Análisis Factorial de Correspondencias Multiples

Marlon Enrique Pérez M.

15/2/2021

Índice

¿Qúe es el AFCM?

EL análisis factorial de correspondencias es una técnica estadística que nos permite investigar alguna relación que pueda existir entre variables cualitativas. En este artículo te mostraré la forma más clara y empírica de lo que es, como funciona y que debes considerar para aplicar esta técnica estadística a tus estudios de investigación.

Para que sirve el análisis factorial de correspondencias múltiples

El análisis factorial de correspondencias múltiples nos ayuda muchísimo dentro del análisis exploratorio de los datos. En lo que respecta entre los modelos estadísticos multivariante el AFCM (análisis factorial de correspondencias múltiples) resulta una poderosísima herramienta de procesamiento de datos de características cualitativas y forma parte uno de los instrumentos más relevantes dentro del Proceso de Data Mining dentro de la investigación en las ciencias sociales.

Ventajas del AFC

- Analizar las relaciones existentes en un grupo de atributos observados.
- Tratar la información de la encuesta con el nivel de síntesis adecuado con el marco conceptual utilizado.
- Reducir las dimensiones del fenómeno observado, sin arbitrariedad.
- El AFCM permite evitar los tres errores más comunes que se cometen tratando los datos de una encuesta los cuales son:
    1. sólo se exploran las relaciones bivariadas.

2. se seleccionan las Tablas de Contingencia “interesantes” entre las p x q tablas creadas, Cuyo criterio de selección es el test chi2, sin tener en cuenta que El test de independencia del Chi2 no prueba la “fuerza de la asociación entre las variables.
3. se construyen “arbitrariamente” las tipologías que resumen una “unidad temática”.

- Veremos también que el A.F.C.M. permite crear, sin arbitrariedad, las tipologías buscadas en un análisis.

Procedimiento de aplicación del AFCM y obtención de resultados en R

1° Paso. - Las Tablas estadísticas creadas a partir de una encuesta o cuestionario

Para obtener las tablas estadísticas que nacen de un cuestionario es de vital importancia considerar primero que nada los objetivos de la investigación ya que en base a ellos se desarrollan las preguntas y el orden en el que se plantean las mismas, ya que este junto a otros factores pueden influir en las respuestas. Pero calma este no es el objetivo de este artículo en este enlace te indicaré cómo desarrollar un buen cuestionario de investigación.

Para los objetivos de este artículo supondremos que tenemos un modelo de preguntas en un cuestionario que se ha realizado a dueños de mascotas en el cual se expresan las siguientes variables de la siguiente forma:

- Primera característica observada **Tamaño**
  1.- Pequeño
  2.- Mediano
  3.- Grande

- Segunda característica observada **Peso**
  1.- Liviano
  2.- Moderado
  3.- Pesado

- Tercera característica observada **Velocidad**
  1.- Lento
  2.- Normal
  3.- Rápido

- Cuarta característica observada **Inteligencia**
  1.- Baja
  2.- Moderada
  3.- Alta

- Quinta característica observada **Afección**
  1.- Afectuoso
  2.- Nada afectuoso

- Sexta característica observada **Agresividad**
  1.- No agresivo
  2.- Agresivo

- Séptima característica observada **Función**
  1.- Hogareño
  2.- Caza
  3.- Guardián

Y es en este momento cuando a partir de un cuestionario se crea una tabla de datos la cual es llamada tabla de códigos condensados TCC

Tabla de códigos condensados (TCC)

Cada línea de la TCC (n x p) contiene todos los códigos correspondientes a las modalidades atribuidas a un individuo, para cada una de las características observadas. Pese a ello, una tabla de códigos condensados no posee propiedades numéricas, lo que se quiere decir con esto es que la suma de sus perfiles fila y columna no tienen sentido interpretativo

