

Análisis Multivariante de Datos

Solución Indicativa de la Primera Prueba de
Evaluación Continua

**Descripción de datos multivariantes,
análisis gráfico y detección de valores
atípicos**

Mayo 2016

Resumen

En este documento planteamos una solución indicativa de la primera prueba de evaluación continua de la asignatura de Análisis Multivariante de Datos.

El objetivo del documento es dar un conjunto de guías e indicaciones sobre como se debe abordar el problema, pero no aporta una solución detallada, ya que no existe una única solución al problema y cada estudiante puede haber escogido opciones diversas e igualmente válidas.

1. Introducción

Durante las primeras 6 semanas de curso nos hemos iniciado en el análisis multivariante. Tras una introducción al tema y un repaso de conceptos de álgebra matricial, hemos estudiado diversas formas de describir los datos multivariantes y representarlos gráficamente. También hemos comenzado el estudio de los valores atípicos. En esta PEC aplicamos los conocimientos adquiridos durante estas 6 primeras semanas de curso.

2. Objetivos

A partir del conjunto de datos CENSUS se pide:

1. Describir el conjunto de datos
 - a) Usar medidas de centralización y analizarlas
 - b) Usar medidas de variabilidad y analizarlas
 - c) Estudiar las dependencias lineales entre las variables
2. Representar gráficamente los datos
3. Estudiar la existencia de valores atípicos
 - a) Identificar y aplicar diversas (al menos dos) técnicas
 - b) Comparar los resultados de las técnicas estudiadas
 - c) Analizar de forma crítica los resultados

3. Descripción del conjunto de datos

El conjunto de datos multivariante estudiado en esta prueba se conoce con el nombre de CENSUS. Este conjunto de datos ha sido usado en el proyecto CASC (<http://neon.vb.cbs.nl/casc/index.htm>) como conjunto de test para estudiar y proponer técnicas de protección de datos en el campo del control de la revelación estadística (en Inglés: Statistical Disclosure Control).

El conjunto CENSUS contiene 1080 registros (filas) y 13 atributos (columnas) con datos extraídos de la base de datos del U.S. Bureau of the Census. La información de CENSUS hace referencia a salarios, impuestos, beneficios y sueldos pagados por empresas y particulares Americanos durante 1995. Para una descripción más completa puede consultarse (<http://neon.vb.cbs.nl/casc/CASCrefmicrodata.pdf>). El conjunto de datos CENSUS

se encuentra públicamente disponible en: <http://neon.vb.cbs.nl/casc/CASCrefmicrodata.zip>.

Para cargar los datos en R podemos usar el siguiente comando, que nos permite escoger el conjunto de datos mediante una ventana emergente (y nos ahorra definir la ruta y el nombre del fichero en el propio comando).

```
> census<-read.csv(file.choose(),header=TRUE,sep=";")
> census
```

Obtenemos el resultado siguiente:

	AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERINVAL
1	270914	45554	4173	4621	45527	1428	30809	27	27	45500	3480	45500	45500
2	250802	57610	2639	6045	42008	1902	39234	1008	808	41000	3136	41000	41000
3	299391	56606	3315	4765	56485	1903	31767	485	485	56000	4284	56000	56000
...													
...													

Una vez cargados los datos podemos comenzar a estudiarlos. Describir datos multivariantes supone estudiar cada variable aisladamente y además las relaciones entre ellas. Para ello primero estudiamos diversas medidas de centralización, diversas medidas de variabilidad, y finalizamos con el estudio de las dependencias lineales.

3.1. Medidas de centralización

La medida de centralización más utilizada para describir datos multivariantes es el vector de medias, que es un vector de dimensión p cuyos componentes son las medias de cada una de las p variables.

La media para cada una de las variables la calculamos usando la siguiente expresión:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$$

En R podemos obtener el vector de medias con el siguiente comando:

```
> VectorMedias <- colMeans(census)
```

que en nuestro caso da como resultado:

AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERINVAL
196039.812	56222.758	3173.135	7544.656	45230.839	2597.184	39712.953	5162.230	1421.411	40068.609	2962.645	39523.375	38444.556

De forma similar podemos calcular el vector de medianas:

```
> medianas<-colMedians(census)
```

que en nuestro caso da como resultado:

```
AFNLWGT AGI      EMCNTRB FEDTAX PTOTVAL STATETAX TAXINC  POTHVAL INTVAL PEARNVAL FICA    WSALVAL ERNVAL
180349.0 58412.5 3215.5   7068.0 43278.0 2322.0   41155.0 1586.5   353.0 39000.0 3002.0 38000.0 36000.0
```

Y también podemos calcular el vector de MEDAS (MEDiana de las Desviaciones absolutas). Para ello usamos la función *mad* (Median of Absolut Deviations) de R, y la aplicamos mediante un bucle a cada una de las variables:

```
> for (i in 1:13){
+ medas[i]<-mad(census[,i])}
> medas
```

cuyo resultado es:

```
AFNLWGT AGI      EMCNTRB FEDTAX PTOTVAL STATETAX TAXINC  POTHVAL INTVAL PEARNVAL FICA    WSALVAL ERNVAL
83267.26 29029.31 1564.88 5669.46 23348.73 1829.53 24903.97 2095.66 466.28 22239.00 1786.53 22239.0 22239.0
```

También conviene calcular los coeficientes de asimetría, que miden la simetría de los datos respecto a su centro, y que se calcula como:

$$A_j = \frac{1}{n} \frac{\sum (x_{ij} - \bar{x}_j)^3}{s_j^3}$$

En R podemos calcular este coeficiente con el comando:

```
> asimetria<-skewness(census)
```

que en nuestro caso da como resultado:

```
AFNLWGT AGI      EMCNTRB FEDTAX PTOTVAL STATETAX TAXINC  POTHVAL INTVAL PEARNVAL FICA    WSALVAL ERNVAL
0.93752 -0.15719 0.01321 0.37576 0.41644 1.00632 -0.08808 4.23016 6.87460 0.32238 -0.00961 0.32378 0.34485
```

3.1.1. Comentarios generales sobre las medidas de centralización obtenidas

Un breve análisis de estas sencillas medidas de centralización nos aporta información relevante de cara al posterior análisis de valores atípicos.

