Practica I: Análisis descriptivo de datos. Modelos estadísticos

Análisis Inteligente de Datos

Universidad: UNIVERSIDAD POLITÉCNICA DE MADRID

Campus: MONTEGANCEDO

Centro: ESCUELA TÉCNICA SUPERIOR DE INGENIEROS INFORMÁTICOS

Identificación del Máster: MÁSTER UNIVERSITARIO EN INGENIERÍA INFORMÁTICA

Asignatura: ANÁLISIS INTELIGENTE DE DATOS

Nombres de los alumnos:

ABEL DE ANDRÉS GÓMEZ

GERMÁN SÁNCHEZ GRANADOS

Fecha de Entrega: 10/01/2016

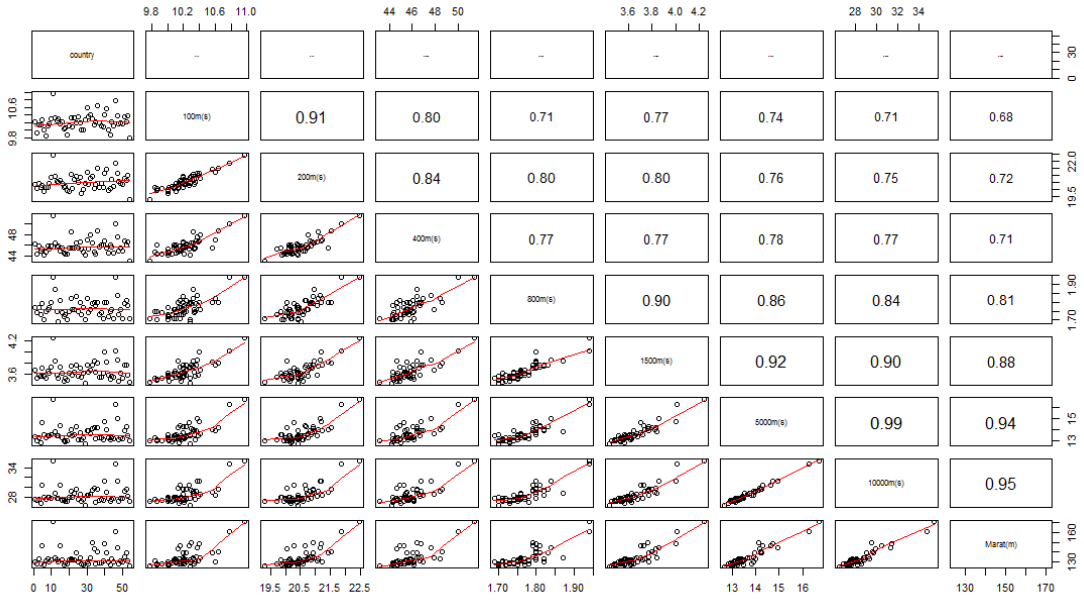
Explicación de los conjuntos de datos.

En cuanto a los datos que se analizaran y se estudiaran en esta práctica, podemos afirmar que se tratan de records nacionales de hombres en 54 países en distintas distancias. Las variables son, en orden:

- País
- 100 m (en segundos)
- 200 m (en segundos)
- 400 m (en segundos)
- 800 m (en minutos)
- 1500 m (en minutos)
- 5000 m (en minutos)
- 10000 m (en minutos)
- Maratón (en minutos)

Análisis descriptivo univariante

Para poder realizar un análisis univariante, deberemos seleccionar las dos variables que presente un mayor coeficiente de correlación de Pearson. Esto nos mostrara "cuanto" de relacionadas están las variables.



Como se puede comprobar, las variables que más correlación tienen entre sí son los 5000m(s) y los 10000m(s). Esto significa que para los corredores de 5000m(s) es más sencillo realizar la prueba de los 10000m(s) que cualquier otra prueba y viceversa.

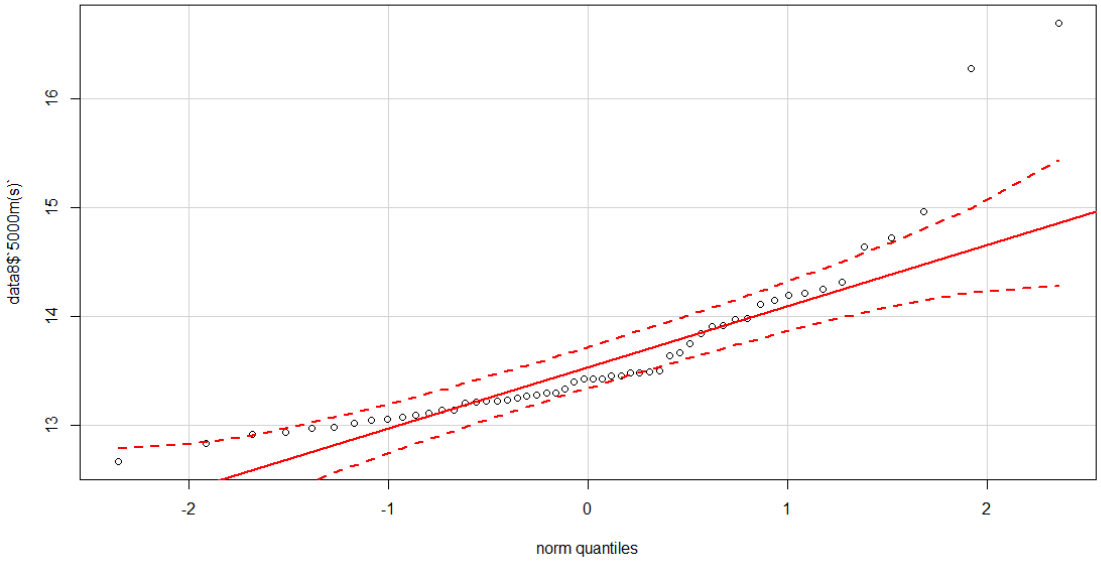
Análisis descriptivo univariante de la variable de los 5000m(s)

Test de normalidad

```
> shapiro.test(data8$`5000m(s)`)

        Shapiro-Wilk normality test

data:  data8$`5000m(s)`
W = 0.7884, p-value = 2.178e-07
```

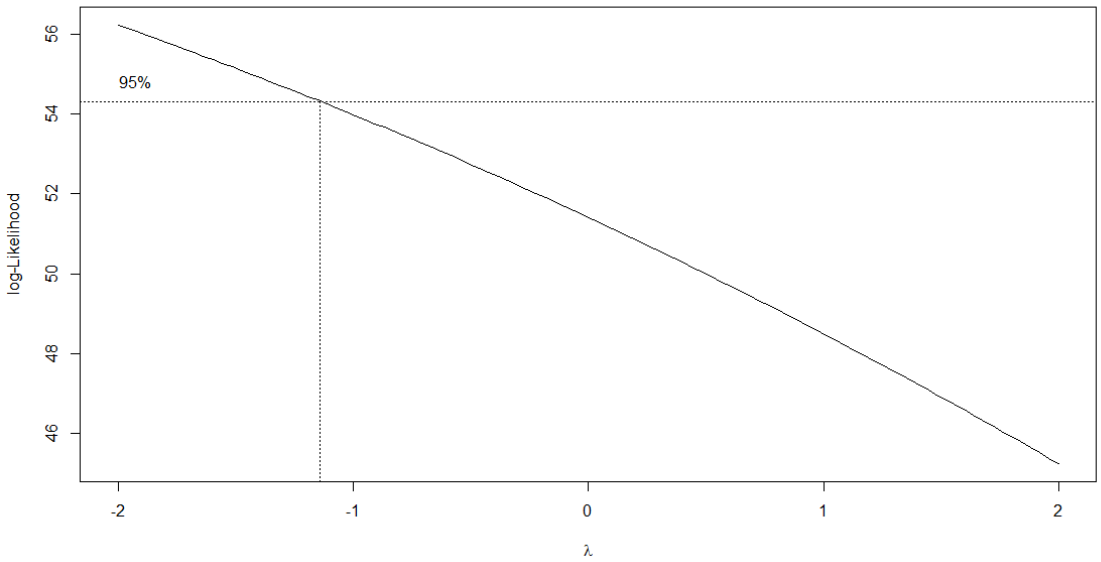
Como se aprecia, el p-valor se puede considerar como nulo y menor que 0.05, por lo tanto podemos decir que los datos no proceden de una distribución normal.

