# Unidad 4 Actividad 1: Análisis de Componentes Principales

Especialidad en Métodos Estadísticos - CIMAT

Edgar Anuar Sánchez Hernández

2023-02-09

## Indicaciones

Suponga que está encargado de realizar un Análisis de Componentes Principales de los datos proporcionados para responder a los siguientes cuestionamientos.

## 1)

¿Cuál es el objeto, individuo o fenómeno que se está analizando en el caso planteado?

## Respuesta 1)

Se están analizando las condiciones de bienestar de los hogares de México. Se analizan datos sobre los servicios y bienes que a juicio del equipo consideran relevantes para medir el bienestar de los hogares en las 32 entidades federativas.

## 2)

Identifique las variables bajo estudio y, para cada una de ellas, mencione el tipo de variables al que pertenece y la escala de medición empleada.

## Respuesta 2)

Todas las variables numéricas reflejan porcentajes de hogares en cada entidad federativa que cuentan con internet, televisión, servicio de TV de paga, entre otros 8 servicios más. Estas variables son continuas y de escala de razón.

La única variable categórica es la entidad federativa y es nominal.

## 3) Análisis exploratorio de datos

### a)

¿Cuáles son los tres servicios/bienes que presentan una mayor dispersión entre los valores reportados por los Estados?

**Respuesta a)**

Para esto podemos calcular la varianza de los datos para cada porcentaje a lo largo de todas las entidades federativas. Esto significa obtener el vector de varianzas.

Importando los datos:

```
library(readxl)
library(reshape2)
library(ggplot2)

Hogares_equipo <- read_excel("Hogares_equipo.xlsx")
vars <- Hogares_equipo[,2:11]
matrizDatos <- as.matrix(vars)
summary(matrizDatos)
```

```
##     Internet        Television       TV de paga       Telefonia
##  Min.   :0.1323   Min.   :0.8006   Min.   :0.3058   Min.   :0.7229
##  1st Qu.:0.3748   1st Qu.:0.9146   1st Qu.:0.4972   1st Qu.:0.8771
##  Median :0.4870   Median :0.9327   Median :0.5774   Median :0.9156
##  Mean   :0.4793   Mean   :0.9258   Mean   :0.5609   Mean   :0.8973
##  3rd Qu.:0.5886   3rd Qu.:0.9519   3rd Qu.:0.6153   3rd Qu.:0.9445
##  Max.   :0.7577   Max.   :0.9874   Max.   :0.7658   Max.   :0.9778
##      Radio          Automovil         Lavadora      Estufa de gas o electrica
##  Min.   :0.3884   Min.   :0.1299   Min.   :0.3957   Min.   :0.6100
##  1st Qu.:0.5182   1st Qu.:0.2198   1st Qu.:0.6406   1st Qu.:0.8565
##  Median :0.5903   Median :0.2874   Median :0.7240   Median :0.9396
##  Mean   :0.5872   Mean   :0.2935   Mean   :0.6807   Mean   :0.8918
##  3rd Qu.:0.6575   3rd Qu.:0.3641   3rd Qu.:0.7731   3rd Qu.:0.9535
##  Max.   :0.7936   Max.   :0.5085   Max.   :0.8407   Max.   :0.9746
##   Refrigerador    Horno de microondas
##  Min.   :0.6200   Min.   :0.1733
##  1st Qu.:0.8313   1st Qu.:0.3345
##  Median :0.8863   Median :0.4405
##  Mean   :0.8632   Mean   :0.4314
##  3rd Qu.:0.9300   3rd Qu.:0.5312
##  Max.   :0.9617   Max.   :0.6726
```

```
#diag(var(vars))
sort(diag(var(matrizDatos)))
```

```
##               Television                 Telefonia              Refrigerador
##               0.001891667               0.004267926               0.007416686
##                 Automovil Estufa de gas o electrica                     Radio
##               0.008507908               0.009044734               0.010125480
##                TV de paga                  Lavadora       Horno de microondas
##               0.011291633               0.017019792               0.019306812
##                   Internet
##               0.023043964
```

Con esto observamos que los 3 servicios que presentan una mayor dispersión entre los valores reportados por los estados son: Internet, Horno de microondas y Lavadora. En este orden.