Por ejemplo, podemos calcular la diferencia entre el vector de medias y el vector de medianas:

```
> diferencia <- medias - medianas
```

```
AFNLWGT AGI      EMCNTRB FEDTAX PTOTVAL STATETAX TAXINC  POTHVAL INTVAL PEARNVAL FICA    WSALVAL ERNVAL
15690.81 -2189.74 -42.36  476.65 1952.83 275.1842 -1442.04 3575.72 1068.41 1068.60  -39.35 1523.37 2444.55
```

Sin embargo puede resultar engañoso estudiar estos valores directamente, ya que se encuentran expresados en sus unidades originales y difícilmente pueden compararse. Por ello puede resultar útil normalizar el resultado de la diferencia por el rango de cada variable.

Para calcular el rango de cada variable, obtenemos su máximo, su mínimo y los restamos:

```
> maximos<- apply(census,2,max)
> minimos<- apply(census,2,min)
> rangos <- maximos-minimos
```

AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
675472	93355	7075	21259	113151	11478	83446	105940	49424	97524	7926	97524	97524

Finalmente normalizamos:

```
> diferenciaNormalizada <- diferencia/rangos
```

AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
0.02322	-0.02345	-0.00598	0.02242	0.01725	0.02397	-0.01728	0.03375	0.02161	0.01095	-0.00496	0.01562	0.02506

A pesar de no ser un indicador especialmente robusto, observamos que la variable POTHVAL es la que presenta una diferencia relativa más alta entre la media y la mediana (0,03375) y ello podría hacernos sospechar de la presencia de valores atípicos.

Otra alternativa sencilla consiste en normalizar usando la media de la variable:

```
> diferenciaNormalizada <- diferencia/medias
```

AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
0.0800	-0.0389	-0.0133	0.0631	0.0431	0.1059	-0.0363	0.6926	0.7516	0.0266	-0.0132	0.0385	0.0635

Observamos que con este procedimiento destacan sobre el resto las variables POTHVAL y INTVAL con valores 0,6926 y 0,7516. Ello podría hacernos sospechar de la presencia de datos atípicos en estas variables.

Además analizando los resultados de la asimetría vemos que las variables POTHVAL y INTVAL con coeficientes 4,23016 y 6,87460 son las más asimétricas.

Así pues, a pesar de que las medidas de centralidad no son las más adecuadas para sacar conclusiones sobre datos atípicos, sí podemos obtener información que nos ayude a guiar nuestra búsqueda.

3.2. Medidas de variabilidad

La medida más habitual usada para evaluar la variabilidad de una variable es la desviación típica, que podemos expresar como:

$$s_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n}}$$

En R podemos calcular la desviación típica de las variables de nuestro conjunto de datos con el comando:

```
> desviacion<-colSds(census)
```

Y obtenemos el resultado:

```
AFNLWGT      AGI      EMCONTRB    FEDTAX    PTOTVAL   STATETAX  TAXINC    POTHVAL  INTVAL  PEARNVAL  FICA      WSALVAL  ERNVAL  
101251.417  24674.843  1401.832  4905.200  21323.470  1826.436  21224.161  9449.640  3750.892  20816.008  1427.234  20601.275  20677.574
```

También podemos calcular la varianza s_j^2 :

```
> varianza <-colVars(census)
```

y obtenemos:

```
AFNLWGT      AGI      EMCONTRB    FEDTAX    PTOTVAL   STATETAX  TAXINC    POTHVAL  INTVAL  PEARNVAL  FICA      WSALVAL  ERNVAL  
10251849465 608847901 1965133 24060985 454690360 3335868 450464989 89295691 14069194 433306196 2036996 424412533 427562061
```

Naturalmente podemos calcular otras medidas relacionadas con la variabilidad como por ejemplo:

- El coeficiente de variación
- La homogeneidad
- El coeficiente de homogeneidad
- La kurtosis

No entramos en detalle en las medidas de homogeneidad puesto que las veremos en el apartado dedicado a la detección de valores atípicos. Valores altos de kurtosis nos indicaran la presencia de valores atípicos.

También podríamos calcular las distancias entre todos los puntos (dimensión a dimensión) y representarlas. Para hacerlo podemos usar la función *dist* de R. Como ejemplo, vemos como calcularíamos las distancias Euclídeas de las variables AFNLWGT, POTHVAL y INTVAL.

```

> DistanciaAFNLWGT<-dist(census[,1],method="euclidean")
> plot(DistanciaAFNLWGT,main=paste("Distancia Euclídea entre puntos
  de AFNLWGT"), xlab=paste("Índice"))
> DistanciaPOTHVAL<-dist(census[,8],method="euclidean")
> plot(DistanciaPOTHVAL,main=paste("Distancia Euclídea entre puntos
  de POTHVAL"), xlab=paste("Índice"))
> DistanciaINTVAL<-dist(census[,9],method="euclidean")
> plot(DistanciaINTVAL,main=paste("Distancia Euclídea entre puntos
  de INTVAL"), xlab=paste("Índice"))

```

El resultado gráfico se muestra en la Figura 1

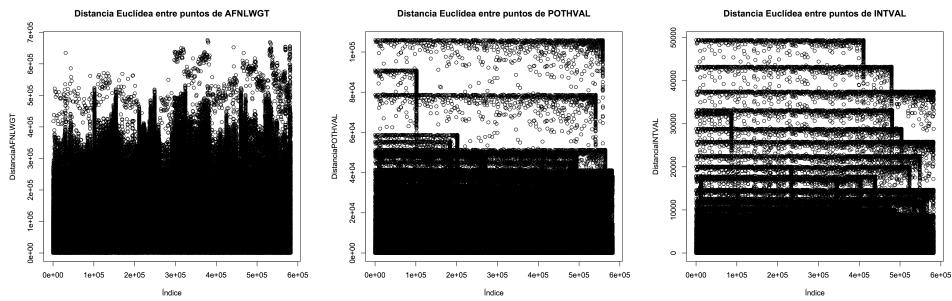


Figura 1: Distancias Euclídeas de las variables AFNLWGT, POTHVAL y INTVAL.