También podemos ver en la siguiente imagen que existen puntos que sobresalen de los limites, por lo tanto no se ajusta a una normal.



Transformación Box – Cox

Como ya hemos visto, nuestros datos no proceden de una transformación normal, puesto que se rechazó la normalidad mediante el test de Shapiro – Wilk, por lo tanto buscaremos una transformación que mejore dicha normalidad. Para ello utilizaremos una transformación del tipo Box – Cox.



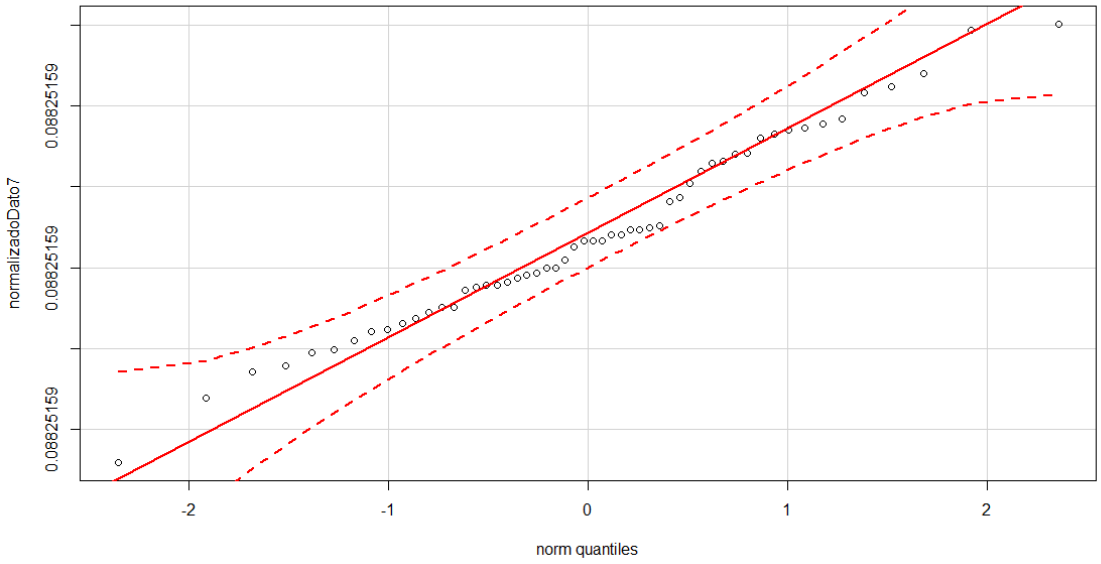
Como podemos comprobar, en la gráfica no vemos el valor máximo de la función, por lo tanto deberemos usar la función powerTransform para estimarlo.

```
> powerTransform(data8[,7])
Estimated transformation parameters
data8[, 7]
-11.33124
```

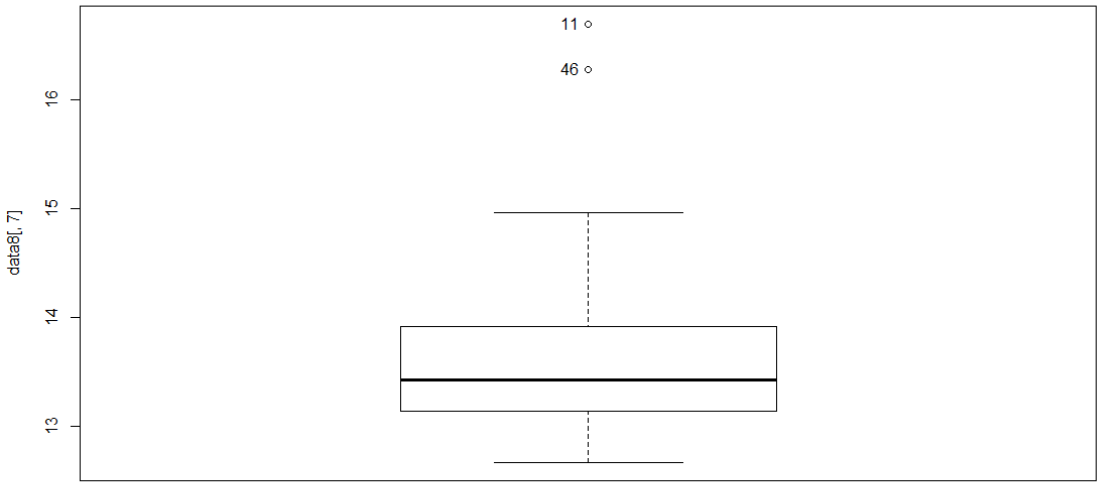
Una vez que tenemos el valor máximo de nuestra función, que se encuentra en λ=-11.33124, podremos normalizar nuestros datos.

```
> shapiro.test(normalizadoDato7)

        Shapiro-Wilk normality test

data:  normalizadoDato7
W = 0.98476, p-value = 0.7202
```

Una vez que se ha hecho la normalización de datos, vamos a comprobar gráficamente que de verdad se ha normalizado:



Datos Atípicos (Outliers)



Como se puede comprobar existen dos datos atípicos en la prueba de los 5000m(s), que corresponden con el dato 46 (Samoa) y el dato 11 (CookIslands). Esto significa que estos países países estan bastante lejos de la media de datos, o dicho de otra forma, son los países que realizan la prueba en grandes tiempos (son malos en la prueba de los 5000m(s)). Siendo Samoa mejor que CookIslands.

Asimetría y Kurtosis

```
> anscombe.test(data8[,7])

        Anscombe-Glynn kurtosis test

data:  data8[, 7]
kurt = 8.6112, z = 3.8721, p-value = 0.0001079
alternative hypothesis: kurtosis is not equal to 3
```

Como se puede observar en este test, podemos comprobar que la variable de los 5000m(s) no corresponde a una distribución normal, ya que el coeficiente de Kurtosis es mayor que 3 (8.6112). Por lo tanto, se trata de una distribución leptocurtica.

Respecto a la simetría, podemos decir que es simétrica por la derecha ya que z=3.8751.

