Técnicas de Análisis Multivariado - Trabajo 1

Justo Manrique Urbina - 20091107

01/12/2019

El presente trabajo tiene como objetivo ilustrar el uso de técnicas de análisis multivariado revisadas en la Maestría de Estadística PUCP. Para ello, se hizo uso de distintas bases de datos orientadas a la aplicación de dichas técnicas, las cuales se presentarán en las secciones correspondientes. Cada tipo de análisis tiene como base un problema de negocio o investigación, así como una base de datos la cual es útil para brindar solución al problema. Posteriormente, se analizan los resultados de dichas técnicas y se concluye sobre la misma.

Las técnicas multivariadas utilizadas en el presente informe son:

- Análisis de Componentes Principales.
- Análisis Discriminante.
- Análisis Factorial.

Ver a continuación el uso de cada técnica.

Análisis de Componentes Principales

Introducción

<>.

Los objetivos del presente estudio son:

- Conocer si existen grupos de autos con perfiles de rendimiento similares e identificar si existen autos de distinta clase.
- Identificar aquellos autos que lideran cada clase.

Datos

Los datos provienen de la revista *Motor Trend*, edición 1974. Dicha base de datos tiene las siguientes variables:

- mpg: Millas por galón.
- cyl: Número de cilindros.
- disp: Desplazamiento (en pies cúbicos).
- hp: Caballos de fuerza.
- drat: Relación del eje trasero.
- wt: Peso (definido en miles de libras).
- qsec: Tiempo para llegar a recorrer un cuarto de libra.
- am: Transmisión (automático o manual).
- vs: Tipo de motor (en forma de V o recto).
- gear: Cantidad de marchas hacia adelante.
- carb: Número de carburadores.

Con el propósito de ejecutar el análisis de los datos, realizamos la carga de librerías e importamos los datos. Posteriormente, realizamos un preprocesamiento de los datos para convertir los valores binarios en variables cualitativas.

```
## Carga de datos y librerías ##
library(FactoMineR)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)

data("mtcars")

## Preprocesamiento de datos ##

mtcars$vs <- factor(mtcars$vs)
mtcars$vs <- recode_factor(mtcars$vs, `0`="V-shaped",`1`="Straight")
mtcars$am <- factor(mtcars$am)
mtcars$am <- recode_factor(mtcars$am, `0`="Automatic",`1`="Manual")
```

Posteriormente, se observa los primeros valores de la base de datos para entender su estructura.

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec       vs        am gear
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46 V-shaped    Manual    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02 V-shaped    Manual    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61 Straight    Manual    4
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44 Straight Automatic    3
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02 V-shaped Automatic    3
## Valiant           18.1   6  225 105 2.76 3.460 20.22 Straight Automatic    3
##                   carb
## Mazda RX4            4
## Mazda RX4 Wag        4
## Datsun 710           1
## Hornet 4 Drive       1
## Hornet Sportabout    2
## Valiant              1
```

Se observa que cada línea corresponde a un modelo de auto específico. Asimismo, se observa que todas las variables son cuantitavias, excepto por las variables 'am' y 'vs'.

Resultados

Posteriormente, se utilizó la matriz de correlación para entender las relaciones lineales que tiene cada variable respecto a otra. Se utilizaron solo las variables cuantitativas para este análisis:

```
corcars <- cor(mtcars[c(1:7,10:11)])
corcars
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
```