Otra forma alternativa de medir la variabilidad de un conjunto de datos multivariante es calcular la distancia entre sus puntos. En nuestro caso dado que contamos con un conjunto de datos de 13 variables, nuestros puntos estarán representados en R^{13} y podremos usar (por ejemplo) la distancia Euclídea para evaluar su variabilidad.

```

> DistanciaConjunta<-dist(census, method="euclidean")
> heatmap(as.matrix(DistanciaConjunta),Colv=NA,Rowv=NA,scale="none")

```

Mostramos el resultado en la Figura 2

3.3. Análisis de dependencias lineales entre variables

Para analizar las dependencias entre las variables podemos estudiar sus correlaciones, para ello usamos el comando de R *cor*, que por defecto nos da la correlación de *Pearson*. Alternativamente podríamos usar la correlación de *Spearman* o *Kendall*.

```
> correlaciones<-cor(census)
```

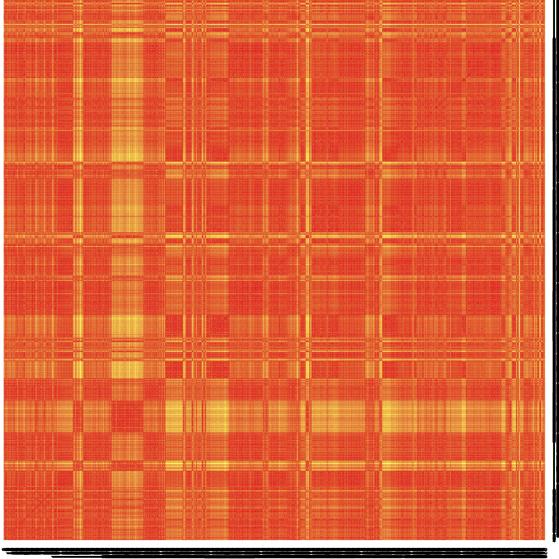


Figura 2: Mapa de calor con las distancias Euclídeas totales (1080×1080).

	AFNLWGT	AGI	EMCONTRB	FEDTAX	PTOTVAL	STATETAX	TAXINC	POTHVAL	INTVAL	PEARNVAL	FICA	WSALVAL	ERNVAL
AFNLWGT	1.000	0.010	0.049	-0.004	0.018	-0.100	0.004	-0.027	-0.024	0.031	0.032	0.039	0.036
AGI	0.010	1.000	0.491	0.945	0.774	0.779	0.980	0.138	0.200	0.730	0.709	0.716	0.701
EMCONTRB	0.049	0.491	1.000	0.382	0.495	0.289	0.436	-0.115	-0.058	0.560	0.547	0.559	0.550
FEDTAX	-0.004	0.945	0.382	1.000	0.798	0.790	0.979	0.200	0.282	0.726	0.688	0.708	0.692
PTOTVAL	0.018	0.774	0.495	0.798	1.000	0.633	0.777	0.275	0.272	0.900	0.849	0.879	0.859
STATETAX	-0.100	0.779	0.289	0.790	0.633	1.000	0.788	0.130	0.220	0.589	0.570	0.572	0.563
TAXINC	0.004	0.980	0.436	0.979	0.777	0.788	1.000	0.168	0.234	0.720	0.696	0.706	0.689
POTHVAL	-0.027	0.138	-0.115	0.200	0.275	0.130	0.168	1.000	0.454	-0.173	-0.189	-0.173	-0.171
INTVAL	-0.024	0.200	-0.058	0.282	0.272	0.220	0.234	0.454	1.000	0.072	0.061	0.070	0.075
PEARNVAL	0.031	0.730	0.560	0.726	0.900	0.589	0.720	-0.173	0.072	1.000	0.955	0.979	0.958
FICA	0.032	0.709	0.547	0.688	0.849	0.570	0.696	-0.189	0.061	0.955	1.000	0.910	0.894
WSALVAL	0.038	0.716	0.559	0.708	0.879	0.572	0.706	-0.173	0.070	0.979	0.910	1.000	0.973
ERNVAL	0.036	0.701	0.550	0.692	0.859	0.563	0.689	-0.171	0.075	0.958	0.894	0.973	1.000

Y podemos representar el resultado mediante un mapa de calor con el comando *heatmap*, cuyo resultado se muestra en la Figura 3.

```
> heatmap(correlaciones, Colv=NA, Rowv=NA, scale="none")
```

Observamos claramente que las variables AFNLWGT, POTHVAL y INTVAL no están correlacionadas con el resto de variables (en función de la variable, en mayor o menor medida).

R cuenta con el paquete *corrgram* que nos permite realizar representaciones gráficas de las correlaciones de forma más potente. Podemos ver un par de ejemplos en la Figura 4.