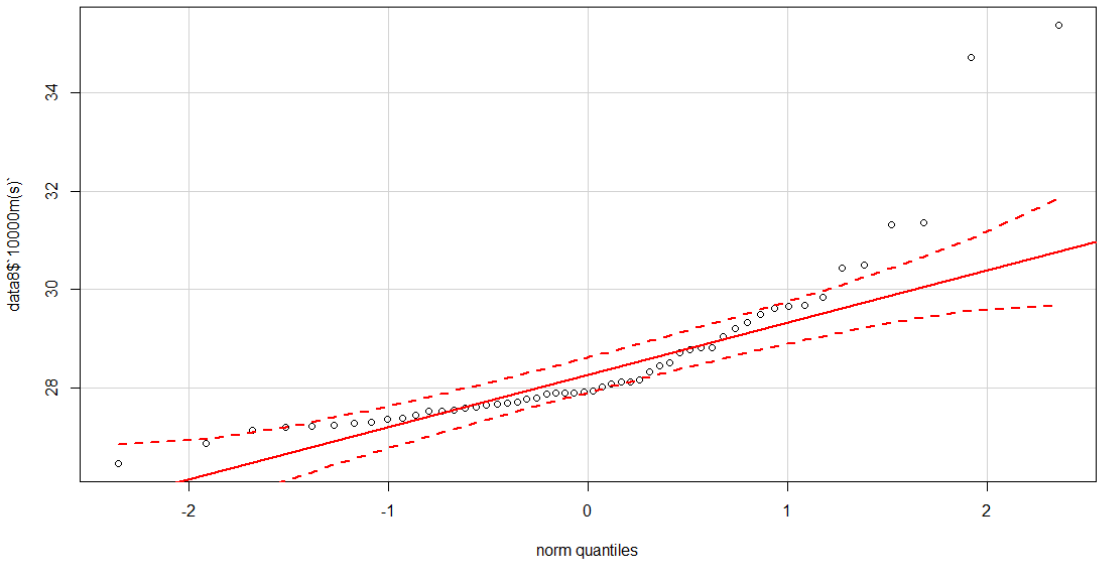
Análisis descriptivo univariante de la variable de los 10000m(s)

Test de normalidad

```
> shapiro.test(data8$`10000m(s)`)

        Shapiro-Wilk normality test

data:  data8$`10000m(s)`
W = 0.7543, p-value = 3.877e-08
```

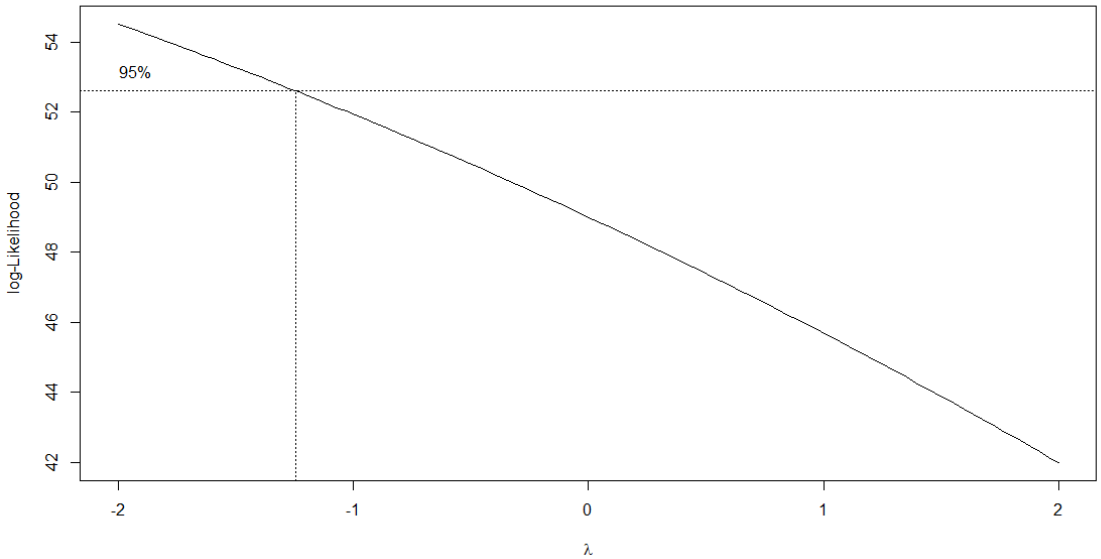
Como se aprecia, el p-valor se puede considerar como nulo y menor que 0.05, por lo tanto podemos decir que los datos no proceden de una distribución normal.

También podemos ver en la siguiente imagen que existen puntos que sobresalen de los límites, por lo tanto no se ajusta a una normal.



Transformación Box – Cox

Como ya hemos visto, nuestros datos no proceden de una transformación normal, puesto que se rechazó la normalidad mediante el test de Shapiro – Wilk, por lo tanto buscaremos una transformación que mejore dicha normalidad. Para ello utilizaremos una transformación del tipo Box – Cox.



Como podemos comprobar, en la gráfica no vemos el valor máximo de la función, por lo tanto deberemos usar la función powerTransform para estimarlo.

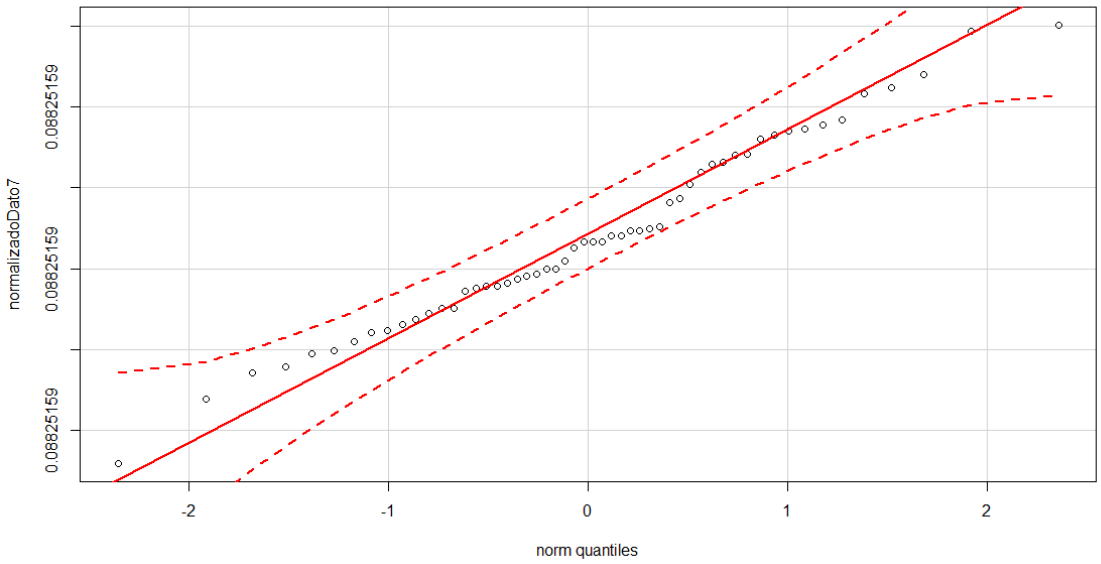
```
> powerTransform(data8[,8])
Estimated transformation parameters
data8[, 8]
```

```
-8.868982
```