```
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec       gear       carb
## mpg   0.41868403  0.4802848 -0.5509251
## cyl  -0.59124207 -0.4926866  0.5269883
## disp -0.43369788 -0.5555692  0.3949769
## hp   -0.70822339 -0.1257043  0.7498125
## drat  0.09120476  0.6996101 -0.0907898
## wt   -0.17471588 -0.5832870  0.4276059
## qsec  1.00000000 -0.2126822 -0.6562492
## gear -0.21268223  1.0000000  0.2740728
## carb -0.65624923  0.2740728  1.0000000
```

Se observa lo siguiente:

- Se observan correlaciones negativas fuertes en los siguientes pares de variables: (mpg) Millas por galón y (cyl) Números de cilindros; (mpg) Millas por galón y (hp) Caballos de fuerza; (mpg) Millas por galón y (disp) Desplazamiento (en pies cúbicos); (mpg) Millas por galón y (wt) Peso (definido en miles de libras).
- Se observan correlaciones positivas fuertes en los siguientes pares de variables: (cyl) Número de cilintros y (disp) Desplazamiento (en pies cúbicos); (cyl) Número de cilindros y (hp) Caballos de fuerza; (disp) Desplazamiento (en pies cúbicos) y (wt) Peso (definido en miles de libras).

En base a este análisis, podemos intuir que aquellas variables que tengan una correlación positiva fuerte formarán parte de un componente, mientras que aquellos con correlación negativa fuerte formarán parte de distintos componentes.

Posteriormente, utilizamos el test de esfericidad de Bartlett:

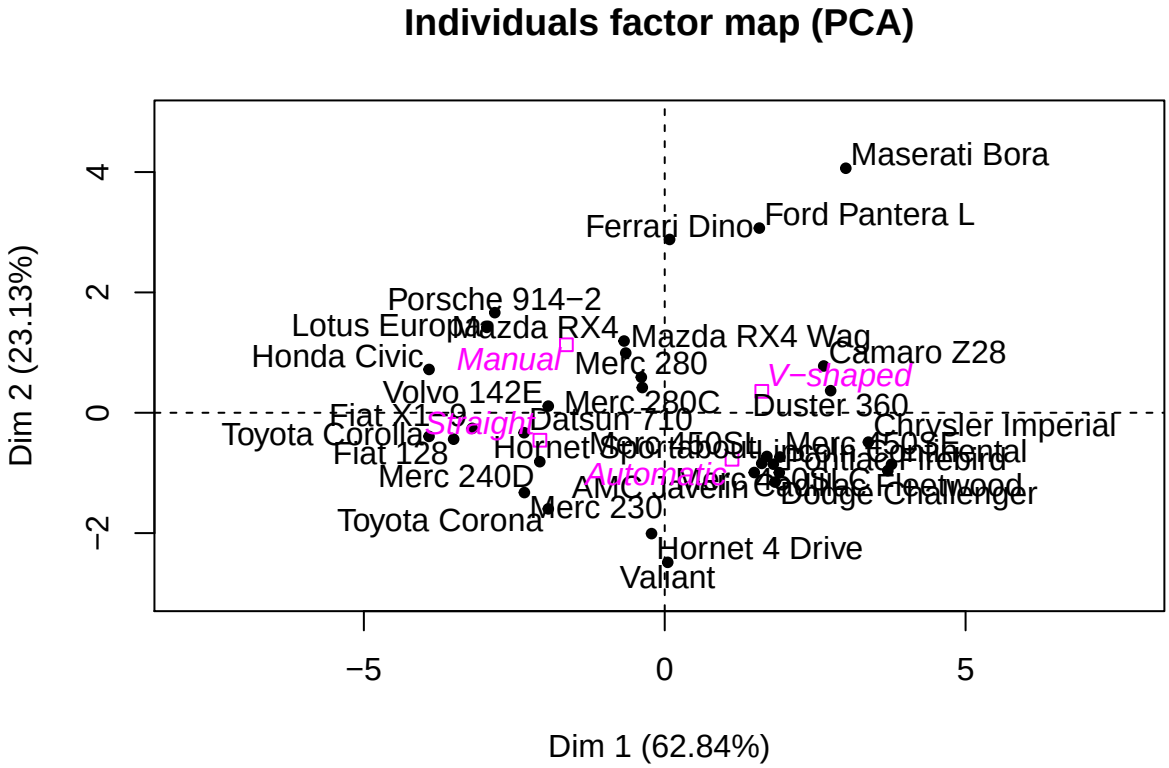
```
cortest.bartlett(corcars,n = 32)
```

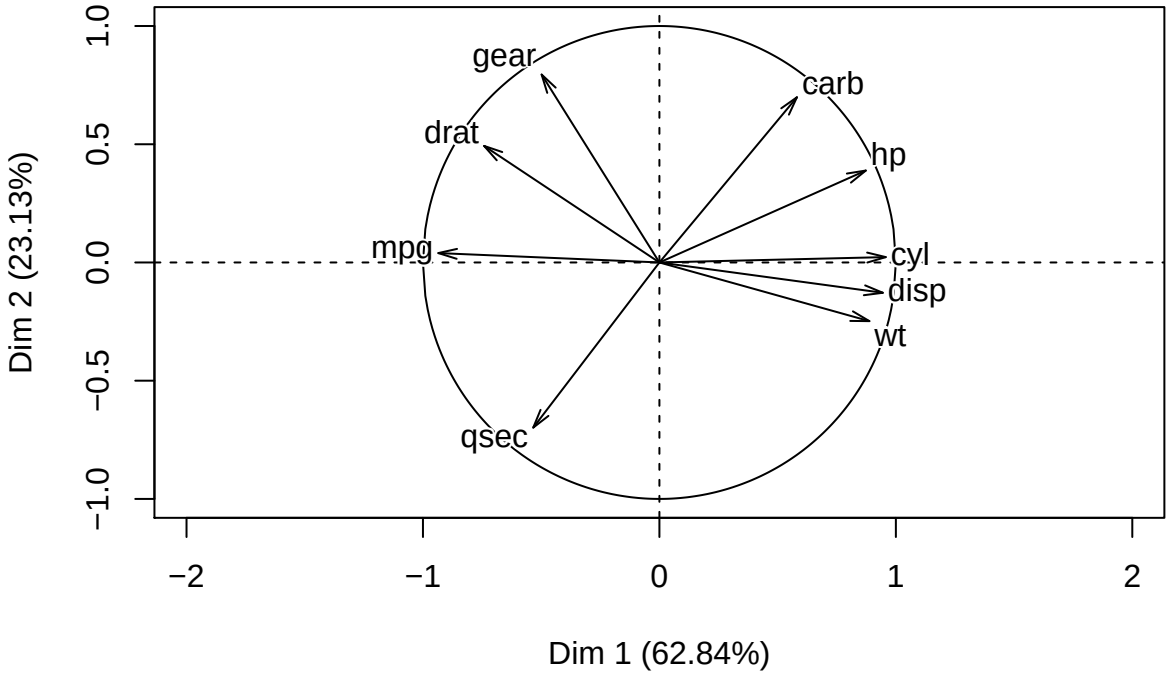
```
## $chisq
## [1] 332.328
##
## $p.value
## [1] 1.203652e-49
##
## $df
## [1] 36
```

De acuerdo a la prueba de esfericidad de Bartlett, observamos que el p-valor es muy pequeño por lo que la hipótesis nula se rechaza. En base a ello podemos concluir que la técnica de componentes principales será de aplicabilidad a la base de datos presentada.

Posteriormente, realizaremos el análisis de componentes principales mediante el siguiente código. Ver a continuación el código y sus salidas:

```
mt_pca <- PCA(mtcars,quali.sup = c(8,9),graph = TRUE,scale.unit = TRUE)
```





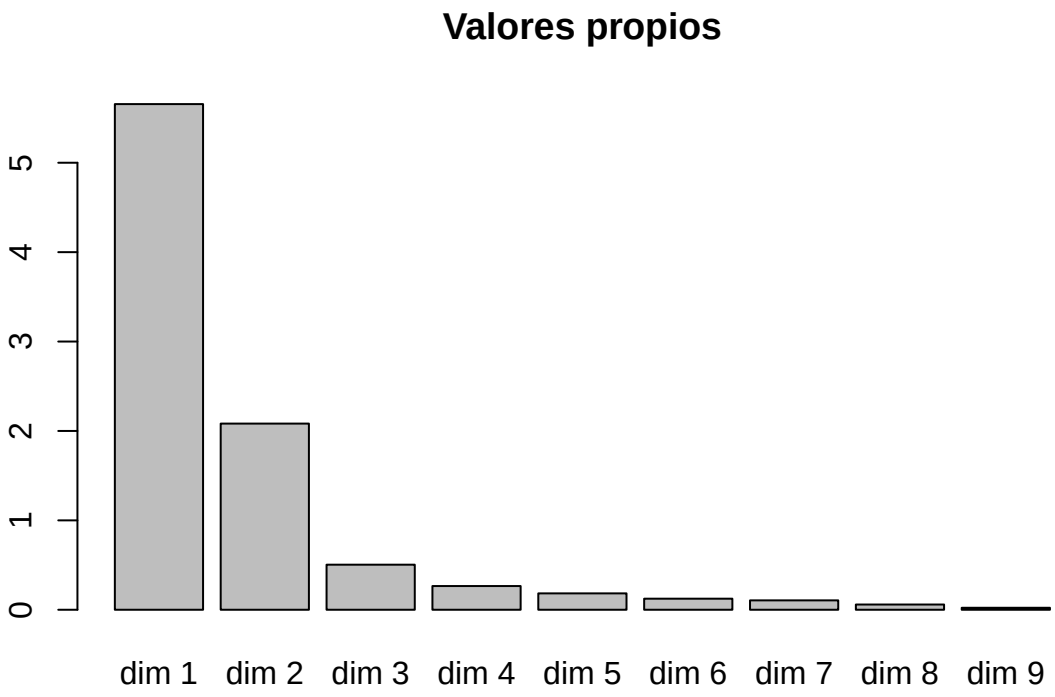
En base a los gráficos mostrados, se observa lo siguiente:

- El primer eje (Dim 1) expresa el 62.84% de la variabilidad de los datos y el segundo eje (Dim 2) el 23.13% de los mismos. En total, los dos primeros ejes expresan juntos el 85.97% de la variabilidad de los datos.
  - Variables factor map (PCA)

    * Se observa que, si un auto se encuentra en el primer cuadrante (esquina superior izquierda), este tendrá mayor millaje por galón, relación del eje trasero y mayor cantidad de marchas hacia adelante. Esto en contraposición del segundo y cuarto cuadrante (parte derecha del gráfico), el cual está asociado a mayor número de cilindros, desplazamiento, peso y caballos de fuerza.
- Individuals factor map (PCA)
    * Se observan que existen autos con perfiles de rendimiento similar. Por ejemplo, en el primer cuadrante se encuentran los autos Porsche 914-2 y Lotus Eurpa, mientras que en el cuarto cuadrante se encuentra el Cadillac Fleetwood y Chrysler Imperial.
    * El perfil de rendimiento similar está asociado a la ubicación del auto en el plano de las dos dimensiones. Para ello, utilizamos el variables factor map para entender qué características tienen en común determinados autos.

Posteriormente a este análisis, realizamos uno más detallado para identificar si existe otro componente (u otros componentes) que puedan incluirse para efectos del estudio. Para ello, generamos un gráfico que nos permita revisar la importancia de todos los componentes.

