COMPLEMENTOS DE FORMACIÓN EN TÉCNICAS DE MINERÍA DE DATOS

CAIO FERNANDES MORENO

UCM - Mineria de Datos

ANÁLISIS DE CLUSTERS

# 1. Los datos.

Los datos utilizados en este trabajo son datos datos de algunos paises.

Para importar los datos se ha utilizado el codigo sas abajo:

**PROC** **IMPORT** OUT= ucm.paises

DATAFILE= "C:\Users\win\Documents\GitHub\ucm\complementos\trabajocomplementos31enero16\DatosPaises-SAS.xls"

DBMS=EXCEL5 REPLACE;

GETNAMES=YES;

**RUN**;

|  |
| --- |
| The SAS System |

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Afganistán | 27963 | 35.6 | 59.8 | 8.6 | -4766 | 566 | 157.13532 |
| **2** | Albania | 2902 | 13.1 | 77.5 | 7.2 | -2861 | 3786 | 577.685262 |
| **3** | Alemania | 80435 | 8.3 | 80.7 | 10.8 | 205408 | 41100 | 2659.28619 |
| **4** | Angola | 21220 | 46.2 | 51.7 | 14.2 | 29864 | 4221 | 19.5415273 |
| **5** | Arabia Saudita | 28091 | 20.8 | 74.1 | 3.4 | 144283 | 19327 | 10.288944 |
| **6** | Argelia | 36036 | 25.1 | 74.4 | 5.1 | 17558 | 4350 | 113.11257 |
| **7** | Argentina | 41223 | 17.8 | 76 | 7.6 | 12057 | 11508 | 310.531847 |
| **8** | Armenia | 2963 | 13.3 | 74.6 | 9 | -2771 | 3125 | 191.307784 |
| **9** | Australia | 22163 | 13.5 | 82.1 | 6.7 | 10724 | 57593 | 82.6334118 |
| **10** | Austria | 8392 | 9.5 | 81.1 | 9.4 | -5712 | 46377 | 1593.5443 |

# 2. Ejecución - Clusters

En el enunciado del ejercicio se pide para cargar el archivo y despues trabajar con una muestra de 100 paises utilizando el procedimiento en SAS llamado **proc surveyselect**, el codigo abajo se ejecuta lo que se ha pedido:

/\*

El archivo \DatosPaises.xlsx" (que podeis descargar del campus) contiene informacion sobre 7 variables socioeconomicas de 133 paises. Seleccionar aleatoriamente una muestra de 100 paises con el procedimiento surveyselect de la siguiente forma:

\*/

libname ucm 'C:\Users\win\Documents\GitHub\ucm\complementos\trabajocomplementos31enero16\';

**data** paises;

set ucm.paises;

**run**;

**proc** **print** data=paises (obs=**100**);

**run**;

**proc** **contents** data=paises out=sa;

**data**;set sa;if \_n\_=**1** then put 'LISTA DE VARIABLES CONTINUAS';if type=**1** then put name @@;**run**;

**data**;set sa;if \_n\_=**1** then put 'LISTA DE VARIABLES CATEGÓRICAS';if type=**2** then put name @@;**run**;

**proc** **surveyselect** data=paises method=srs n=**100** out=sample\_paises;

**run**;

Tambien se pide para trabajar solo con las variables POBL, NATALIDA, ESPERANZ e MORTALID.

/\*

Para la muestra obtenida, realizar un Analisis Cluster incluyendo SOLO

las variables demograficas (Pobl Natalidad EsperanzaVida Mortalidad

), que debe incluir como minimo:

\*/

**proc** **print** data=sample\_paises (obs=**100**);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Por las variables no teneren la misma unidad de medida es necesario normalizar los datos. Esto se hace en SAS utilizando el procedimiento **proc stdize** para normalizar las variables POBL, NATALIDA, ESPERANZ e MORTALID.

Codigo SAS:

**proc** **stdize** data=sample\_paises out=sample\_paisesnorm;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=sample\_paisesnorm;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=sample\_paisesnorm;

**run**;

Otra etapa que se ha hecho ha sido estudiar las correlaciones entre las 4 variables (POBL, NATALIDA, ESPERANZ e MORTALID), pero en esta etapa se ha utilizado la herramienta R y SAS.