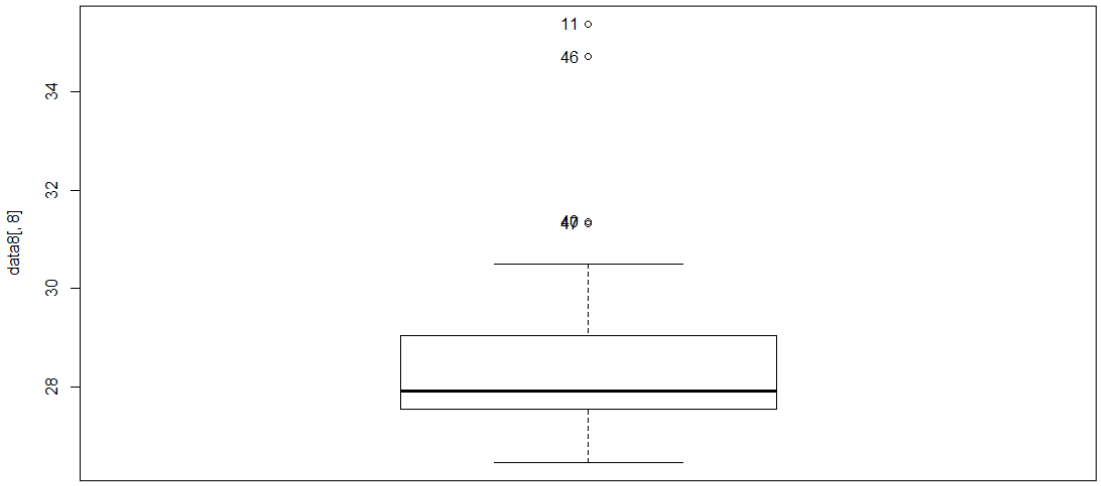
Una vez que tenemos el valor máximo de nuestra función, que se encuentra en λ=-8.868982, podremos normalizar nuestros datos.

```
> shapiro.test(normalizadoDato8)

        Shapiro-Wilk normality test

data:  normalizadoDato8
W = 0.96294, p-value = 0.09338
```

Una vez que se ha hecho la normalización de datos, vamos a comprobar gráficamente que de verdad se ha normalizado:



Datos Atípicos (Outliers)



Como se puede comprobar existen dos datos atípicos en la prueba de los 10000m(s), que corresponden con el dato 46 (Samoa), 47 (Singapore), 40(PapuaNewGuinea) y el dato 11 (CookIslands). Esto significa que estos países países están bastante lejos de la media de datos, o dicho de otra forma, son los países que realizan la prueba en grandes tiempos (son malos en la prueba de los 10000m(s)).

Asimetría y Kurtosis

```
> anscombe.test(data8[,8])

Anscombe-Glynn kurtosis test

data:  data8[, 8]
kurt = 9.3496, z = 4.0514, p-value = 5.09e-05
alternative hypothesis: kurtosis is not equal to 3
```

Como se puede observar en este test, podemos comprobar que la variable de los 10000m(s) no corresponde a una distribución normal, ya que el coeficiente de Kurtosis es mayor que 3 (9.3496). Por lo tanto, se trata de una distribución leptocurtica.

Respecto a la simetría, podemos decir que es simétrica por la derecha ya que z=4.0514.

Análisis Bivariante

En cuanto a los datos que se utilizaran en el análisis bivariante, serán aquellas que tengan mayor coeficiente de correlación. Recordamos que eran las variables de los 5000m(m) y 10000m(m).

Datos Atípicos (Outliers)

Para poder obtener los datos atípicos en el análisis bivariante, tendremos que realizar el grafico de dispersión.

Para ello, primero es necesario calcular la matriz de correlaciones entre las variables a estudiar:

```
> cor(data8[,7:8])
           5000m(s) 10000m(s)
5000m(s)   1.0000000 0.9882324
10000m(s) 0.9882324 1.0000000
```

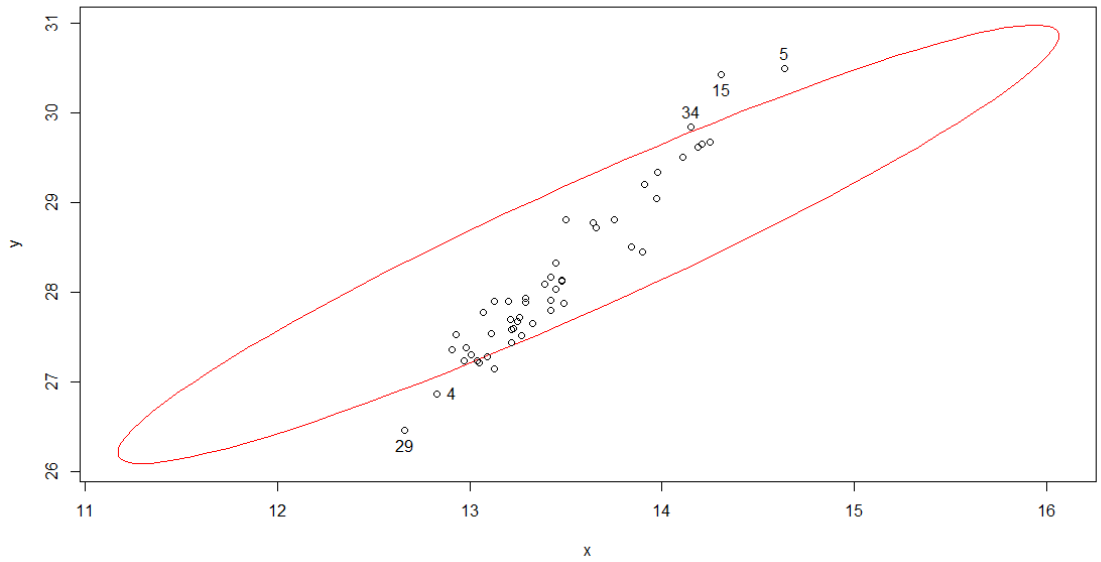
Se puede comprobar de Nuevo que ambas variables están muy correlacionadas entre sí, existiendo casi una correlación perfecta.

Por otro lado necesitamos el vector de medias que será el centro de nuestra elipse:

```
> colMeans(data8[,7:8])
  5000m(s) 10000m(s)
  13.61759  28.53519
```

A continuación vamos a diseñar la elipse con un 95% de confianza por defecto:

```
> plot(ellipse(0.95,centre=c(13.61759,28.53519)),type='l',col=2)
> points(data8[,7],data8[,8])
```



Como se puede comprobar, los paises que corresponden con la posicion 4(Belgica), 5(Bermuda),15(Republica Dominicana),29(Kenya),34(Mauritania) son aquellos que estan fuera de lo normal en las pruebas analizadas. Concretamente los paises 4 y 29 destacan por ser mejores en ambas pruebas mientras que los paises 5, 15 y 34 destacan por ser peores (las medias de las marcas son peores).