```
barplot(mt_pca$eig[,1], main="Valores propios", names.arg=paste("dim",1:nrow(mt_pca$eig)))
```



Se observa que existen en total 9 componentes, de los cuales los dos primeros componentes (los cuales explican el 85.97% de los datos) son los que explican en mayor proporción la variabilidad. Los componentes del 3 al 9 tienen un bajo autovalor, por lo que no serán utilizados para el análisis.

Asimismo, identificaremos si existen individuos (en este caso, autos) o variables que contribuyen mucho a los componentes elegidos. Ver código a continuación:

```
round(mt_pca$ind$contrib[,1:2],2)
```

```
##                   Dim.1 Dim.2
## Mazda RX4          0.25  2.13
## Mazda RX4 Wag      0.23  1.48
## Datsun 710         3.02  0.17
## Hornet 4 Drive     0.03  6.05
## Hornet Sportabout  1.44  1.06
```

```
## Valiant              0.00  9.27
## Duster 360           4.20  0.20
## Merc 240D            2.38  0.99
## Merc 230             3.00  2.64
## Merc 280             0.08  0.52
## Merc 280C            0.08  0.26
## Merc 450SE           2.03  0.81
## Merc 450SL           1.59  0.79
## Merc 450SLC          1.80  1.10
## Cadillac Fleetwood   7.60  1.39
## Lincoln Continental  7.85  1.10
## Chrysler Imperial    6.33  0.36
## Fiat 128             6.80  0.29
## Honda Civic          8.47  0.78
## Toyota Corolla       8.48  0.23
## Toyota Corona        2.07  3.83
## Dodge Challenger     1.86  1.99
## AMC Javelin          1.22  1.48
## Camaro Z28           3.86  0.91
## Pontiac Firebird     2.00  1.49
## Fiat X1-9            5.65  0.10
## Porsche 914-2        4.41  4.15
## Lotus Europa         4.83  3.02
## Ford Pantera L       1.37 14.14
## Ferrari Dino         0.00 12.45
## Maserati Bora        5.01 24.78
## Volvo 142E           2.07  0.02
```