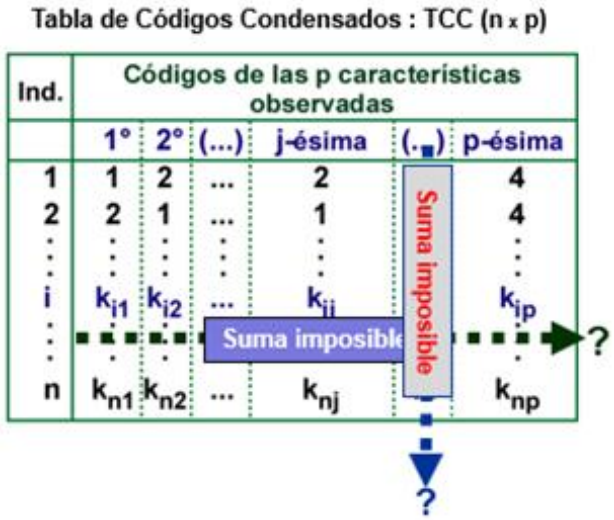


Tabla Disyuntiva completa (TDC)

Y es en este punto cuando surge la necesidad de producir una representación gráfica cuando se transforma una TCC tabla de datos condensados en una tabla lógica de ceros y unos. Considerando que las modalidades u opciones de respuesta deben ser mutuamente excluyentes es decir que para cada pregunta que existen varias opciones de respuesta se pueda responder únicamente una sola opción. Esta tabla lógica es conocida también como una tabla disyuntiva completa TDC la cual es de orden (n x K) siendo K la suma de las modalidades de las P variables o características observadas.

	1era característica modalidades				j-ésima característica modalidades				p-ésima característica modalidades			mar-gen colum-na
IND.	11	(...)	1k	(...)	j1	(...)	jr	(...)	p1	(...)	ps	
	1	...	...	...	...	j	...	...	...	...	K	
1	0	...	1	...	1	...	0	...	1	...	0	p
2	1	...	0	...	0	...	1	...	0	...	1	p
3	0	...	1	...	0	...	1	...	0	...	1	p
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	0	...	1	...	0	x_{ij}	0	...	1	...	0	p
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	1	...	0	...	0	x_{nj}	0	...	0	...	1	p
mar-gen línea	n_1	...	...	...	...	n_j	...	...	...	...	n_K	np

Tabla de Burt matriz de Burt (B)

La tabla de Burt surge de multiplicar la tabla disyuntiva completa por si transpuesta con lo cual se obtendría una matriz de orden (K x K).

B = A x A´

Notas importantes sobre el AFCM

En este punto hay que tener claro que el AFCM se puede aplicar a dos de las tablas anteriormente mencionadas

1. AFCM aplicado a la tabla de BURT

2. AFCM aplicado a la TDC Tabla disyuntiva completa

Teniendo claro que:

- Las modalidades (columnas) de una Tabla Lógica y de una Tabla de Burt tienen el mismo peso.

- La nube de puntos-individuos (definida a partir de una Tabla Lógica) y los baricentros de esos puntos (definidos a partir de una Tabla de Burt) están ubicados en el mismo espacio euclidiano, de K dimensiones

- Los puntos-individuos en la representación de una Tabla Lógica tienen todo el mismo peso, mientras que los puntos-modalidades en la representación de la Tabla de Burt están afectados de un peso que es proporcional a la importancia de la clase.

- En consecuencia, por la propiedad de equivalencia distribucional que cumplen los espacios dotados de la distancia del Chi2…

El análisis de la nube de puntos-columna N(J), baricentros de los individuos de una Tabla Lógica, puede ser hecho mediante el Análisis Factorial de Correspondencias de una Tabla de Burt. Por lo tanto: El Análisis Factorial de Correspondencias de una Tabla Lógica Y de una Tabla de Burt tienen que producir resultados equivalentes

AFCM aplicado a la tabla disyuntiva completa

Una vez obtenidos los datos de la encuesta o cuestionario procedemos a preparar los datos para su respectivo análisis:

```
library(readxl)

## Warning: package 'readxl' was built under R version 4.0.3

ACM_PERRu<- read_excel("C:/Users/k_kep/OneDrive/Escritorio/ACM
PERRuño.xlsx")
View(ACM_PERRu)
vi<-names(ACM_PERRu)[c(3,4,5,6,7,8,9)]
```