Test de Normalidad

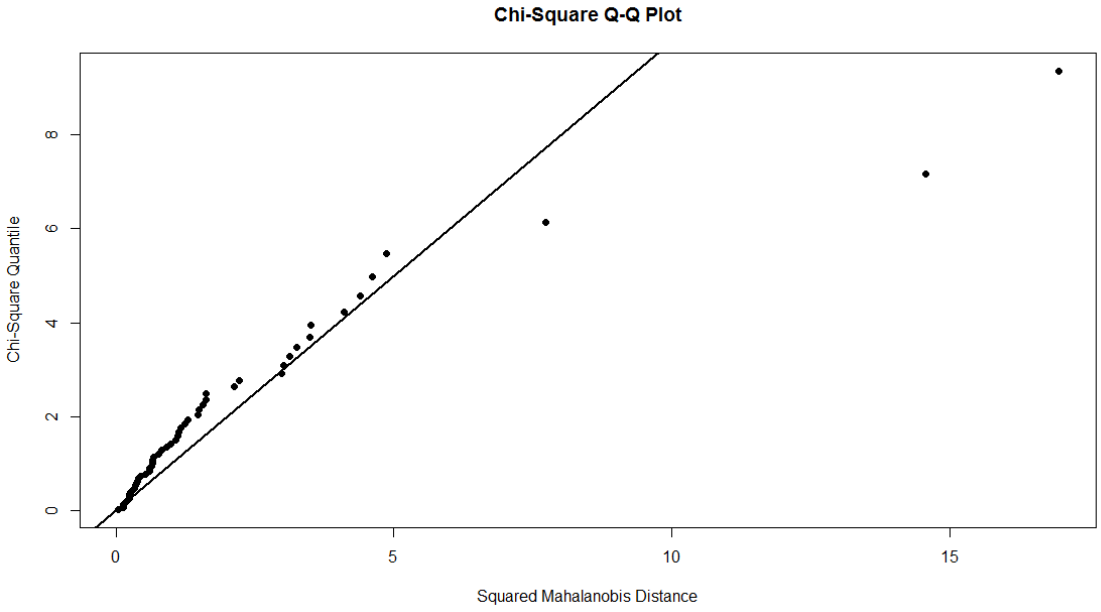
Para realizar un test de normalidad utilizaremos el test de Mardia.

```
> mardiaTest(data8[,7:8],qqplot=T)
   Mardia's Multivariate Normality Test
---------------------------------------
   data : data8[, 7:8]

   g1p            : 5.618585
   chi.skew       : 50.56727
   p.value.skew   : 2.748799e-10
```

```
   g2p            : 13.70787
   z.kurtosis     : 5.243016
   p.value.kurt   : 1.579733e-07

   chi.small.skew : 55.39077
   p.value.small  : 2.690664e-11

   Result          : Data are not multivariate normal.
---------------------------------------
```

El p-valor de la Kurtosis es próximo a 0, por lo tanto rechazamos la hipótesis de la normalidad.

Como se puede comprobar en la gráfica, los datos no se ajustan a una normal.



Para comprobar el máximo grado de verosimilitud deberemos utilizar la función powerTransform.

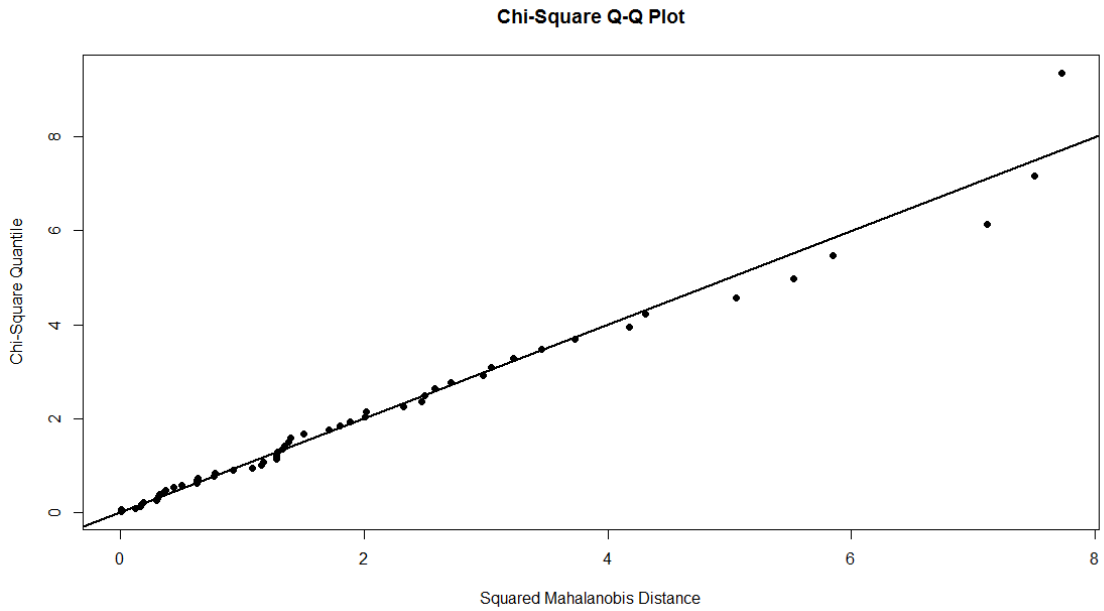
```
> powerTransform(data8[,7])
Estimated transformation parameters
    data8[, 7]
     -11.33124
> powerTransform(data8[,8])
Estimated transformation parameters
    data8[, 8]
     -8.868982
```

```
> summary(powerTransform(data8[,7]))
bcPower Transformation to Normality

Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
data8[, 7]  -11.3312    0.012         -11.3547         -11.3077

Likelihood ratio tests about transformation parameters
LRT df         pval
LR test, lambda = (0) 27.50671  1 1.565505e-07
LR test, lambda = (1) 33.33486  1 7.757957e-09
> summary(powerTransform(data8[,8]))
bcPower Transformation to Normality
```

```
Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
data8[, 8]    -8.869      NaN              NaN              NaN

Likelihood ratio tests about transformation parameters
LRT df         pval
LR test, lambda = (0) 30.55537  1 3.244726e-08
LR test, lambda = (1) 37.18672  1 1.073421e-09
```

A continuación se realizara una normalización mediante el uso de los lambdas obtenidos.

```
> transformacion=bcPower(data8[,7:8], c(-11.3312,-8.869))
> mardiaTest(transformacion,qqplot=T)

   Mardia's Multivariate Normality Test
---------------------------------------
   data : transformacion

   g1p            : 1.086281
   chi.skew       : 9.776527
   p.value.skew   : 0.0443651

   g2p            : 7.77932
   z.kurtosis     : -0.2027072
   p.value.kurt   : 0.8393639

   chi.small.skew : 10.70909
   p.value.small  : 0.03003554

   Result          : Data are not multivariate normal.
---------------------------------------
```



Como se puede comprobar, mediante la transformación hemos conseguido que nuestros datos estén más normalizados. Pero no se encuentran todos normalizados, ya que los p-valores de la Kurtosis y de la asimetría son distintos de 0, por lo tanto no se podría aceptar del todo la transformación.

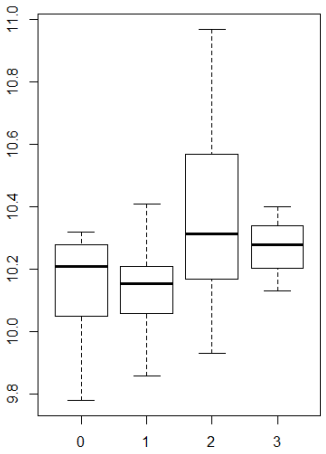
Análisis Multivariante

Medias por Continentes

En primer lugar, se va a realizar un breve estudio sobre las medias de las diferentes variables por continente.

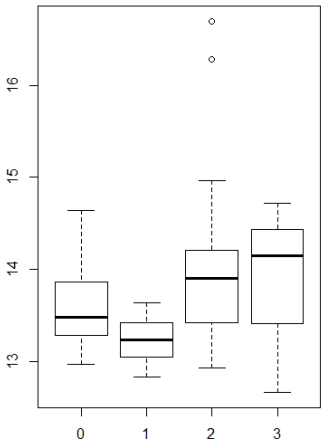
Para ello, en primer lugar lo que se realizara en esta práctica es dividir los países por sus respectivos continentes. Para que no exista grandes diferencias en los datos de los distintos continentes, vamos a crear 4 continentes que serán: América=0, Europa=1, Asia=2 y África=3.