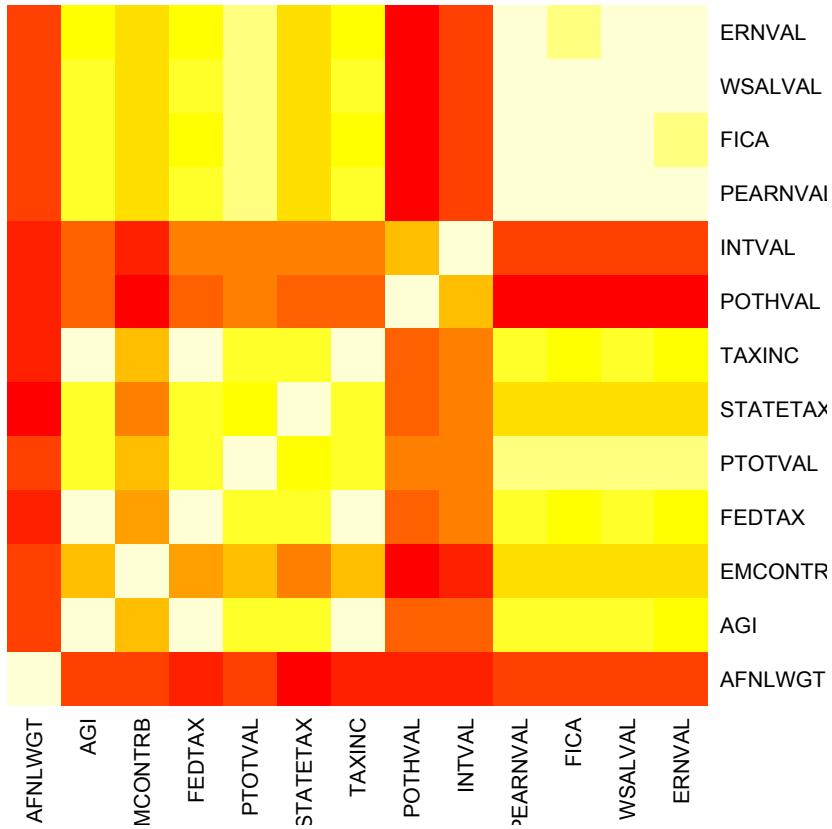


Figura 3: Mapa de calor de las correlaciones de las variables de Census.

4. Representaciones gráficas de los datos

Hemos visto que existen muchas formas de representar gráficamente los datos con los que trabajamos. A continuación mostramos algunos ejemplos especialmente relevantes.

4.1. Histogramas

Para representar los histogramas de cada variable del conjunto de datos podemos usar el siguiente comando de R cambiando en cada caso el valor de la columna y el nombre que queremos que aparezca en el título y la etiqueta del eje x :

```
> hist(census[,1],main=paste("Histograma de la Variable AFNLWGT"),
      xlab=paste("AFNLWGT"))
```

La Figura 5 muestra el resultado del comando anterior, y la Figure 6 muestra otros tres histogramas a modo de ejemplo.

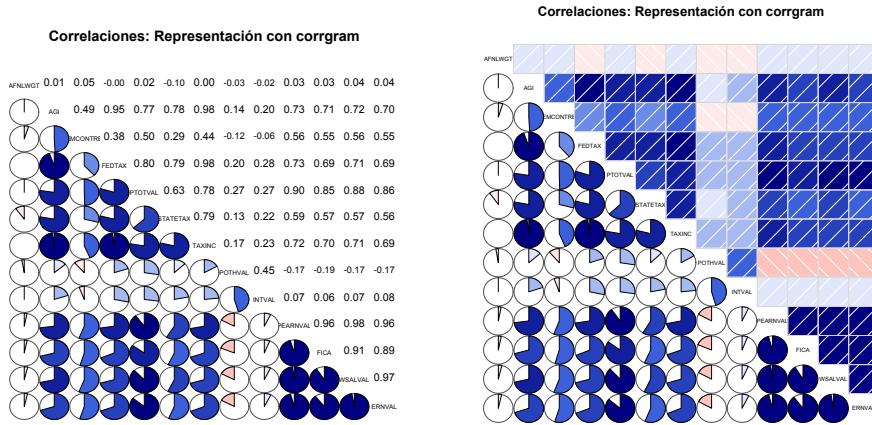


Figura 4: Ejemplos del uso de *corrgram* sobre CENSUS.

4.2. Representación de la estimación de la función de densidad por kernels

R nos proporciona una forma sencilla de estimar la función de densidad de una variable aleatoria a partir de una muestra de valores. No entraremos en los detalles matemáticos porque quedan fuera del objeto de esta asignatura pero mostramos algunos resultados obtenidos con R.

Usamos el comando siguiente (adaptándolo en cada caso a la variable que queremos representar).

```
> plot(density(census[,1]),main=paste("Estimación de la función de densidad de AFNLWGT"),xlab=paste("1080 muestras"))
```

El resultado de este comando se muestra en la Figura 7.

4.3. Representación tipo "plot"

Resulta sencillo representar mediante el comando *plot* los valores de una determinada variable, ello puede ayudarnos a ver gráficamente (de forma intuitiva) si la variable es homogénea y si presenta valores muy alejados de la media.

Podemos usar el siguiente comando para mostrar los valores de las variables POTHVAL y INTVAL:

```
> plot(census[,8],main=paste("Plot de los valores de la variable POTHVAL"),ylab=paste("POTHVAL"),xlab=paste("Índice"))

> plot(census[,9],main=paste("Plot de los valores de la variable INTVAL"),ylab=paste("INTVAL"),xlab=paste("Índice"))
```

