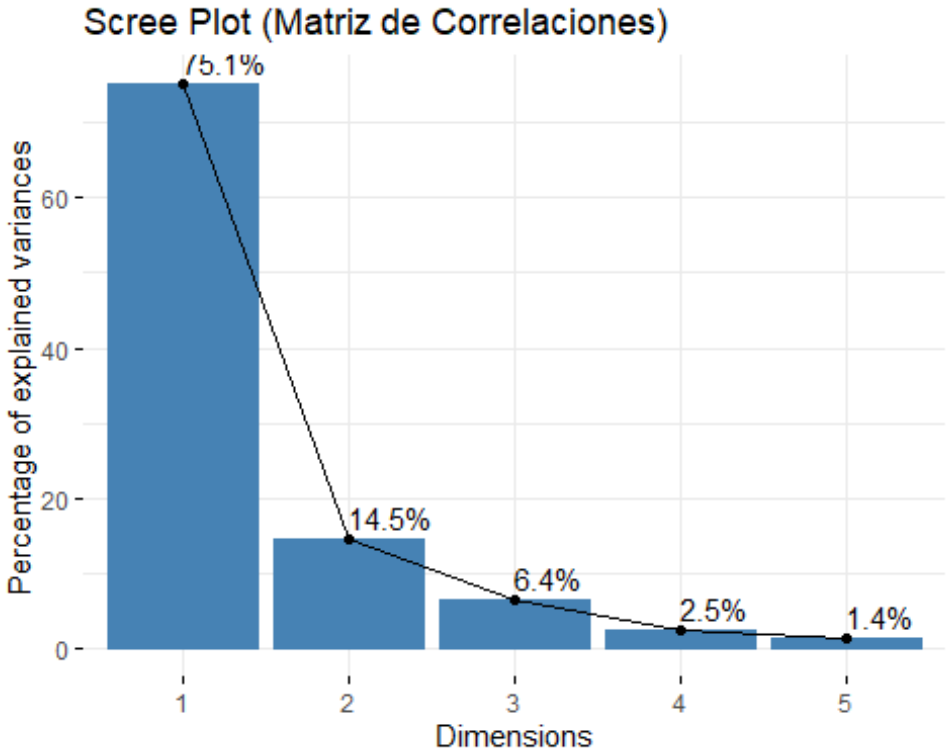
```
cpR <- PCA(data, scale.unit = TRUE, graph = FALSE)
fviz_pca_ind(cpR, col.ind = "blue", addEllipses = TRUE, repel = TRUE,
title = "Gráfico de Individuos (Matriz de Correlaciones)")
```



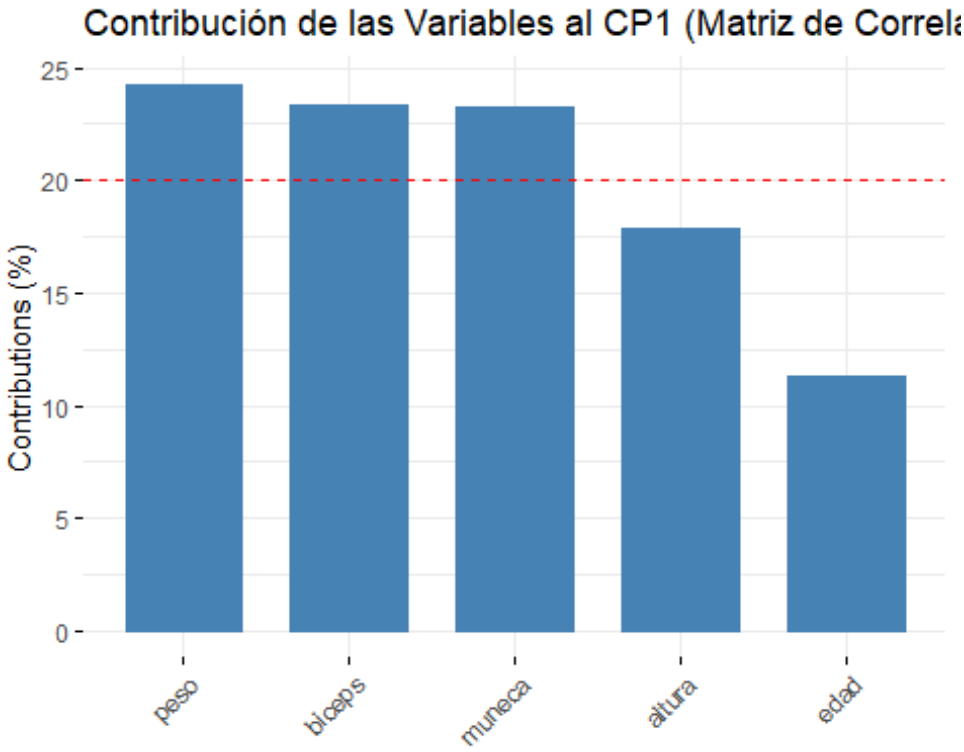

```
fviz_pca_var(cpR, col.var = "red", addEllipses = TRUE, repel = TRUE,
title = "Gráfico de Variables (Matriz de Correlaciones)")
```




```r
fviz_screeplot(cpR, addlabels = TRUE, title = "Scree Plot (Matriz de Correlaciones)")
```



```r
fviz_contrib(cpR, choice = "var", axes = 1, top = 10, title =
"Contribución de las Variables al CP1 (Matriz de Correlaciones)")
```




```r
fviz_pca_biplot(cpR, repel = TRUE, col.var = "red", col.ind = "blue",
title = "Biplot de PCA (Matriz de Correlaciones)")
```