**b)**

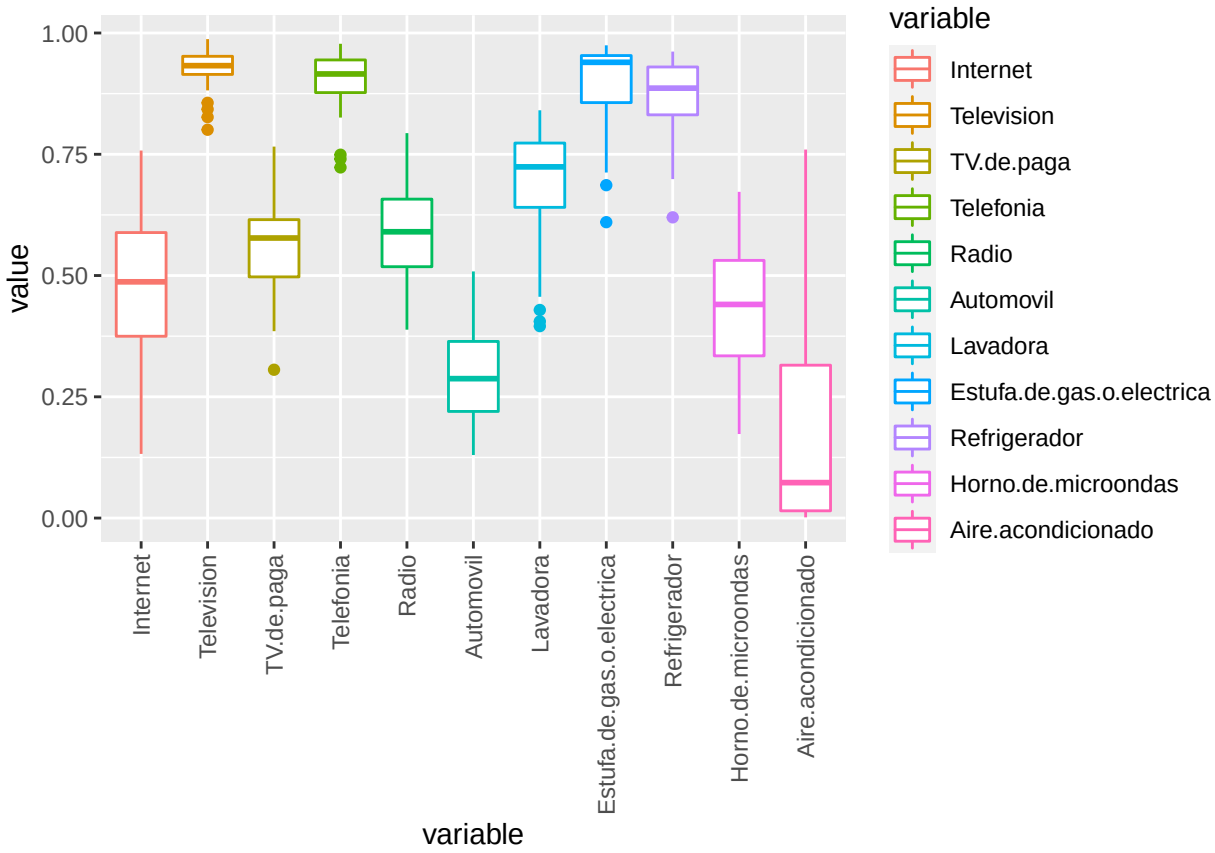
¿Existen servicios/bienes que presentan datos atípicos?

**Respuesta b)**

Elaboramos un diagrama de cajas y bigotes para identificar outliers:

```
#outliers
df <- data.frame(Hogares_equipo)
data_mod <- melt(df, id ="Entidad.federativa")

ggplot(data_mod) +
  geom_boxplot(aes(x=variable, y=value, color=variable)) +
  theme(axis.text.x = element_text(angle=90, vjust=.5, hjust=1))
```




Observamos que las variables: Televisión, TV de paga, Telefonía, Lavadora, Estufa y Refrigerador son las que presentan datos atípicos