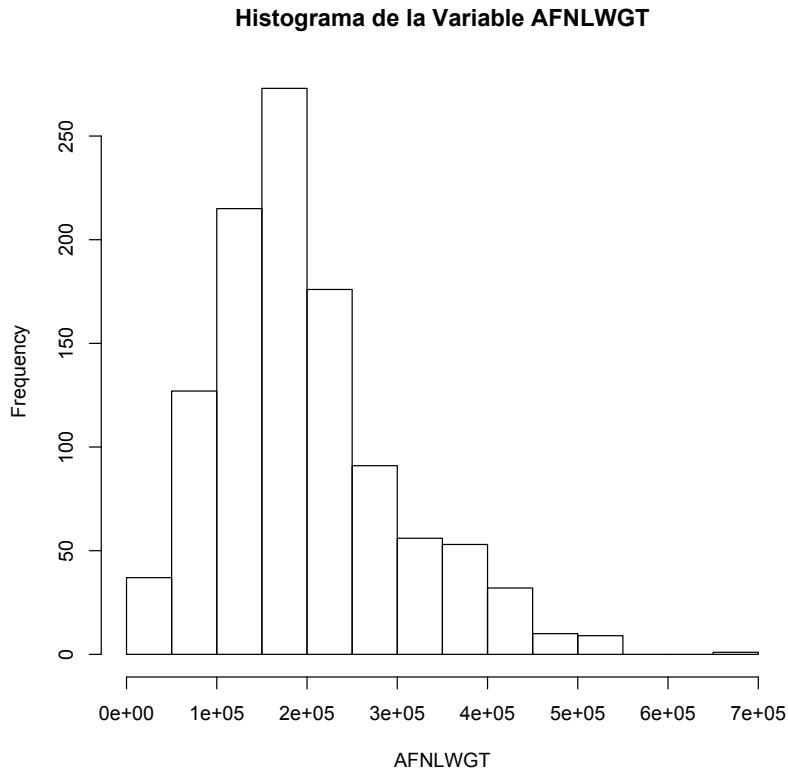


Figura 5: Ejemplo de representación de datos mediante un histograma. La figura muestra el caso de la variable AFNLWGT del conjunto de datos CENSUS.

El resultado se muestra en la Figura 8

Podemos representar de forma sencilla los puntos que resultan de combinar las dimensiones dos a dos mediante el siguiente comando de R:

```
> pairs(census)
```

El resultado que obtenemos se muestra en la Figure 9.

4.4. Otras representaciones interesantes

Existen muchos tipos de gráficos interesantes que podemos obtener mediante el uso de funciones sencillas de R. A continuación se muestran algunos ejemplos.

Diagramas de estrellas y segmentos (véase Figura 10).

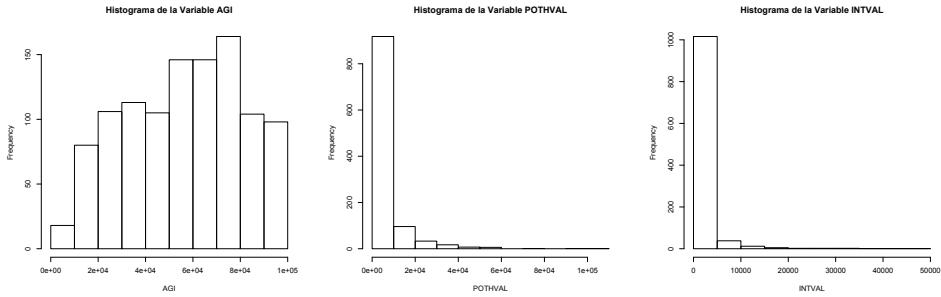


Figura 6: Otros ejemplos con las variables AGI, POTHVAL y INTVAL.

```
> stars(correlaciones, main=paste("Diagrama de Estrellas con las
correlaciones"), draw.segments=FALSE)
> stars(correlaciones, main=paste("Diagrama de Estrellas con las
correlaciones"), draw.segments=TRUE)
```

Diagramas de Caras de Chernoff (Véase la Figura 11).

```
> faces(t(census),main=paste("Caras de Chernoff de las 13 variables
de CENSUS"))
> faces(correlaciones,main=paste("Caras de Chernoff sobre la matriz
de correlaciones"))
```

5. Estudio de valores atípicos

Para estudiar los valores atípicos puede resultar de interés estudiar en primer lugar la homogeneidad de los datos. Si las desviaciones $d_{ij} = (x_{ij} - \bar{x}_j)^2$ son muy distintas, esto sugiere que hay datos que se separan mucho de la media y que tenemos por tanto alta heterogeneidad. Una posible medida de homogeneidad es la varianza de las d_{ij} , que podemos calcular mediante la siguiente expresión:

$$\frac{1}{n} \sum_{i_1}^n (d_{ij} - s_j^2)^2$$

o, si lo preferimos, el coeficiente de homogeneidad que calculamos dividiendo la expresión anterior por s^4 :

$$H_j = \frac{\frac{1}{n} \sum_{i_1}^n (d_{ij} - s_j^2)^2}{s_j^4}$$

Desarrollando el cuadrado del numerador, H_j puede expresarse como:

$$H_j = \frac{\frac{1}{n} \sum (x_{ij} - \bar{x}_j)^4}{s_j^4} - 1 = K_j - 1$$

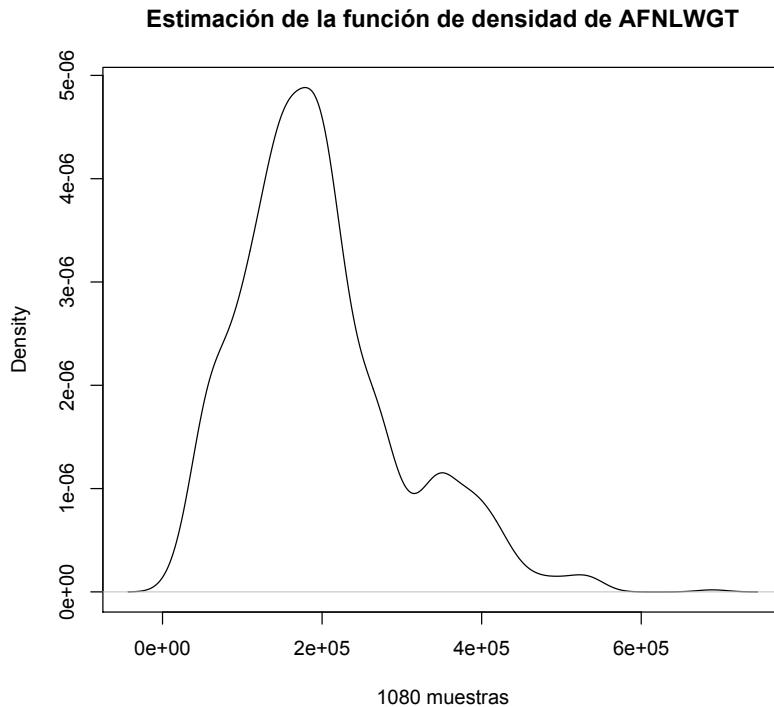


Figura 7: Ejemplo de representación de datos mediante una estimación de la función de densidad con 1080 muestras de la variable AFNLWGT.

donde K_j es el coeficiente de Kurtosis, el cual puede usarse como una forma alternativa de medir la homogeneidad.