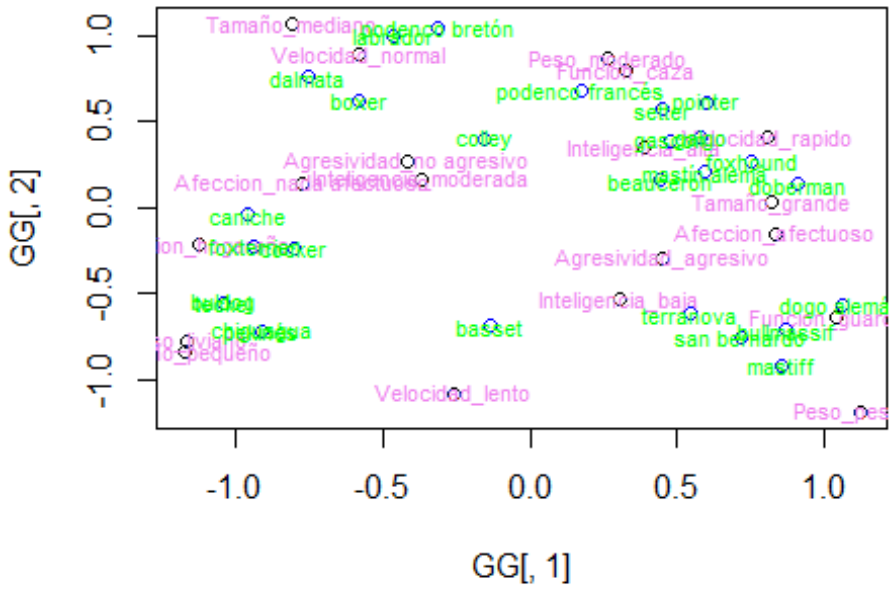
Una vez hecha la lectura de la base de datos la cual es nada más que una tabla de códigos condensados se procede a transformar una base de datos TCC a una TDC para lo cual utilizamos la librería **FastDummies**

```
library(fastDummies)

## Warning: package 'fastDummies' was built under R version 4.0.3

##  SE DEBEN COLOCAR LOS LABELS EN LAS VARIABLES PARA LOS
CORRESPONDIENTES ETIQUETADOS
        ACM_PERRu$Tamaño<-factor(ACM_PERRu$Tamaño,  levels=c("1","2","3")
                                                 ,  labels=c("pequeño",
"mediano", "grande"))
        ACM_PERRu$Peso<-factor(ACM_PERRu$Peso,  levels=c("1","2","3")
                                              ,  labels=c("liviano",
"moderado", "pesado"))
        ACM_PERRu$Velocidad<-factor(ACM_PERRu$Velocidad,
levels=c("1","2","3")
                                                       ,
labels=c("lento", "normal", "rapido"))
        ACM_PERRu$Inteligencia<-factor(ACM_PERRu$Inteligencia,
levels=c("1","2","3")
                                                            ,
labels=c("baja", "moderada", "alta"))
        ACM_PERRu$Afeccion<-factor(ACM_PERRu$Afeccion,  levels=c("1","2")
                                                      ,
labels=c("afectuoso", "nada afectuoso"))
        ACM_PERRu$Agresividad<-factor(ACM_PERRu$Agresividad,
levels=c("1","2")
                                                           ,
labels=c("no agresivo", "agresivo"))
        ACM_PERRu$Funcion<-factor(ACM_PERRu$Funcion,
```

```
levels=c("1","2","3")
                                                    ,
labels=c("hogareño", "caza", "guardián"))
## ELABORAMOS LA PSEUDO TABLA DE CODIGOS CONDESADOS LA CUAL TIENE LA
INFORMACION CON LOS LABELS DE LAS VARIABLES
        PSEUDO_TCC<-ACM_PERRu
        View(PSEUDO_TCC)
        var_TDC<-names(PSEUDO_TCC)[c(3,4,5,6,7,8,9)]
## SELECCIONAMOS UNICAMENTE AQUELLASVARIABLES QUE SON CATEGÓRICAS QUE SE
VAN A GRAFICAR EN EL AFCM
        TCC<-select(PSEUDO_TCC,3:9)
## GRACIAS A LA FUNCIÓN dummy.cols del paquete fastdummies PODEMOS
ELABORAR LA MATRIZ DISYUNTIVA COMPLETA
        PSEUDO_TDC<-dummy_cols(TCC,var_TDC)
## Y FINALMENTE SELECCIONAMOS UNICAMENTE LAS VARIABLES DUMMIES QUE FUERON
OBTENIDAS
        TDC<-select(PSEUDO_TDC,8:26)
        col.sums<-apply(TDC, 2, sum)

# MATRIZ EN TERMINOS DE PROPORCIONES
      n<-sum(TDC)
      P<-TDC/n
## OBTENEMOS LOS MARGINALES FILAS Y COLUMNAS
      P<-as.matrix(P)
      rr<-margin.table(P,1)
      cc<-margin.table(P,2)
### OBTENCIÓN DE LA MATRIZ ESTANDARIZADA (S)
      S<-diag(rr^(-0.5)) %*% (P-rr %*%t (cc))%*%diag(cc^(-0.5))
## REALIZAMOS LA DESCOMPOSICIÓN DE LA MATRIZ (S)
      u<-svd(S)$u
      v<-svd(S)$v
      Da<-diag(svd(S)$d)
# OBTENEMOS LAS COORDENADAS PRINCIPALES PARA LOS PUNTOS INDIVIDUOS  LOS
PUNTOS VARIABLES
      FF<-diag(rr^(-0.5))%*% u %*%Da
      GG<-diag(cc^(-0.5))%*% v %*%Da
# CÁLCULO DE LA INERCIA
      cumsum(svd(S)$d)/sum(svd(S)$d)

##  [1] 0.1850366 0.3383906 0.4591491 0.5621030 0.6562400 0.7456616
0.8134529