```
round(mt_pca$var$contrib[,1:2],2)
```

```
##      Dim.1 Dim.2
## mpg  15.46  0.08
## cyl  16.20  0.02
## disp 15.79  0.79
## hp   13.47  7.26
## drat  9.72 11.67
## wt   13.95  2.96
## qsec  5.03 23.43
## gear  4.39 30.34
## carb  5.98 23.46
```

Se observa lo siguiente:

- Respecto a la importancia de variables en la construcción de componentes:
  - Se observa que, para el segundo componente, la cantidad de carburadores, de marchas hacia adelante y el tiempo para recorrer un cuarto de milla son las variables más importantes.
  - Se observa que, para el primer componente, la cantidad de millas por galón, el número de cilindros, los caballos de fuerza, el peso y el desplazamiento (en pies cúbicos) son las variables más importantes.
  - La relación del eje trasero contribuye a ambos componentes de forma similar.
- Respecto a la importancia de los individuos en la construcción de componentes:
  - Los autos de marca Ford Pantera L, Ferrari Dino y Maserati Bora son los que contribuyen en gran manera a la construcción del segundo componente. Se observa que, en relación a los demás individuos, el aporte de estos individuos es mucho mayor.
  - El aporte de los autos al componente 1 es más equilibrado que el componente 2. Se observa que los autos con mayor aporte son el Toyota Corolla, Honda Civic y Lincoln Continental, sin embargo

en relación a los demás individuos el aporte es regular.

Finalmente, realizamos la descripción de los ejes a través de la correlación de las variables de cada componente:

```
dimdesc(mt_pca,axes = c(1,2))
```

```
## $Dim.1
## $Dim.1$quanti
##      correlation      p.value
## cyl    0.9573620 9.987998e-18
## disp   0.9449932 4.195104e-16
## wt     0.8882114 1.186867e-11
## hp     0.8730011 7.238862e-11
## carb   0.5816671 4.798744e-04
## gear  -0.4981777 3.711660e-03
## qsec  -0.5335561 1.662320e-03
## drat  -0.7415688 1.197792e-06
## mpg   -0.9349924 4.804756e-15
##
## $Dim.1$quali
##           R2      p.value
## vs 0.5916018 2.698147e-07
## am 0.3231649 6.875337e-04
##
## $Dim.1$category
##                Estimate      p.value
## vs=V-shaped   1.843685 2.698147e-07
## am=Automatic  1.376372 6.875337e-04
## am=Manual    -1.376372 6.875337e-04
## vs=Straight  -1.843685 2.698147e-07
##
##
## $Dim.2
## $Dim.2$quanti
##      correlation      p.value
## gear   0.7947510 5.574498e-08
## carb   0.6988388 8.637452e-06
## drat   0.4929872 4.146805e-03
## hp     0.3887501 2.788365e-02
## qsec  -0.6984510 8.780023e-06
##
## $Dim.2$quali
##           R2      p.value
## am 0.4193267 6.169841e-05
##
## $Dim.2$category
##                 Estimate      p.value
## am=Manual      0.9512586 6.169841e-05
## am=Automatic -0.9512586 6.169841e-05
```

En base a lo identificado, la primera dimensión está asociada a las características relacionadas a la potencia del motor (número de cilindros, caballos de fuerza, carburadores y cilindrada) mientras que la segunda dimensión está asociada a la maniobrabilidad del auto (cantidad de marchas, tipo de transmisión, entre otros).

Discusión y conclusiones

En base al análisis presentado, y en relación a los objetos de estudio, se concluye lo siguiente:

- Existen dos perfiles de autos: aquellos orientados a ser autos potentes (tienen mayor índice en la dimensión 1) y aquellos que son maniobrables (tienen mayor índice en la dimensión 2)
- Utilizando las variables cualitativas, se observa que aquellos autos orientados a ser autos potentes tienen el motor en forma de V, mientras que aquellos que no tienen el motor de forma recta.
- De igual forma, se observa que aquellos autos orientados a ser autos maniobrables tendrían mayor propensión a tener transmisión manual.
- Existen autos que serían tanto maniobrables como potentes: los casos específicos serían el Maserati Bora y Ford Pantera.

Análisis Discriminante

Introducción

Datos

Resultados

Discusión y conclusiones

Análisis Factorial

Introducción

Datos

Resultados

Discusión y conclusiones