Si hay unos pocos datos atípicos muy alejados del resto, la variabilidad de las desviaciones será grande, debido a estos valores y los coeficientes de kurtosis o de homogeneidad serán altos.

Un caso especialmente importante de heterogeneidad es la presencia de una pequeña proporción de observaciones atípicas (outliers). El coeficiente de kurtosis puede ayudar a detectar la presencia de valores atípicos, ya que tomará un valor alto, mayor que 7 u 8. Siempre que observemos un valor alto de la kurtosis para una variable esto implica heterogeneidad por uno o dos atípicos muy alejados del resto.

Podemos calcular fácilmente el coeficiente de Kurtosis usando la función *kurtosis* del paquete *moments* de R.

```
> kurt<-kurtosis(census)
```

Y obtenemos como resultado:

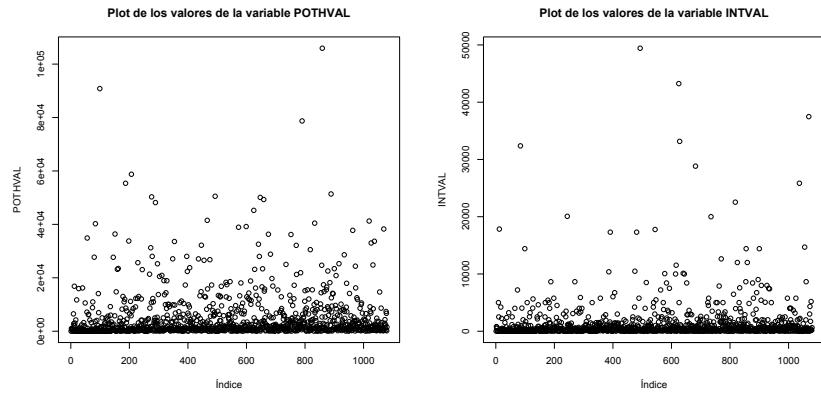


Figura 8: Plot de las variables POTHVAL y INTVAL

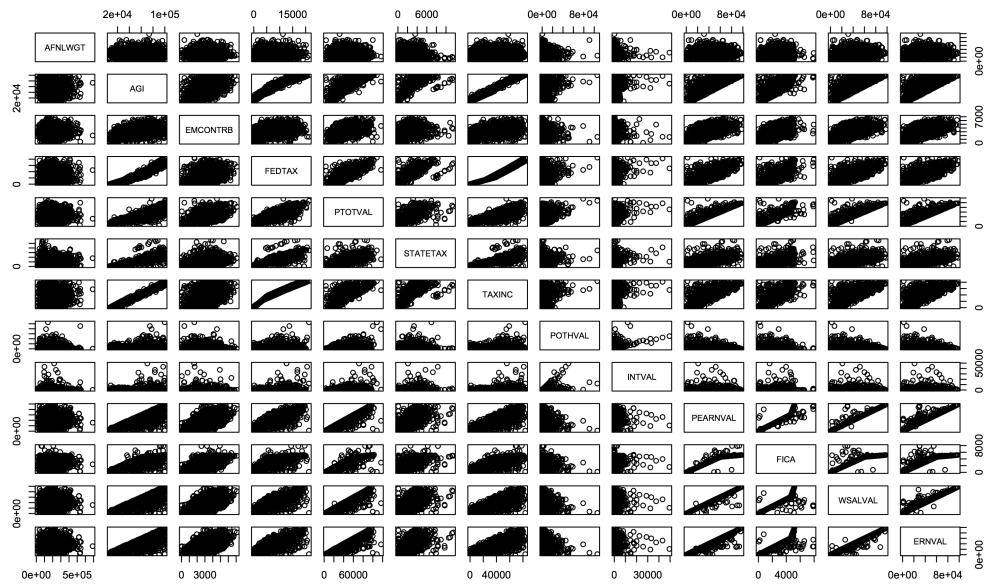


Figura 9: Representación de las variables del conjunto Census dos a dos

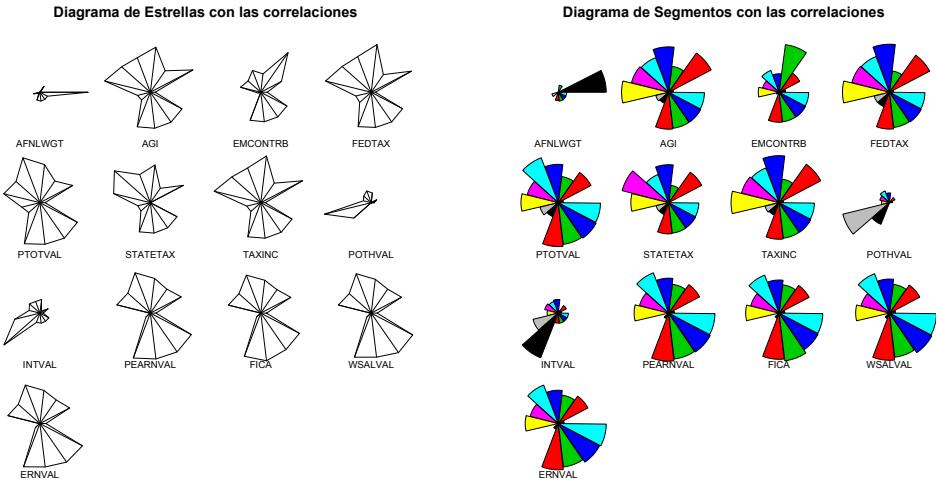


Figura 10: Diagrama de estrellas y segmentos

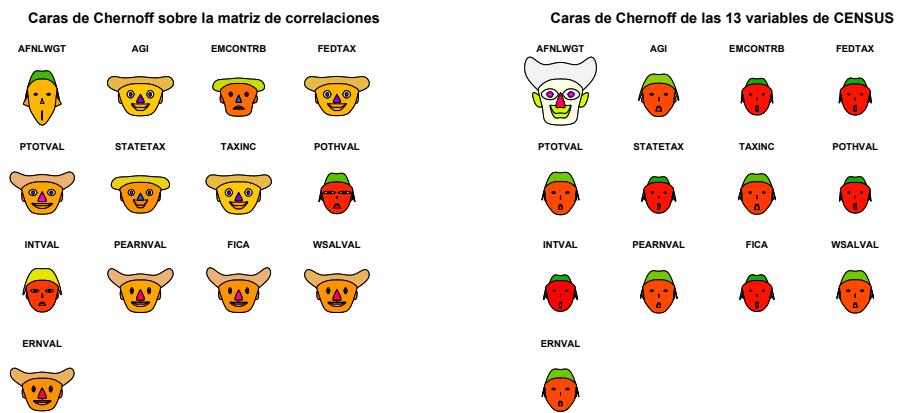


Figura 11: Ejemplo de diagrama de caras de Chernoff

```

AFNLWGT AGI      EMCONTRB FEDTAX   PTOTVAL STATETAX TAXINC   POTHVAL INTVAL    PEARNVAL FICA      WSALVAL ERNVAL
3.965949 1.983583 2.371602 2.242560 2.560684 4.569022 2.030749 29.582254 65.030361 2.485896 2.259823 2.503381 2.470991

```

Este resultado nos indica que existen valores atípicos en las variables POTHVAL y INTVAL. Podemos confirmar visualmente esta indicación mediante la representación gráfica de las variables con el comando de R *boxplot*:

```
> boxplot(census)
```

El resultado se muestra en la Figura 12.

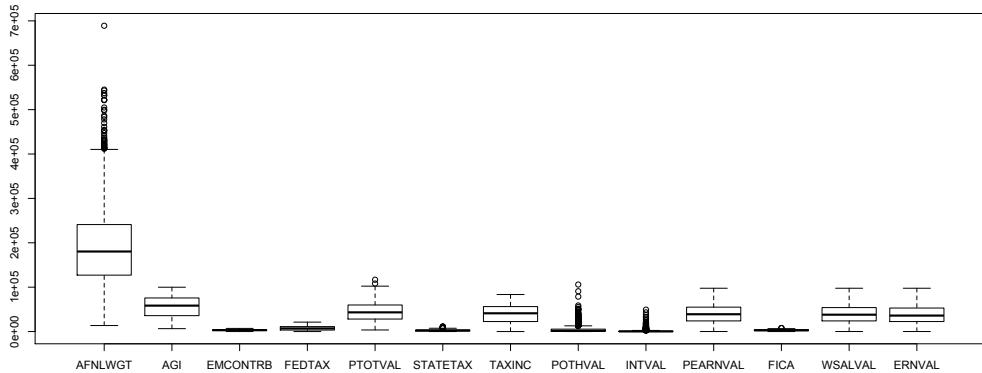


Figura 12: Diagrama de cajas del conjunto Census)

Tal como esperábamos encontramos valores atípicos en las variables POTHVAL y INTVAL. Además, parece que existan valores atípicos también en la variable AFNLWGT, sin embargo (dado que la kurtosis es baja) ello podría deberse a la existencia de dos poblaciones en la variable.

Para detectar los valores atípicos podemos usar una aproximación univariante y considerar sospechosas todas aquellas observaciones tales que:

$$\frac{|x_i - \text{med}(x)|}{\text{MEDA}(x)} > 4,5$$

Para hacer este análisis podemos usar el código de R siguiente:

```

atipicos <- matrix(0,nrow=13,ncol=1080)
total<-0
for (j in 1:13)
{
  for (i in 1:1080)

```

```

{
  z <- abs(census[i,j]-medianas[j])
  z <- z/medas[j]
  if (z > 4.5)
  {
    atipicos[j,i]<-z;
    total<-total+1
  }
}
}
print(total)

```

Este código nos devuelve el número de atípicos detectados, que podemos representar mediante el comando *myImagePlot*¹:

```
> myImagePlot(atipicos, yLabels=colnames(census))
```

y que se muestra en la Figure 13.

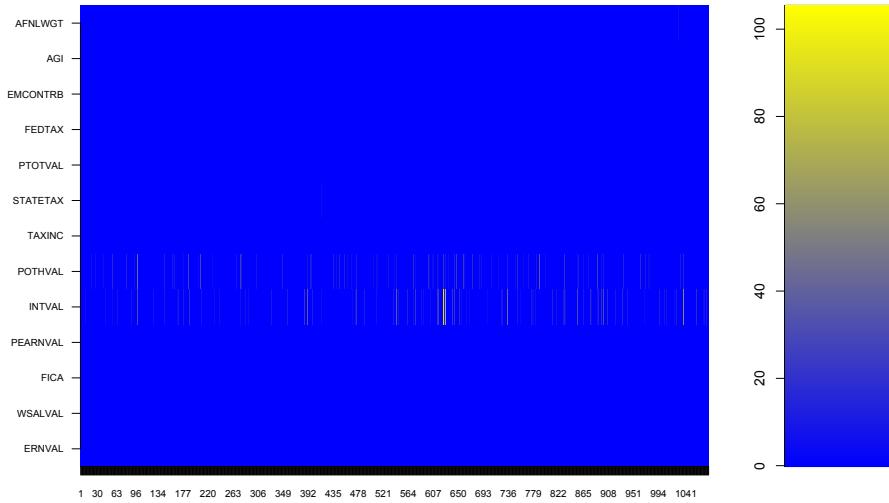


Figura 13: Valores atípicos detectados. Las líneas de color representan los valores obtenidos por $\frac{|x_i - \text{med}(x)|}{\text{MEDA}(x)}$. Obsérvese que los valores atípicos se concentran en las variables POTHVAL y INTVAL.