##  [8] 0.8756395 0.9187329 0.9550959 0.9806346 1.0000000 1.0000000
1.0000000
## [15] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000

# PROCEDIMIENTO PARA ELABORAR EL GRÁFICO
  # OBTENEMOS LAS COORDENADAS DE LAS DIMENSIONES
      head(GG)

##            [,1]        [,2]        [,3]        [,4]        [,5]
[,6]
```

```
## [1,] -1.1766370 -0.83057452 -0.63118578 -0.31814689  0.17405018 -
0.08505787
## [2,] -0.8086295  1.06888021  0.97133875  0.74236170  0.47512440
0.03047557
## [3,]  0.8186404  0.03130804 -0.02922622 -0.09898535 -0.23959822
0.02953515
## [4,] -1.1682555 -0.77559599 -0.40003490 -0.20277276  0.18986074
0.06675840
## [5,]  0.2669075  0.86703831 -0.06512362 -0.08979627 -0.07418589
0.16823922
## [6,]  1.1218678 -1.18675368  0.82240199  0.57586596 -0.09605669 -
0.57788325
##             [,7]         [,8]        [,9]      [,10]        [,11]
[,12]
## [1,]  0.36109019 -0.054143135  0.19489708 -0.1418671 -0.117156027
0.20129249
## [2,] -0.83317239  0.311042606 -0.11983090 -0.2349034 -0.146600080 -
0.01967802
## [3,]  0.10921537 -0.078414072 -0.05100834  0.1445058  0.103539506 -
0.08737715
## [4,]  0.14860856  0.267212816 -0.03615692 -0.1583619 -0.007452018 -
0.21691656
## [5,] -0.04413446  0.002895897  0.06061985  0.2041835 -0.075689600
0.08008208
## [6,] -0.11419720 -0.435649018 -0.11188451 -0.3183347  0.223854109
0.12283669
##              [,13]         [,14]         [,15]         [,16]
[,17]
## [1,] -1.615504e-16 -3.850928e-17  4.242215e-17  7.131186e-17
5.762531e-17
## [2,] -1.615504e-16 -3.850928e-17  4.242215e-17  7.131186e-17
5.762531e-17
## [3,] -1.615504e-16 -3.850928e-17  4.242215e-17  7.131186e-17
5.762531e-17
## [4,] -7.437446e-17 -1.564519e-16 -4.923870e-17 -3.759259e-17 -
8.489143e-18
## [5,] -7.437446e-17 -1.564519e-16 -4.923870e-17 -3.759259e-17 -
8.489143e-18
## [6,] -7.437446e-17 -1.564519e-16 -4.923870e-17 -3.759259e-17 -
8.489143e-18
##              [,18]         [,19]
## [1,] -6.258895e-18  1.035159e-17
## [2,] -6.258895e-18  1.035159e-17
## [3,] -6.258895e-18  1.035159e-17
## [4,] -2.277529e-17 -2.068434e-17
## [5,] -2.277529e-17 -2.068434e-17
## [6,] -2.277529e-17 -2.068434e-17

  # GRAFICAMOS LA PRIMERA Y LA SEGUNDA DIMENSIÓN QUE SON LAS QUE TIENEN
MAYOR CONTRIBUCION A LA INERCIA
```

```
      plot(GG[,1],GG[,2])
      points(FF[,1],FF[,2],col="brown")
      kvar<-
c("Tamaño_pequeño","Tamaño_mediano","Tamaño_grande","Peso_liviano","Peso_
moderado"

,"Peso_pesado","Velocidad_lento","Velocidad_normal","Velocidad_rapido"

,"Inteligencia_baja","Inteligencia_moderada","Inteligencia_alta","Afeccio
n_afectuoso"
           ,"Afeccion_nada afectuoso", "Agresividad_no
agresivo","Agresividad_agresivo"
           ,"Funcion_hogareño","Funcion_caza","Funcion_guardián")
      text(GG[,1],GG[,2],labels =kvar,cex=0.7, col="violet")
  # GRAFICAMOS LOS PUNTOS INDIVIDUOS

        points(FF[,1],FF[,2],col="blue")
        ind<-PSEUDO_TCC$Raza
        text(FF[,1],FF[,2],labels =ind,cex=0.7,col = "green")
```