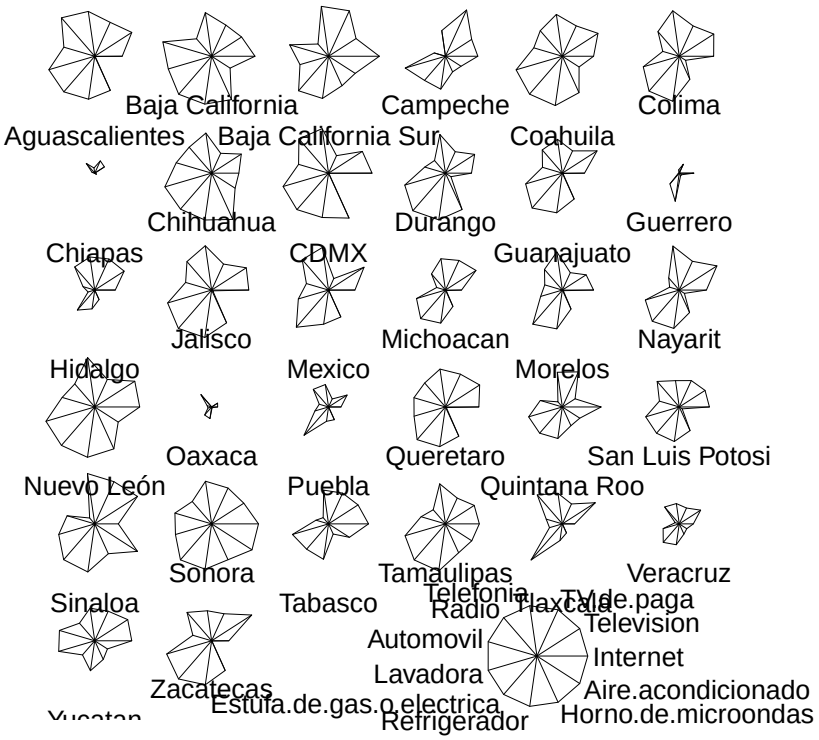
**c)**

¿Es posible identificar Estados con comportamiento similar en los servicios/bienes medidos a los hogares dentro de su territorio?

**Respuesta c)**

Elaboramos una gráfica de polígonos para cada entidad para identificar comportamientos similares.

```
stars(scale(df[2:12]), key.loc = c(11,2), radius = T, labels = df$Entidad.federativa,
      flip.labels = T, len = 1, cex.lab = 0.5, cex.main = 0.5, cex.axis = 0.5)
```




Con esta gráfica identificamos estados con comportamientos similares. Por ejemplo Sonora y Nuevo León tienen porcentajes similares en los 11 porcentajes. De hecho muestran una distribución aproximadamente uniforme de porcentajes de los 11 servicios/bienes.

## 4.

¿Considera que los datos utilizados para el presente estudio son adecuados para realizar un Análisis de Componentes Principales?

## Respuesta 4.

Para responder esta pregunta podemos recurrir al índice de Kaiser-Meyer-Olkin. En R se calcula como sigue:

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
KMO(matrizDatos)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = matrizDatos)
## Overall MSA =  0.8
## MSA for each item =
##                  Internet               Television                TV de paga
##                      0.73                     0.74                      0.79
##                 Telefonia                    Radio                 Automovil
##                      0.80                     0.66                      0.81
##                  Lavadora Estufa de gas o electrica              Refrigerador
##                      0.82                     0.90                      0.89
##       Horno de microondas
##                      0.78
```

Se observa que el valor del índice KMO para los datos es de 0.8, lo cual indica que el ACP será aceptable. Sólo tener en consideración que el índice KMO para la variable Radio es menor a lo mínimo, por lo cual se sugiere no utilizar esta variable.

### 5

Utilizando la matriz de correlación, determine el número de componentes a ser considerados en el estudio si se empleara cada uno de los métodos enlistados:

a. El porcentaje de Variación Total Acumulada (al menos el 75%)
b. El criterio de Kaiser
c. El método gráfico

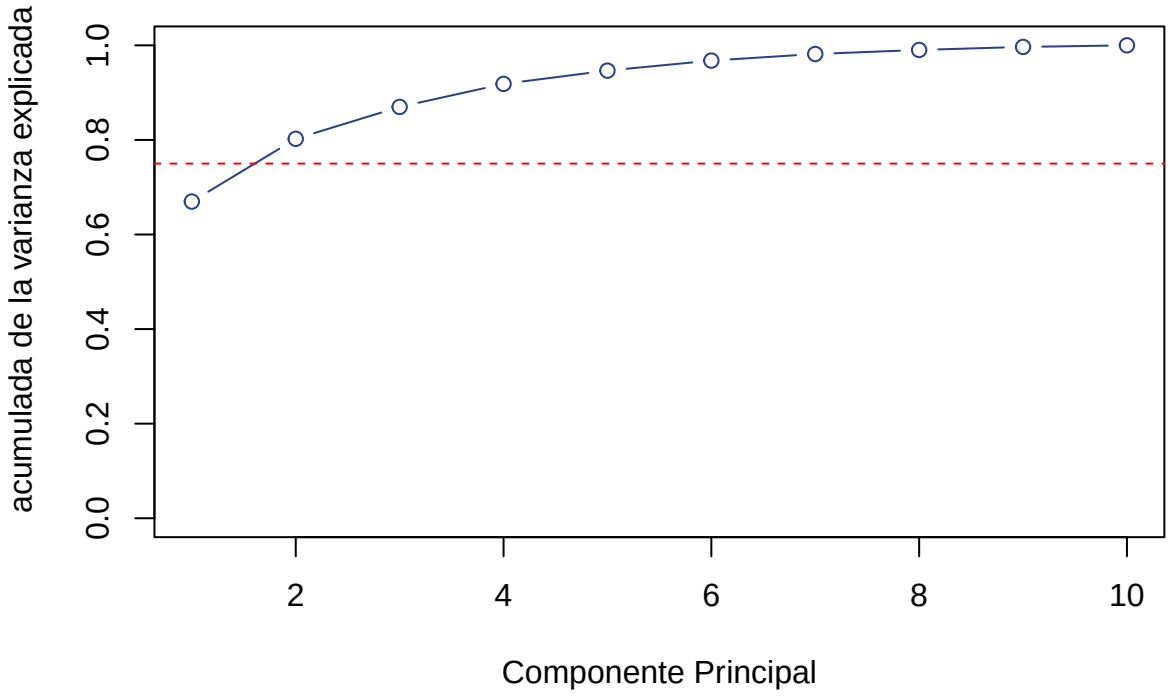
## Respuesta 5.

Mediante R realizamos primero el análisis de componentes principales utilizando la matriz de correlación con el siguiente código:

```
#Análisis de componentes principales
porcentajes_pca <- prcomp(matrizDatos, center = T, scale = T)
```

Ahora realizamos el gráfico de variación acumulada en R:

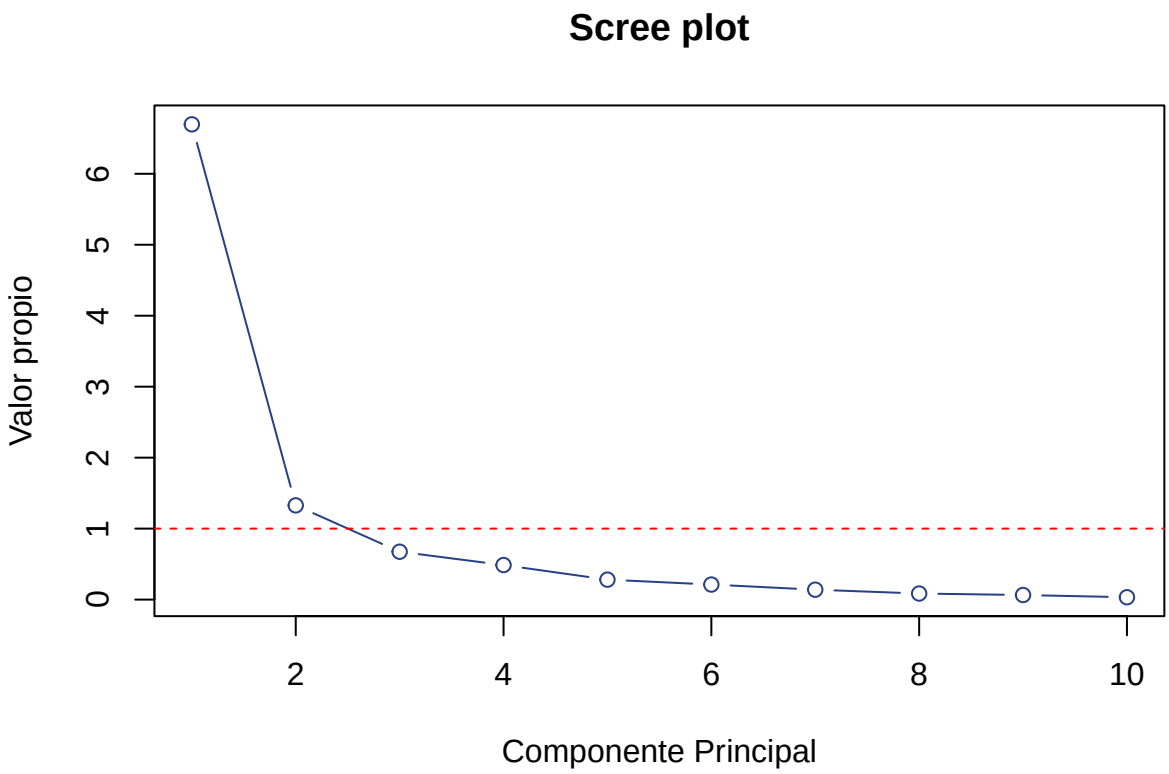
```
pve <- porcentajes_pca$sdev^2/sum(porcentajes_pca$sdev^2)
plot(cumsum(pve), xlab = "Componente Principal", ylab = "Porción
     acumulada de la varianza explicada", ylim = c(0,1), type = 'b', col = "royalblue4")
abline(h = 0.75, lty = 2, col = "red")
```




Se observa que para que el porcentaje de Variación Total Acumulada sea de al menos el 75%, se deben incluir los primeros dos componentes.

Ahora, para utilizar el criterio de Kaiser, realizamos el gráfico de sedimentación (scree plot) en R con el siguiente código:

```
plot(porcentajes_pca$sdev^2, main = "Scree plot", xlab = "Componente Principal",
     ylab = "Valor propio", type = 'b', col = "royalblue4")
abline(h = 1, lty = 2, col = "red")
```




Se observa que los primeros dos valores propios son mayores a 1. Por lo tanto este criterio coincide en que tomemos los dos primeros componentes principales.

Finalmente, el método gráfico indica que debemos utilizar sólo el primer componente principal, ya que el puntio de inflexión más pronunciado se encuentra en el segundo componente principal.

### 6.

El equipo decide trabajar con los dos primeros componentes. Realice la gráfica Bi-plot para estos dos componentes e interprete lo que observa en ella.

a. ¿Existen variables que puedan considerarse contradictorias (opuestas)? Si la respuesta es afirmativa, mencionar al menos un ejemplo.
b. ¿Existen variables que pueda asumirse que no tienen relación entre ellas? Si la respuesta es afirmativa, mencionar al menos un ejemplo.
c. ¿Existen variables con fuerte relación entre ellas? Si la respuesta es afirmativa, mencionar al menos un ejemplo.
d. ¿Cuál es el servicio/bien que, de acuerdo con lo reflejado en el biplot, tiene una mayor relevancia o incidencia en los resultados que se despliegan en el gráfico?

### Respuesta 6.

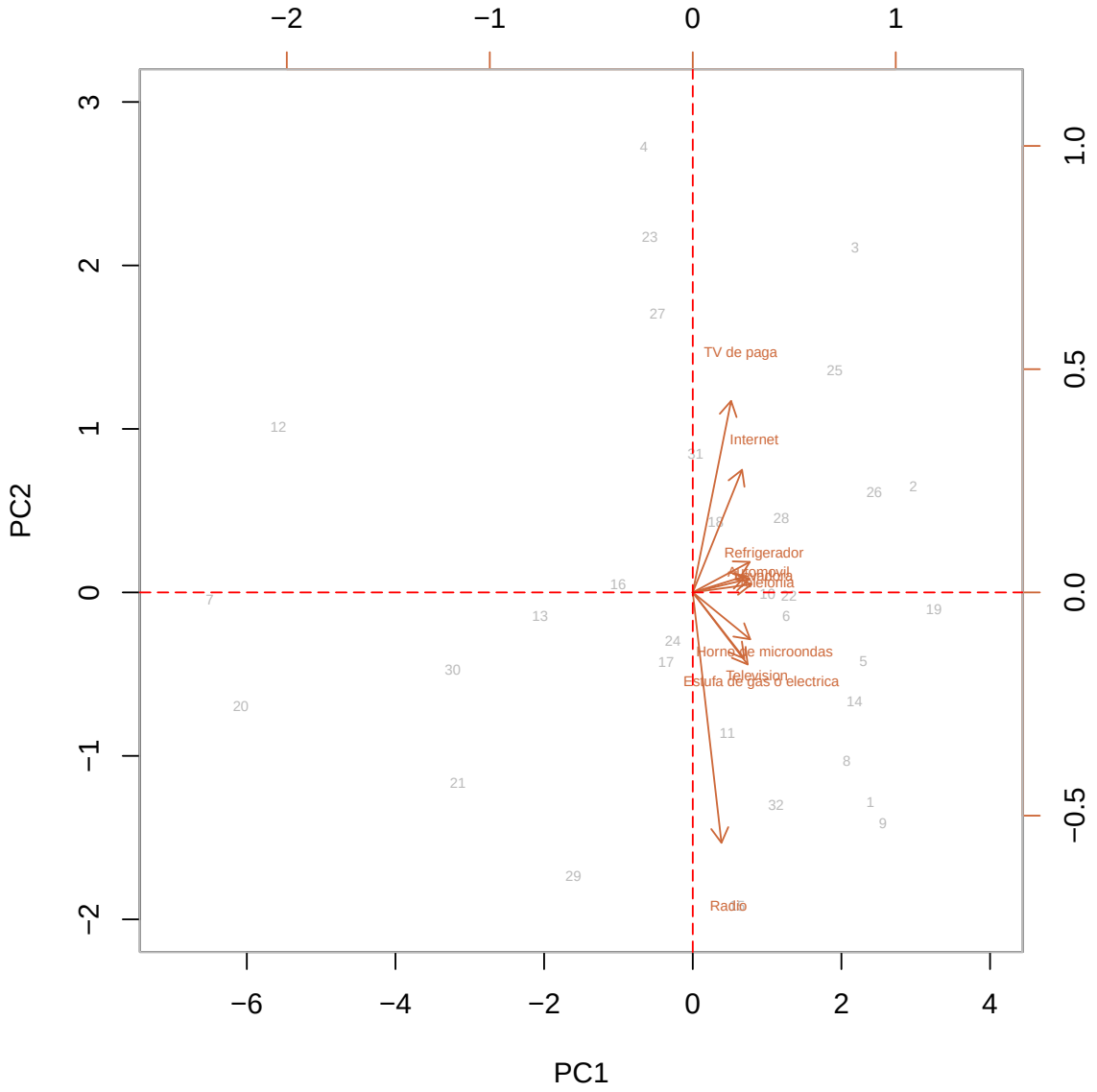
Primero obtendremos los scores de cada observación en los componentes seleccionados, los primeros dos:

```
Z1 <- round(scale(matrizDatos) %*% porcentajes_pca$rotation[,1],4)
Z2 <- round(scale(matrizDatos) %*% porcentajes_pca$rotation[,2],4)
scores <- cbind(Z1, Z2)
colnames(scores) <- c("Z1","Z2")
scores
```

```
##            Z1      Z2
##  [1,]  2.3923 -1.2814
##  [2,]  2.9695  0.6494
##  [3,]  2.1824  2.1116
##  [4,] -0.6579  2.7266
##  [5,]  2.2968 -0.4189
##  [6,]  1.2570 -0.1469
##  [7,] -6.4975 -0.0418
##  [8,]  2.0715 -1.0318
##  [9,]  2.5587 -1.4119
## [10,]  1.0080 -0.0075
## [11,]  0.4665 -0.8615
## [12,] -5.5751  1.0156
## [13,] -2.0519 -0.1450
## [14,]  2.1770 -0.6683
## [15,]  0.5963 -1.9145
## [16,] -1.0033  0.0513
## [17,] -0.3575 -0.4231
## [18,]  0.3110  0.4306
## [19,]  3.2460 -0.1027
## [20,] -6.0801 -0.6963
## [21,] -3.1590 -1.1634
## [22,]  1.2963 -0.0190
## [23,] -0.5768  2.1769
## [24,] -0.2669 -0.2942
## [25,]  1.9094  1.3605
## [26,]  2.4421  0.6110
## [27,] -0.4732  1.7042
## [28,]  1.1933  0.4546
## [29,] -1.6052 -1.7357
## [30,] -3.2290 -0.4730
## [31,]  0.0384  0.8475
## [32,]  1.1210 -1.3027
```

Con esto realizamos el biplot:

```
biplot(porcentajes_pca, choices = 1:2, col = c("gray73","sienna3"), scale = 0,
       xlim = c(-7,4), ylim = c(-2,3), cex = 0.5)
abline(h = 0, v = 0, col = "red", lty = 5)
```




**Respondiendo las preguntas:**

**a.**

Sí. Por ejemplo la variable Radio y la variable TV de paga están representadas por vectores que forman un ángulo cercano a 180°. Esto significa una correlación negativa entre ellas, o bien que pueden considerarse como opuestas.

Lo mismo podemos indicar con las variables Internet y Radio.

**b.**

Sí. La variable Radio y las variables Refrigerador, Automóvil, Lavadora y Telefonía, están representadas por vectores que forman un ángulo cercaco a 90°. Por lo que se considera que su correlación es prácticamente nula.

**c.**

Sí. Las variables Refrigerador, Automóvil, Lavadora y Telefonía están representadas por vectores cuyos ángulos son cercanos a cero, por lo cual son variables altamente correlacionadas.

Además, las variables Horno de microondas, Televisión y Estufa de gas o eléctrica, forman otro grupo de variables con alto grado de correlación.

Finalmente podemos agregar que TV de paga e Internet también tienen correlación alta.

**d.**

El servicio/bien que tiene una mayor relevancia o incidencia en los resultados que se despliegan en el gráfico es "Radio", al estar representada por el vector de mayor longitud.

## 7.

A partir de la información arrojada después de aplicar el método de ACP, responda y justifique basado en la evidencia ofrecida por el análisis realizado, las siguientes preguntas:

a. ¿Se comprueba que Nuevo León, CDMX y Jalisco presentan un comportamiento similar en las variables medidas? ¿Por qué?

b. ¿Lo mismo ocurre con Chiapas, Guerrero y Oaxaca? ¿Por qué?

c. ¿El desempeño general del Estado de Aguascalientes es similar a la de otros estados? En caso de ser afirmativo mencione al menos uno.

d. Considerando la hipótesis de que un hogar con mayores ingresos contará con la mayoría de los bienes y servicios enlistados,

e. ¿cuál es el Estado que cuenta con los hogares de mayores ingresos? y,

ii. ¿cuál es el Estado que tiene los hogares con ingresos más bajo?

## Respuesta 7.

**a)**

Podemos observar que Nuevo León, CDMX y Jalisco, que son las observaciones 9, 14 y 19, se encuentran en el mismo cuadrante del biplot. Por lo cual se sugiere que tienen comportamiento similar en las variables medidas, aunque no completamente igual.

**b)**

Chiapas, Guerrero y Oaxaca son las observaciones 7, 12 y 20. Localizándolas en el gráfico biplot, se observa que no se encuentran en el mismo cuadrante, sin embargo se ubican relativamente cerca una de otra. Por lo cual concluimos que estas 3 entidades sí tienen comportamientos similares.

**c)**

La entidad de Aguascalientes está representada por la observación 1, que se encuentra cerca de las observaciones 8 y 9, que representan a las entidades de Chihuahua y CDMX. Es decir, estas 3 entidades tienen comportamiento similar.

**d) i)**

Observando el bibplot, podemos deducir que la observación 19 (Nuevo León) representa aquella con el mayor número de bienes en sus hogares (y con mayores ingresos), ya que se encuentra ubicada hacia donde la mayoría de vectores apuntan, en la parte central de las direcciones de estas variables.

**d) ii)**

Del lado diametralmente opuesto a Nuevo León (19) se encuentran Chiapas (7) y Oaxaca (20), que son los estados con menor número de bienes en sus hogares y por lo tanto son los estados con ingresos más bajos.