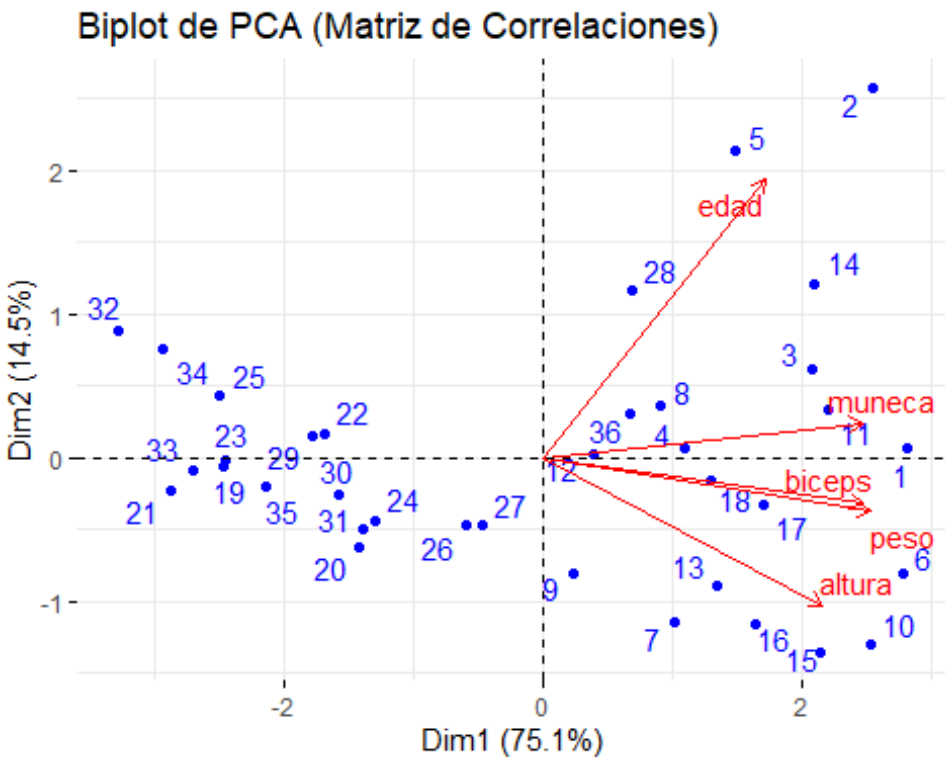


Grafico de individuos: En este gráfico de individuos, las dos primeras dimensiones explican el 75.1% y el 14.5% de la varianza, respectivamente.

Grafico de variables: Las variables están bien distribuidas, lo que significa que no hay una sola variable dominante que explique toda la varianza en las primeras dos dimensiones.

Scree plot: Este gráfico muestra que las dos primeras dimensiones explican la mayor parte de la varianza. La primera dimensión explica el 75.1% y la segunda el 14.5%. Las siguientes dimensiones explican porcentajes menores (6.4% y menos).

Contribución de las Variables al CP1: Las variables peso, biceps, y muñeca tienen contribuciones similares y son las que más aportan a la primera componente principal (CP1).

Biplot de PCA: Los individuos con valores más altos en la dimensión 1 (Dim1) tienden a tener valores altos en peso, biceps, muñeca y altura. La edad está más relacionada con la segunda dimensión (Dim2), por lo que los individuos en la parte superior del gráfico probablemente tienen una mayor edad.

Parte 4

Compare los resultados obtenidos con la matriz de varianza-covarianza y con la correlación . ¿Qué concluye? ¿Cuál de los dos procedimientos aporta componentes con de mayor interés?

Matriz de varianza-covarianza: Los resultados muestran que las variables con mayor variabilidad (como el peso y la altura) son las que más contribuyen a las primeras componentes principales. La primera componente principal (CP1) explica el 76.2% de la varianza, lo que indica que con esta matriz se está capturando mucha información en la primera dimensión, pero está fuertemente influenciada por las variables de mayor magnitud.

Matriz de correlación: La primera componente principal (CP1) explica el 75.1% de la varianza, lo cual es muy similar a la matriz de varianza-covarianza, pero las variables que más contribuyen están más balanceadas, dándole más peso a variables como biceps y muñeca, que tienen menor escala en términos absolutos.

¿Qué variables son las que más contribuyen a la primera y segunda componentes principales del método seleccionado?

PC1:

En el análisis con varianza-covarianza, las variables que más contribuyen son peso y altura, debido a que tienen mayor magnitud en términos de varianza.

En el análisis con correlaciones, las variables que más contribuyen son peso, biceps, y muñeca, mostrando una mayor igualdad en la contribución de diferentes medidas corporales.

PC2:

En ambos métodos, la variable edad es la que más contribuye a la segunda componente principal, lo que indica que esta variable captura una variabilidad importante que no está explicada en la primera componente.

Escriba las combinaciones finales que se recomiendan para hacer el análisis de componentes principales.

CP1 = X2(peso) + X4(muñeca) + X5(biceps) + X3(altura)

CP2 = X1(edad) + X5(biceps) + X3(altura)