Y con los resultados obtenidos podemos evidenciar el comportamiento que poseen los individuos frente a ciertas características, podemos compararlos entre ellos, obtener cierta información del comportamiento y fidelidad de los datos en determinados tipos de encuestas cuestionarios y estudios. esta herramienta es fundamental para estudios de caracter de investigación cualitativa.

2° Paso AFCM aplicado a la tabla de BURT

El análisis de correspondencia múltiple implica la descomposición de la matriz de BURT en sus respectivos autovalores y autovectores mediante la descomposición espectral la cual queda representada como:

B=E V E´´

DONDE:

- E: Matriz ortogonal de autovectores
- V: Matriz diagonal de autovalores

Los autovalores (λ1> λ2> λ3> …> λm) son obtenidos a partir de la siguiente Expresión:

|B- λI|=0

Y a los autovalores λ les corresponde un autovector que son obtenidos de

|B- λI| a=0

Para la selección de los ejes factoriales, también es necesario definir la inercia total de la nube Nk con respecto al centro (0,0) a partir de la siguiente expresión:

Definida la inercia total de la nube de puntos Nk y sabiendo que es igual a la suma de las m autovalores podemos expresar el aporte de cada un eje a lo inercia total de la nube por medio de la tasa de inercia del factor

para el analisis de correspondencias multiples en una tabla de burt es mucho más sencillo si utilizamos la librería FactorMiner con la cual podemos realizar este tipo de analisis multivariante

```
# EN ESTE PASO PODEMOS ASIGNAR UNA TABLA DISYUNTIVA COMPLETA A LA MATIZ
DE BURT (afcm).
    afcm <- MCA(TCC, graph = FALSE)
    afcm
```