Nosotros vamos a tomar la variable de los 100m(s), se podría tomar otra cualquiera. Y la graficaremos mediante un gráfico de caja.



Como se puede comprobar, la media del continente 1 (Europa) es la más baja, por lo tanto, los corredores de este continente son los más rápidos. Por el contrario, los corredores del continente 2 (Asia) son los que peor media tienen (más alta) y por lo tanto son los más lentos.

Sin embargo, si tomamos la variable de los 5000m(s), vamos a comprobar que el continente Europeo son los más rápidos de nuevo, pero los más lentos ya no son los asiáticos, sino que son los Africanos. Existen también unos puntos (valores atípicos) que significa que la media de esos países asiáticos está fuera de la media del continente Asiático, los motivos podrían ser que los países fueran tan pequeños, o tan poco poblados, que no tendrían la capacidad de tener atletas profesionales y son simplemente atletas no profesionales.



Matriz de correlaciones (R)

La matriz de correlaciones es aquella que además de ser simétrica y cuadrada, contiene los coeficientes de correlación entre los pares de variables fuera de la diagonal. Tiene en cuenta el efecto no solo de las variables que se están correlacionando sino del resto de variables.

	100m(s)	200m(s)	400m(s)	800m(s)	1500m(s)	5000m(s)	10000m(s)	Marat(m)
100m(s)	1.0000000	0.9147554	0.8041147	0.7119388	0.7657919	0.7398803	0.7147921	0.6764873
200m(s)	0.9147554	1.0000000	0.8449159	0.7969162	0.7950871	0.7613028	0.7479519	0.7211157
400m(s)	0.8041147	0.8449159	1.0000000	0.7677488	0.7715522	0.7796929	0.7657481	0.7126823
800m(s)	0.7119388	0.7969162	0.7677488	1.0000000	0.8957609	0.8606959	0.8431074	0.8069657
1500m(s)	0.7657919	0.7950871	0.7715522	0.8957609	1.0000000	0.9165224	0.9013380	0.8777788
5000m(s)	0.7398803	0.7613028	0.7796929	0.8606959	0.9165224	1.0000000	0.9882324	0.9441466
10000m(s)	0.7147921	0.7479519	0.7657481	0.8431074	0.9013380	0.9882324	1.0000000	0.9541630
Marat(m)	0.6764873	0.7211157	0.7126823	0.8069657	0.8777788	0.9441466	0.9541630	1.0000000

Como se puede comprobar en la matriz de correlaciones, la mayor correlación existente va a ser entre la variable de 10000 m (s) y la variable de Maratón (m). Esto significa que si crece una variable, también crecerá la otra.

Esto significa que para un corredor que realiza la prueba de los 5000 m(s) es más fácil obtener mejores resultados en la prueba de los 10000 m(s) que por ejemplo realizar la prueba de los 100 m (s).

Esto se debe a que físicamente, los corredores de largas distancias como pueden ser los 5000 m (s), 10000 m(s) y las pruebas de maratón están acostumbrados a la resistencia física.

Por el contrario, un corredor de velocidad, como podrían ser aquellos que realizan las pruebas de los 100 m(s), 200m(s), 400m(s) están preparados para las pruebas en las que se tiene que recorrer una distancia corta a la máxima velocidad posible y no para recorrer aquellas que son de largas distancias (ya que no tienen tanta resistencia).

Matriz de correlaciones parciales

La matriz de correlaciones parciales es aquella que mide las relaciones entre pares de variables eliminando el efecto de las restantes.

Coeficientes de correlación múltiple al cuadrado

El coeficiente de determinación mide la proporción de variabilidad total de la variable dependiente respecto a su media que es explicada por el modelo de regresión. Es usual expresar esta medida en tanto por ciento, multiplicándola por cien.

- r2multv(data8[,-1])

100m(s) 200m(s) 400m(s) 800m(s) 1500m(s) 5000m(s) 10000m(s) Marat(m)
0.8661801 0.8981543 0.7666577 0.8479119 0.8950478 0.9820568 0.9815583 0.9161253

Como se puede comprobar con los coeficientes de correlación múltiple la variable de los 5000m(s) es la que más "explicada" se ve linealmente por todas las demás variables, seguida de la variable de 10000m(s).

Coeficientes de dependencia efectiva

El coeficiente de dependencia efectiva es una medida global de dependencia lineal de los datos.

```
1-det(cor(data8[,-1]))^{1/8}
[1] 0.8321275
```

En este caso, la dependencia lineal explica el 83% de la variabilidad del conjunto de datos.

Autovalores y Autovectores

An eigenanalysis of the previous subset or other you may be interested in. Estudiar la matriz S o R (creo que en nuestro caso S) de ese conjunto de variables y sus auto valores para ver si existe dependencia lineal entre las variables. Identificar los datos atípicos.

Las matrices simétricas con coeficientes reales pueden recomponerse simplemente multiplicando la matriz de autovectores por su matriz de autovalores y ésta, a su vez, por la transpuesta de la matriz de autovectores. Este tipo de matrices y sus operaciones son importantes porque aparecen con frecuencia en el análisis estadístico multivariado.

El rango r =rango (S) determina la dimensión del espacio vectorial generado por las variables observables, es decir, el número de variables linealmente independientes es igual al rango de S y, a su vez, es igual al número de autovalores distintos de 0 de S.

Dicho de otra forma, que el determinante de la matriz de covarianza sea cero (|S|=0) significa que las medidas en una o varias variables del estudio pueden ser eliminadas. Una o varias variables son combinaciones lineales de las otras y, por tanto, redundantes en cuanto a la información proporcionada.

Realizando el análisis de autovalores y autovectores de la matriz de covarianza (S), obtenemos lo siguiente:

Autovalores

8.450724e+01
1.141345e+00

2.287648e-01
8.056601e-02
1.177928e-02
6.256523e-03
3.064681e-03
3.971719e-04

Autovectores

	V1	V2	V3	V4	V5	V6	V7	V8
1	-0.016521364	-0.09994775	0.008058901	-0.32409133	-0.312057689	0.882721015	0.0288828651	-0.0850102738
2	-0.043603961	-0.25282207	0.080820783	-0.89738426	0.172464956	-0.292153978	-0.0736065787	0.0428152608
3	-0.113581901	-0.91633442	0.253458358	0.28818053	0.011540098	0.001880073	0.0003886092	0.0020880219
4	-0.004643738	-0.01405243	-0.012268779	-0.02637904	-0.043536918	-0.126535307	0.1944547253	-0.9711927207
5	-0.014583470	-0.03076566	-0.037064086	-0.06485429	-0.205722936	-0.110116229	0.9446648850	0.2154569931
6	-0.078593405	-0.11653932	-0.376765236	-0.03502143	-0.826194160	-0.304718790	-0.2456938201	0.0353176224
7	-0.175189411	-0.20893763	-0.872775053	0.01809892	0.381997503	0.120192128	0.0563118120	-0.0071143206
8	-0.973561706	0.16745845	0.154755649	0.01273865	-0.002528303	0.003109968	-0.0026157492	-0.0008844997

Como podemos ver, a partir de la matriz de covarianza obtenida de nuestro conjunto de datos, se han obtenido dos autovalores bastante pequeños, muy próximos a cero.