En el resultado del estudio se ha encontrado una correlacion muy alta entre EsperanzaVida y Natalidad de **-0.87063840**, se percibe que cuanto mayor es la Esperanza de Vida menos Natalidad hay en un pais. En R he utilizado las 133 observaciones, pero con SAS solo estudiamos las correlaciones de una amuestra de 100 observaciones.

Abajo la tabla de correlacion hecha en R y una figura para representar la correlacion:

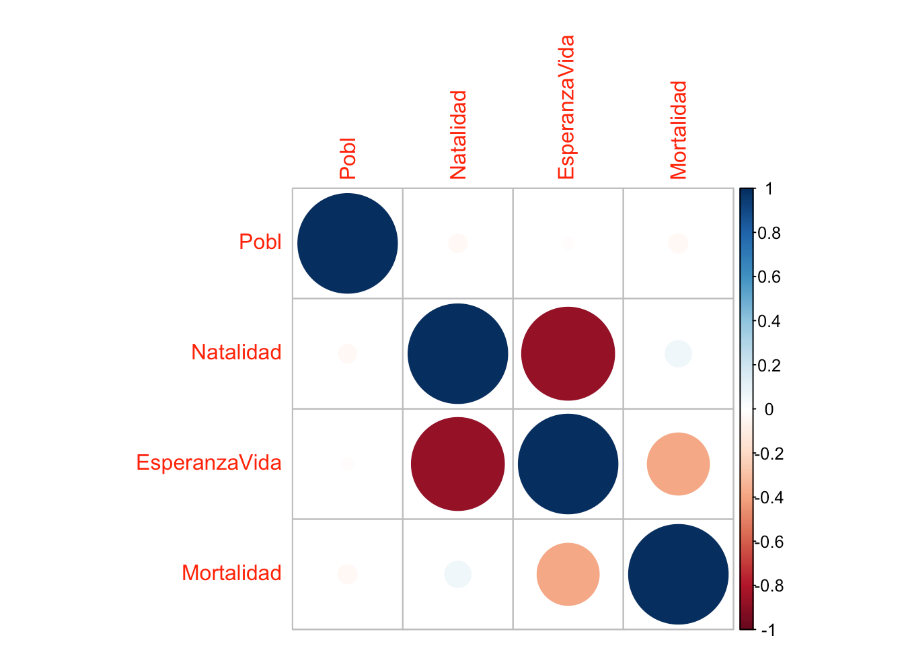
## Pobl Natalidad EsperanzaVida Mortalidad

## Pobl 1.00000000 -0.03243038 -0.01445153 -0.03519014

## Natalidad -0.03243038 1.00000000 -0.87063840 0.06833273

## EsperanzaVida -0.01445153 -0.87063840 1.00000000 -0.38664292

## Mortalidad -0.03519014 0.06833273 -0.38664292 1.00000000

****

En SAS para mirar la correlacion se hace con con el codigo abajo, con esto se analisa la correlacion y se puede ver tambien si hay datos atipicos o no.

**proc** **corr** data=sample\_paises outp=sample\_paisescorr;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados:

The CORR Procedure

|  |  |
| --- | --- |
| **4 Variables:** | POBL NATALIDA ESPERANZ MORTALID |

| **Simple Statistics** | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** | **Label** |
| **POBL** | 100 | 27906 | 39987 | 2790573 | 622.00000 | 198615 | POBL |
| **NATALIDA** | 100 | 22.21200 | 11.28550 | 2221 | 8.30000 | 49.80000 | NATALIDAD |
| **ESPERANZ** | 100 | 70.68200 | 9.00345 | 7068 | 49.50000 | 83.30000 | ESPERANZAVIDA |
| **MORTALID** | 100 | 8.50600 | 2.84679 | 850.60000 | 2.70000 | 15.30000 | MORTALIDAD |

Se puede ver en las cuatro variables que hay datos atipicos en casos donde son muy altos y muy bajos.