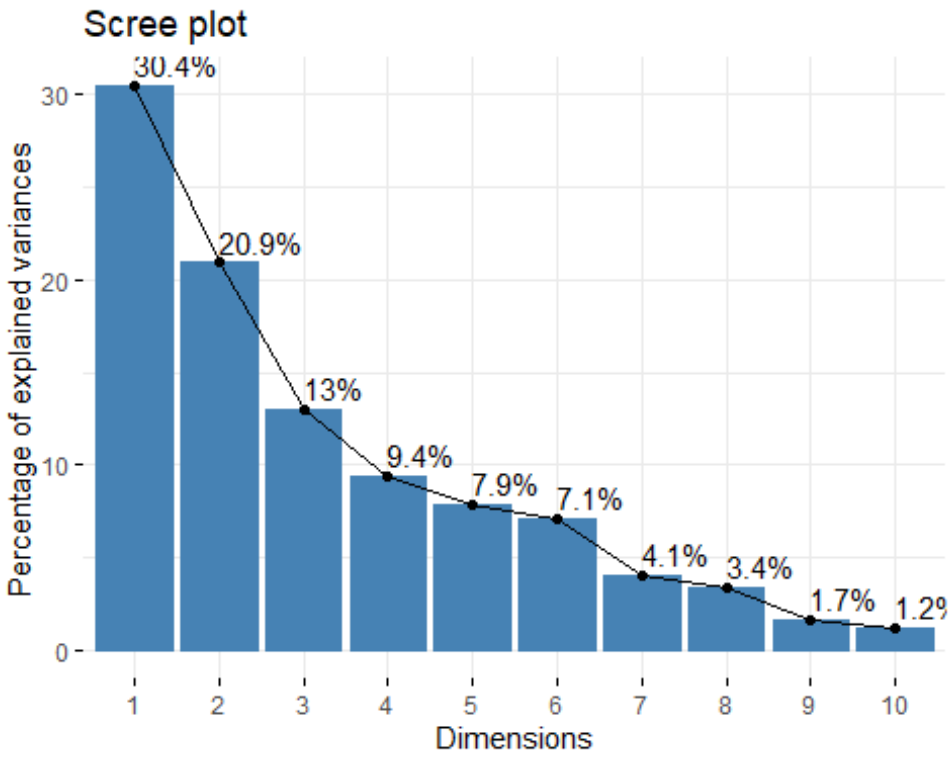
```
## **Results of the Multiple Correspondence Analysis (MCA)**
## The analysis was performed on 27 individuals, described by 7 variables
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$var"            "results for the variables"
## 3  "$var$coord"      "coord. of the categories"
## 4  "$var$cos2"       "cos2 for the categories"
## 5  "$var$contrib"    "contributions of the categories"
## 6  "$var$v.test"     "v-test for the categories"
## 7  "$ind"            "results for the individuals"
## 8  "$ind$coord"      "coord. for the individuals"
## 9  "$ind$cos2"       "cos2 for the individuals"
## 10 "$ind$contrib"    "contributions of the individuals"
```

```
## 11 "$call"           "intermediate results"
## 12 "$call$marge.col" "weights of columns"
## 13 "$call$marge.li"  "weights of rows"
```

```
## OBTENEMOS LOES EIGEN VALORES QUE NOS SERVIRÁN PARA ELEGIR LOS FACTORES
O LAS DIMENSIONES QUE UTILIZAREMOS PARA REPRESENTAR A LOS PUNTOS
VARIABLES Y A LOS PUNTOS INDIVIDUOS.
eig.val<-get_eigenvalue(afcm)
eig.val
```

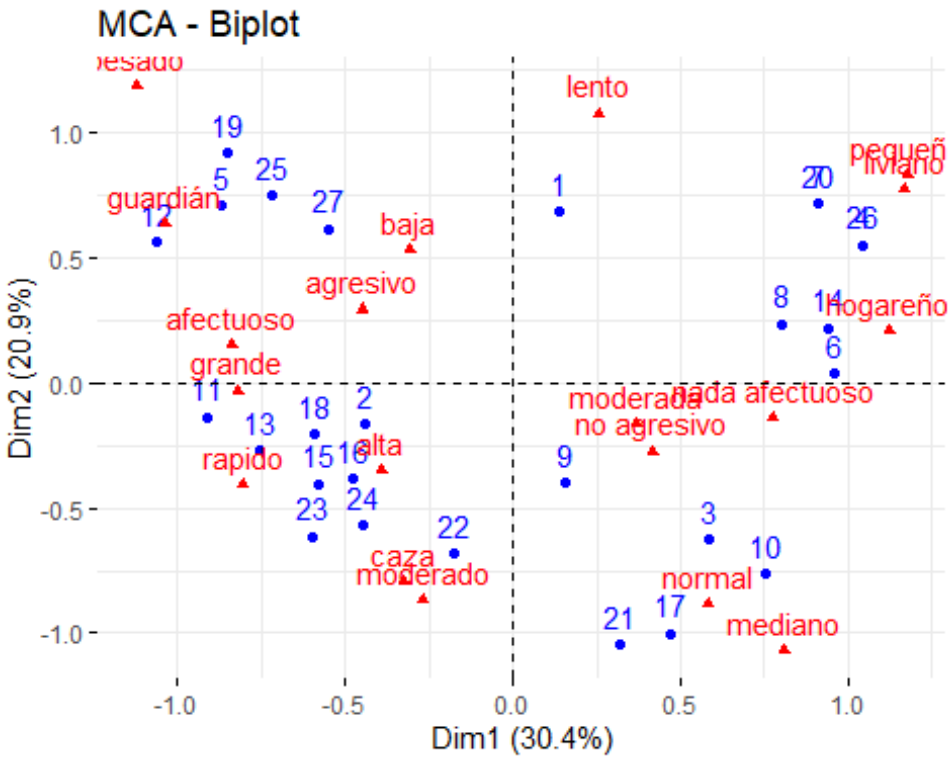
```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1  0.521890019       30.4435845                    30.44358
## Dim.2  0.358471438       20.9108339                    51.35442
## Dim.3  0.222279319       12.9662936                    64.32071
## Dim.4  0.161565866        9.4246755                    73.74539
## Dim.5  0.135077985        7.8795491                    81.62494
## Dim.6  0.121884489        7.1099285                    88.73487
## Dim.7  0.070050713        4.0862916                    92.82116
## Dim.8  0.058946345        3.4385368                    96.25969
## Dim.9  0.028306494        1.6512121                    97.91091
## Dim.10 0.020154984        1.1757074                    99.08661
## Dim.11 0.009941711        0.5799331                    99.66655
## Dim.12 0.005716351        0.3334538                   100.00000
```

```
# REALIZAMOS UN HISTOGRAMA DE LA CONCENTRACIÓN DE LA INERCIA
CORRESPONDIENTE A LOES EIGENVALORES PARA MEDIANTE LA REGLA DEL CODO
ELEGIR LAS DIMENSIONES
fviz_screeplot(afcm, addlabels = TRUE)
```