En primer lugar 3.971719e-04 con autovector asociado {-0.973561706, 0.16745845, 0.154755649, 0.01273865, -0.002528303, 0.003109968, -0.0026157492, -0.0008844997}.

Los coeficientes más altos que se encuentran en el vector, indican las variables que están involucradas en la relación lineal. Como podemos comprobar, la primera variable es la que más valor tiene, por lo tanto la variable 100m (s) está muy involucrada en la relación lineal. Esta relación lineal significa que la variable de los 100 m (s) tiene una varianza muy pequeña, siendo casi constante.

En segundo lugar 3.064681e-03 con autovector asociado {-0.175189411, -0.20893763, -0.872775053, 0.01809892, 0.381997503, 0.120192128, 0.0563118120, -0.0071143206}.

Los coeficientes más altos que se encuentran en este último vector, también indican las variables más involucradas en la dependencia lineal. Como se puede observar, en este caso sería la tercera variable y también la quinta. Por lo tanto, podríamos decir que la variable de 400m(s) y 1500m(s) son las variables más involucradas en la dependencia lineal.

Como conclusiones, podríamos decir que las variables de los 100m(s), 400m(s) y 1500m(s) son las que menos información aportan al conjunto de datos e incluso podrían ser redundantes ya que son combinaciones lineales del resto de variables.

Datos Atípicos (Outliers)

Mediante la función *mvoutlier* obtendremos los datos atípicos:

```
> data8.out=pcout(data8[,-1])
> which(data8.out$wfinal01==0)
 5  6 11 12 15 21 30 31 33 34 36 40 41 46 47
```

Análisis de Componentes Principales

Un problema en el análisis de datos multivariante es la reducción de la dimensionalidad: es decir, si se puede conseguir con precisión los valores de las variables (p) con un pequeño subconjunto de ellas (r<p), habremos conseguido reducir la dimensión a costa de una pequeña perdida de información.

El análisis de componentes principales tiene este objetivo. Dada n observaciones de p variables, se analiza si es posible representar adecuadamente esta información con un conjunto menor de variables (construidas como combinaciones lineales de las originales).

El primer paso en el análisis de componentes principales consiste en la obtención de los valores y vectores propios de la matriz de covarianzas muestral o de la matriz de coeficientes de correlación que se obtienen a partir de los datos. En nuestro caso vamos a utilizar la matriz de correlaciones.

```
Importance of components:
                          PC1    PC2     PC3     PC4    PC5     PC6     PC7     PC8
Standard deviation     2.5891 0.7990 0.47700 0.45371 0.3124 0.26587 0.21666 0.09858
Proportion of Variance 0.8379 0.0798 0.02844 0.02573 0.0122 0.00884 0.00587 0.00121
Cumulative Proportion  0.8379 0.9177 0.94615 0.97188 0.9841 0.99292 0.99879 1.00000
```

Para elegir nuestras componentes principales, podremos utilizar dos métodos:

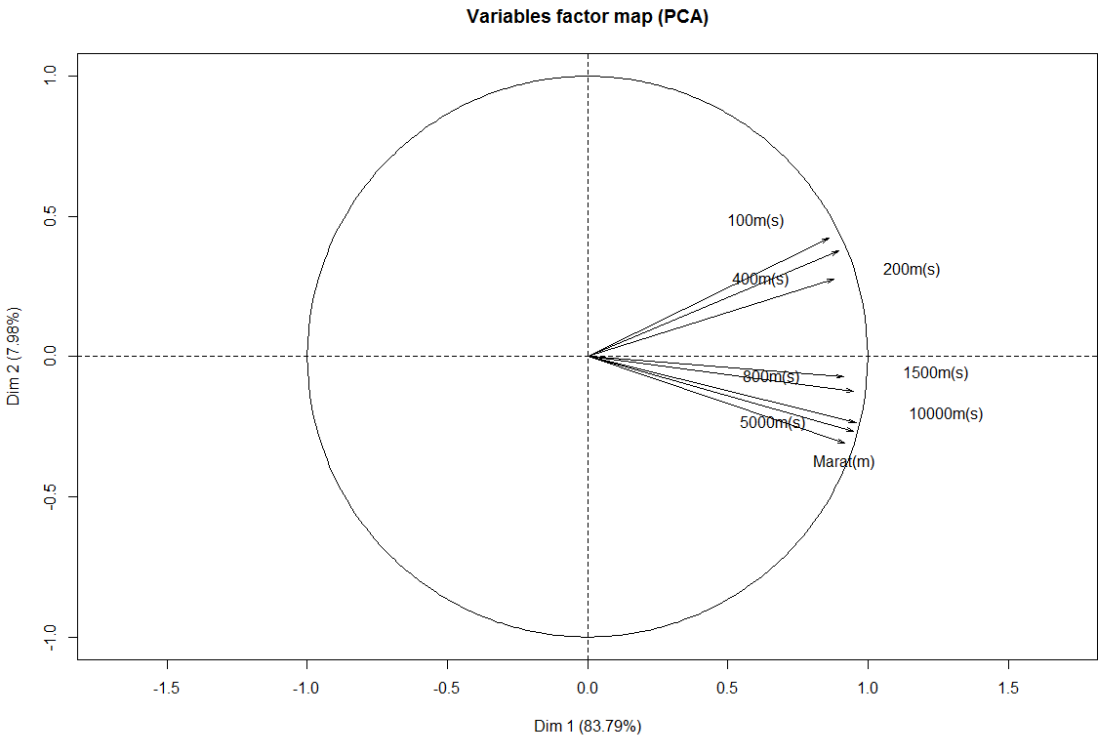
Por un lado, podemos utilizar **el criterio de Kaiser**, que consiste en conservar aquellos factores cuya desviación estándar al cuadrado asociada sea mayor que 1.

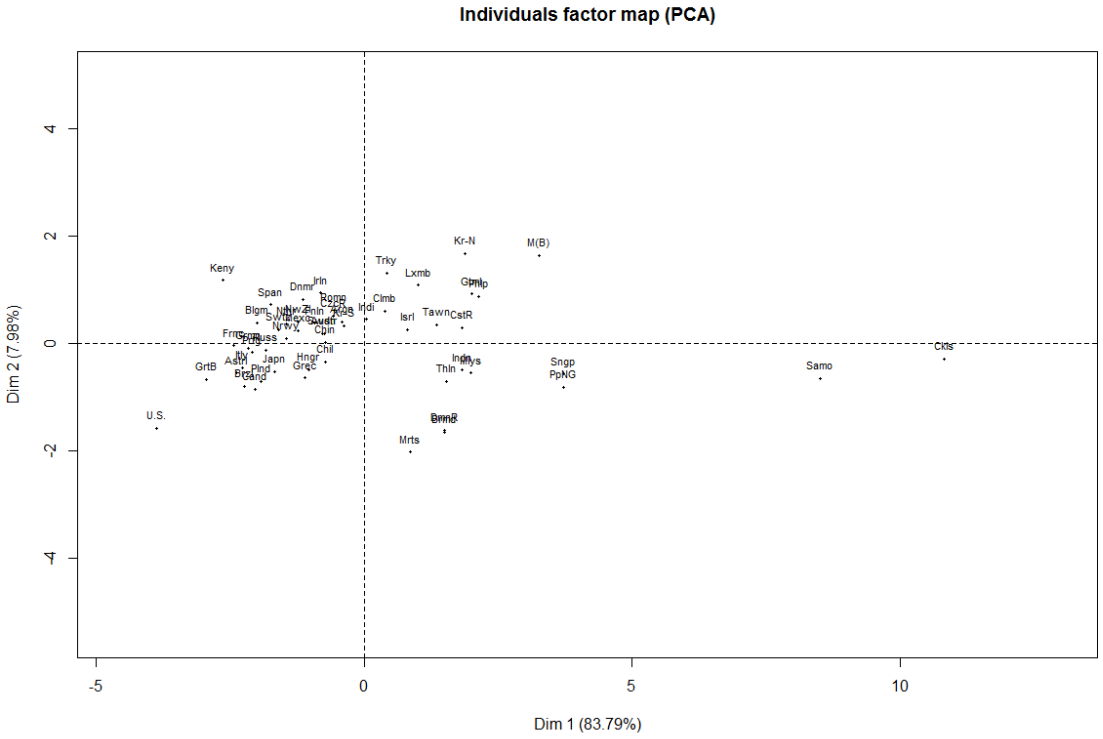
```
(data8.pca$sdev)^2
[1] 6.703289951 0.638410110 0.227524494 0.205849181 0.097577441 0.0706
87912 0.046942050 0.009718862
```