Usando como criterio de corte el valor 4.5 obtenemos 297 valores atípicos, que afectan a 250 registros distintos (23%). Si usamos un criterio más

¹http://www.phaget4.org/R/image_matrix.html

conservador con un valor de corte de 9 obtenemos 146 valores atípicos, que afectan a 123 registros distintos (11 %).

No reproducimos aquí los resultados exactos obtenidos, sin embargo bastaría mostrar la matriz “*atipicos*”.

La aproximación anterior es esencialmente univariante y puede interesar-nos tener una visión multivariante, para ello podemos usar la distancia de Mahalanobis y calcular que *filas* (registros) de nuestro conjunto CENSUS se encuentran más alejadas del resto.

Para ello podemos usar el comando de R:

```
> MahalDist<-mahalanobis(census,medias,covariancia,tol=1e-19)
```

Podemos representar gráficamente el resultado con un diagrama tipo “plot” para ver que registros están más alejados, y podemos representar la aproximación de la función de densidad de la distancia de Mahalanobis obtenida para tener una idea del número de atípicos.

```
> plot(density(MahalDist, bw=0.1),main="Est. fun. de densidad distancias de Mahalanobis sobre CENSUS",xlab=paste(" "), ylab=paste("Densidad"))
> plot(MahalDist,main="Diagrama de puntos Distancia Mahalanobis sobre CENSUS",xlab=paste("Registro"), ylab=paste("Distancia"))
```

Obtenemos el resultado de la Figura 14. Puede observarse con una simple exploración visual que existen diversos valores atípicos (p.e. aquellos por encima de 50).

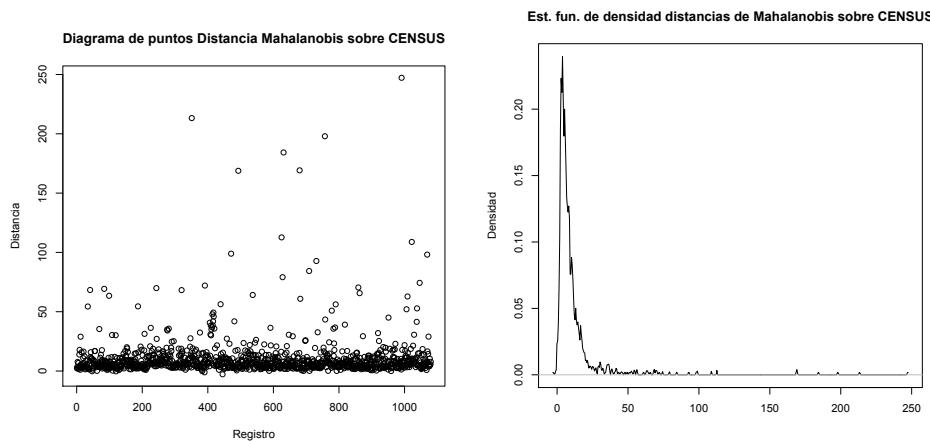


Figura 14: Representación de las distancias de Mahalanobis (por puntos y por densidad)

Los registros que podemos considerar atípicos fijado el criterio anterior son un total de 32 (3 %), que es un valor habitual en conjuntos de datos bien trabajados. En concreto, los registros que consideraríamos outliers son:

34, 41, 84, 99, 187, 243, 320, 351, 391, 439, 471, 493, 537, 625, 628, 631, 680 682, 709, 731, 757, 778, 790, 859, 863, 991, 1006, 1009, 1022, 1038, 1046, 1069

También se podría realizar un estudio más preciso mediante una aproximación iterativa en la que el proceso de detección y eliminación de k outliers se repetiría hasta que la kustosis se estabilizara por debajo de 8 en todas las variables (una a una) o hasta que no existiesen datos x_i tales que $\frac{|x_i - \text{med}(x)|}{\text{MEDA}(x)} > 4,5$.

Técnicas más avanzadas de detección de valores atípicos implican el uso de proyecciones, contrastes de hipótesis (e.g. test de la χ^2) y análisis de componentes principales. Dado que estos temas no se han visto hasta el momento, solo mencionamos la posibilidad de usarlas para que aquellos estudiantes que lo deseen puedan profundizar en el tema.

Se puede profundizar en el estudio de datos atípicos mediante R usando los siguientes paquetes:

- outliers (<http://cran.r-project.org/web/packages/outliers/outliers.pdf>)
- Data Mining with R (<http://cran.r-project.org/web/packages/DMwR/index.html>)
- MASS (<http://cran.r-project.org/web/packages/MASS/MASS.pdf>)

Para más información complementaria sobre detección de outliers, el estudiante puede consultar:

1. Breunig, Markus M., et al. "LOF: identifying density-based local outliers." ACM sigmod record. Vol. 29. No. 2. ACM, 2000.
2. Kriegel, Hans-Peter, Peer Kröger, and Arthur Zimek. "Outlier detection techniques." Tutorial at the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining. 2009.
3. <http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf>
4. Aggarwal, Charu C., and Philip S. Yu. "Outlier detection for high dimensional data." ACM Sigmod Record. Vol. 30. No. 2. ACM, 2001.