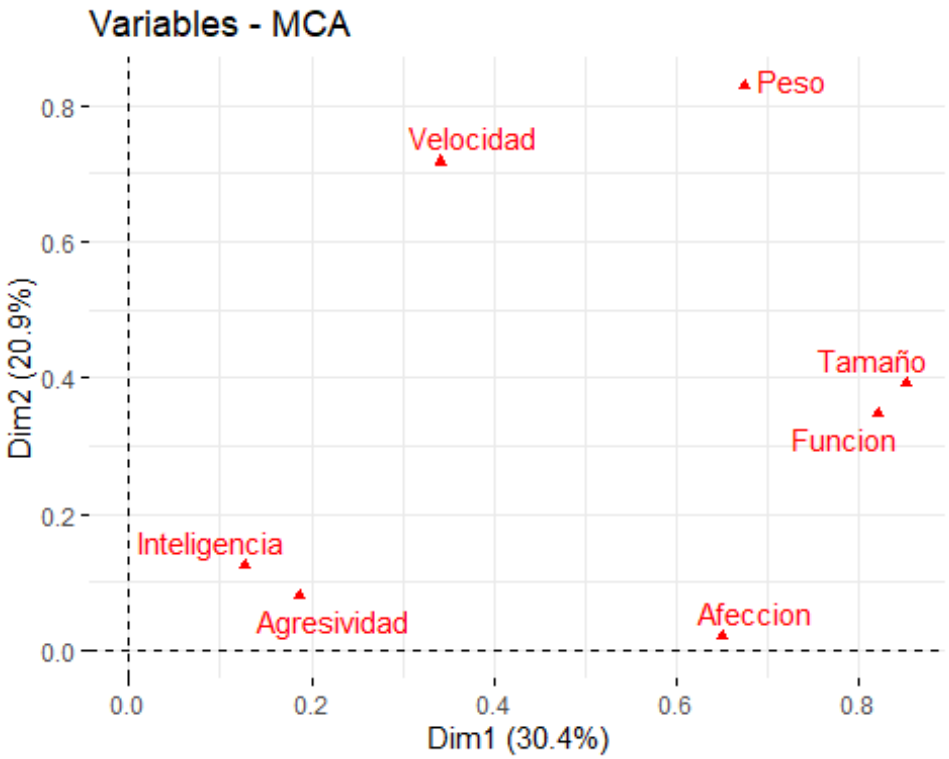
El histograma de la inercia explicada nos indica que dimensiones se pueden considerar a la hora de realizar las representaciones de los individuos y las variables en los planos factoriales, gracias a la regla del codo se puede evidenciar que la primera y la segunda dimensión concentran el 51,3% de la inercia explicada, pese a ello tambien se puede jugar con la tercera dimension ya que la variación porcentual de la inercia es de un 12 % entre la segunda y tercera dimension frente a un 20 % entre la primera ay segunda. pero pese a ello en este estudio realizaremos unicamente los graficos representados en las dos primeras dimensiones.

```
# REALIZAMOS UN PLOTING DE LOS PUNTOS INDIVIDUOS Y LOS PUNTOS VARIABLES
PARA VISUALIZAR DE QUE FORMA ESTAN  RELACIONADOS LOS INDIVIDUOS CON
REFERENCIA A LAS DIFERENTES VARIABLES
fviz_mca_biplot(afcm,ggtheme = theme_minimal())
```