Otra forma para saber cuántos componentes tener en cuenta es mantener el número de componentes necesarios para explicar al menos un porcentaje del total de la varianza. Por ejemplo, es importante explicar al menos un 80% de la varianza. En nuestro caso, podríamos mantener con el primer componente (*proportion of variance* (PC1)) que es del 0,83 -->83%. Se obtiene del *summary.*

Para conseguir una mayor varianza (más información en los datos), nosotros vamos a utilizar P C1 y PC2, cuya proporción de información que contienen es del 91%.

Una vez que tenemos los dos componentes principales, vamos a pasar a analizar el grafico de i ndividuos y de variables obtenidos en el Análisis de Componentes Principales.





Para realizar el análisis, es necesario tener en cuenta ambos gráficos. En general, podemos ver que la nube de puntos se encuentra cercana al centro de los ejes, esto nos quiere decir que los países que se encuentran en dicha nube se encuentran en la misma media, es decir no destaca n en algo en particular.

Si nos centramos en Kenya, vemos que se encuentra fuera de la nube, por lo tanto, destaca en algo en particular. Concretamente destaca en las distancias largas, como pueden ser la maratón, los 5000m(s), etc…

Por el contrario, si tenemos en cuenta los Estados Unidos (U.S), destaca por lo contrario que Kenya, es un país cuyos atletas son mejores en las distancias cortas como los 100m(s), etc.

Si comparamos España (Span) y Reino Unido (GrtB), podemos decir que España es mejor que Reino Unido en las distancias largas, pero sin embargo, Reino Unido es mejor en las distancias cortas.

Por último, podemos destacar a Samoa, que es un país que es malo en todo, sobre todo en las distancias cortas.

Modelo de regresión lineal múltiple

Para comenzar con el análisis del modelo de regresión múltiple, primero es necesario destacar a que este es idéntico al modelo de regresión lineal simple, la única diferencia es que en el modelo de regresión múltiple aparecen más variables explicativas.

El modelo de regresión múltiple viene dado por la siguiente ecuación (conocida como hiperplano):

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Donde:

- ε es el error.
- β_0 es la media.
- B_k k=1…p: son los coeficientes de regresión.

Como en el caso de la regresión simple, con la regresión múltiple también se desea encontrar aquella ecuación que más se ajuste a los datos. Los coeficientes serán elegidos de forma que la varianza residual sea mínima.

Siguiendo con nuestro ejemplo, hemos considerado que la variable de los 5000m(s) será la variable dependiente, ya que es la que mayor correlación tiene con el resto de variables (coeficiente de correlación múltiple al cuadrado). Por lo tanto el resto de variables se utilizaran como variables explicativas.

Uno de los métodos tradicionales aprendidos en esta asignatura es la de suponer un plano de regresión con todas las variables e ir eliminando (aquellas variables cuyo p-valor sea el mayor) y comprobando simultáneamente los valores de la R-cuadrado, para ver cuáles son las variables con las que construiremos nuestra plano.

Para estimar los parámetros B_0, B_1,…, B_P usaremos la función lm(), después utilizaremos la función summary(), para visualizar la información obtenida.

```
Call:
lm(formula = data8$`5000m(s)` ~ `10000m(s)` + `1500m(s)`, data = data8
[,
    -1])
```

```
Residuals:
      Min        1Q    Median        3Q       Max
-0.242797 -0.064828 -0.001089  0.071341  0.241460

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07552    0.39667  -0.190  0.84976
`10000m(s)`  0.39163    0.02064  18.974  < 2e-16 ***
`1500m(s)`   0.68923    0.22835   3.018  0.00396 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1093 on 51 degrees of freedom
Multiple R-squared:  0.9801,  Adjusted R-squared:  0.9794
F-statistic:  1259 on 2 and 51 DF,  p-value: < 2.2e-16
```

Por consiguiente, nuestra recta de regresión será el siguiente:

$$Y = -0.07552 + 0.39163\text{`10000m(s)'} + 0.68923\text{`1500m(s)'}$$

Como se puede observar, vemos que las variables de los 10000m(s) y 1500m(s) tienen un efecto sobre la variable de los 5000m(s). Teniendo el resto de variables un efecto nulo sobre esta.

Por otro lado, como los p-valor de las variables de los 10000m(s) y 1500m(s) son menores que 0.05, estas variables son estadísticamente significantes para la regresión lineal múltiple.

Además, si nos fijamos en la R-cuadrado (que nos indica cuanto se ajustan los datos a la recta de regresión) nos damos cuenta que tiene un valor del 98%, por lo tanto nuestros datos se ajustan muy bien a la recta de regresión propuesta.

Test de residuos

Breusch-Pagan test

Se utiliza para determinar la heterocedasticidad en un modelo de regresión lineal. Analiza si la varianza estimada de los residuos de una regresión depende de los valores de las variables independientes.

```
bptest(Step.for)

        studentized Breusch-Pagan test

data:  Step.for
BP = 0.52481, df = 2, p-value = 0.7692
```

Como se puede observar, hemos obtenido un valor de p bastante alto, por consiguiente, podemos decir que el modelo es heterocedastico. Al tratarse de un modelo heterocedastico, no es necesario realizar ninguna transformación para mejorar dicho modelo (por ejemplo Box-Cox)

Durbin-Watson test

Es una estadística de prueba que se utiliza para detectar la presencia de en los residuos (errores de predicción) de un análisis de la regresión.

```
dwtest(Step.for, alternative="two.sided")
```

```
        Durbin-Watson test

data:  Step.for
DW = 2.0267, p-value = 0.9592
alternative hypothesis: true autocorrelation is not 0
```

Predicción

Una vez que tenemos una recta o un plano, podemos hacer predicciones sobre cómo podrían ser los futuros datos, obviamente existe un error. En nuestro caso, le hemos dado un valor de 10 a nuestra variable de los 100000m(s) y un valor de 5 a nuestra variable de los 1500m(s), por lo tanto, obtendremos un valor de 7.28 en la variable de los 5000m(s).

```
> predict(datos.lm, newdata, interval="confidence")
       fit      lwr      upr
1 7.286909 5.935383 8.638435
```

Además de mostrarnos el valor exacto, también nos avisa sobre un intervalo de error de (5.93,8.64), con una cota de error de 1.35.