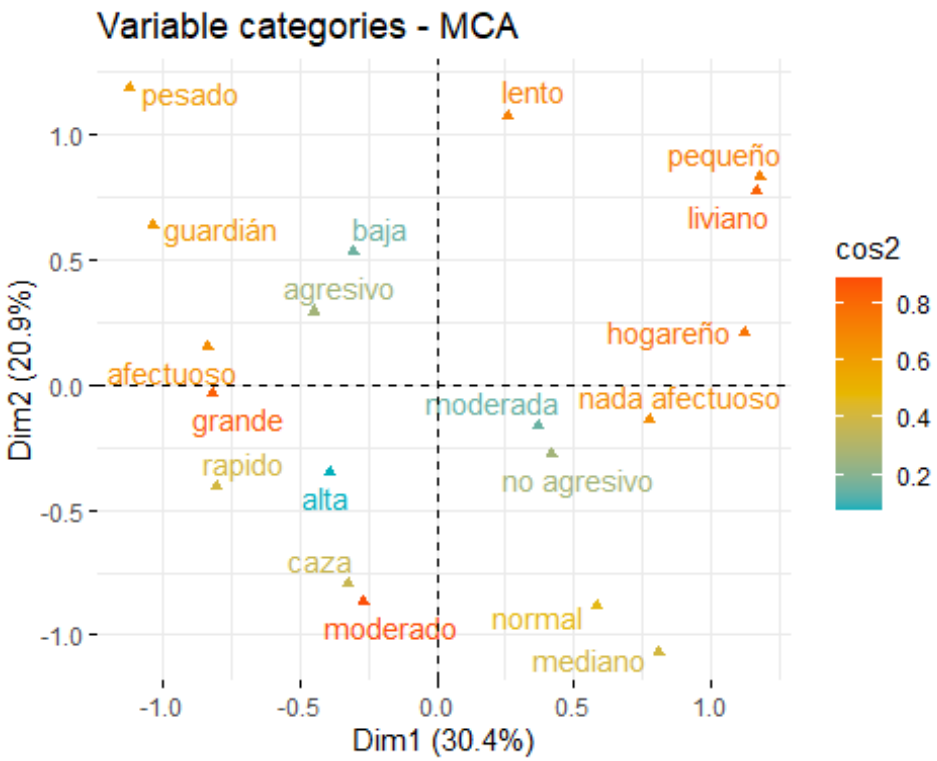
Este gráfico es el mas importante ya que en el se concentra la médula espinal del estudio ya que asocia los individuos es decir la raza de los canes en este caso con las variables de estudio, en el se puede interpretar que razas (observaciones) estan as intimanete ligadas a determinadas características

```
# REALIZAMOS UN  PLANO PARA VISUALIZAR LA RELACIÓN QUE GUARDAN LAS
VARIABLES EN TORNO A LOS PLANOS FACTORIALES ESCOGIDOS
fviz_mca_var(afcm, choice = "mca.cor",
             repel = TRUE, # Avoid text overlapping (slow)
             ggtheme = theme_minimal())
```



Una vez establecidas las dimensiones se puede comenzar a realizar la representación grafica de los puntos individuos y variables en el caso de los puntos variables o perfiles columna podemos ver que la inteligencia de los canes esta muy asociada con la agresividad, mientras que el tamaño esta asociado con la función que desempeña el animal, por otro lado podemos ver que la velocidad esta relativamente asociada con el peso del can esto poca asociación puede deberse a la concentración de masa muscular de la mascota.

```
# FINALMENTE VEMOS COMO ESTAN RELACIONALADAS LAS CATEGORIAS QUE SE
REGISTRAN EN LAS VARIABLES DE ESTUDIO ENTORNO A LAS DIMENSIONES
ESCOGIDAS.
fviz_mca_var(afcm, axes = c(1, 2), col.var = "cos2",
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, # Avoid text overlapping
             ggtheme = theme_minimal())
```



En el plot de las categorías podemos evidenciar mas claramente como estan relacionadas las categorias, esto nos muestra que categorías estan mas intimamente relacionadas con otras lo que resulta en una profundización de las variables de estudio realizadas