Abajo se puede ver la correlacion hecha en SAS, el resultado es lo mismo con R con una pequena diferencia que con SAS tenemos una amuestra de 100 observaciones y con R he hecho con las 133 observaciones. Pero la correlacion es con Natalidad y Esperanza de Vida.

| **Pearson Correlation Coefficients, N = 100  Prob > |r| under H0: Rho=0** | | | | |
| --- | --- | --- | --- | --- |
|  | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** |
| |  | | --- | | **POBL** | | POBL | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.04365 | | 0.6663 | | |  | | --- | | 0.00360 | | 0.9716 | | |  | | --- | | 0.00055 | | 0.9957 | |
| |  | | --- | | **NATALIDA** | | NATALIDAD | | |  | | --- | | -0.04365 | | 0.6663 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.88465 | | <.0001 | | |  | | --- | | 0.09054 | | 0.3703 | |
| |  | | --- | | **ESPERANZ** | | ESPERANZAVIDA | | |  | | --- | | 0.00360 | | 0.9716 | | |  | | --- | | -0.88465 | | <.0001 | | |  | | --- | | 1.00000 | |  | | |  | | --- | | -0.37909 | | 0.0001 | |
| |  | | --- | | **MORTALID** | | MORTALIDAD | | |  | | --- | | 0.00055 | | 0.9957 | | |  | | --- | | 0.09054 | | 0.3703 | | |  | | --- | | -0.37909 | | 0.0001 | | |  | | --- | | 1.00000 | |  | |

1. Analisis Cluster jerarquico con al menos dos metodos de agrupamiento. A partir de los procedimientos estudiados en clase, determinar el numero (o numeros) adecuado de grupos.

*“El análisis cluster (o de conglomerados) tiene como objetivo formar grupos de individuos con características similares con respecto a determinadas variables.”*

*“La idea básica es, a partir de un conjunto de individuos, crear grupos exluyentes y exhaustivos tales que:*

* *Los individuos de cada grupo deben ser lo más parecidos que sea posible (homogeneidad interna).*
* *Los grupos deben ser lo más diferentes que sea posible (heterogeneidad entre grupos).”*

Para hacer una analisis de cluster es necesario tres fases:

* Normalizar las variables;
* Ver la correlación entre las variables;
* Corregir el problema de los atípicos ya que distorsionarían la generación de cluster;

**Analisis de clusteres jerárquico**

Importante: No hace falta hacer la analise factorial porque son pocas variables, he preguntado para la profesora Aida y ella me ha contestado que en este ejemplo no hace falta.

Se ha utilizado los codigos SAS abajo para probar 2 metodos de agrupamiento (centroid y ward) que son los 2 mas utilizados.

El procedimento SAS para analise de cluster jerarquico es el proc cluster, se ha anadido los parametros pseudo, RSQUARE para que en los resultados se muestre la pseudo-T2 y la pseudo-F y el R2.

Metodo de centroide.

**proc** **cluster** data=sample\_paisesnorm method=centroid pseudo ccc RSQUARE

outtree=sample\_paisesnormC

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados con el metodo centroide:

The CLUSTER Procedure

Centroid Hierarchical Cluster Analysis

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 1.96555872 | 0.94160487 | 0.4914 | 0.4914 |
| **2** | 1.02395385 | 0.08070507 | 0.2560 | 0.7474 |
| **3** | 0.94324878 | 0.87601013 | 0.2358 | 0.9832 |
| **4** | 0.06723865 |  | 0.0168 | 1.0000 |

|  |  |
| --- | --- |
| **Root-Mean-Square Total-Sample Standard Deviation** | 1 |

|  |  |
| --- | --- |
| **Root-Mean-Square Distance Between Observations** | 2.828427 |

| **Cluster History** | | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Approximate Expected R-Square** | **Cubic Clustering Criterion** | **Pseudo F Statistic** | **Pseudo t-Squared** | **Norm Centroid Distance** | **Tie** |
| **10** | **CL21** | **OB57** | 3 | 0.0036 | .824 | .821 | 0.26 | 46.8 | 2.4 | 0.515 |  |
| **9** | **CL10** | **CL54** | 6 | 0.0087 | .815 | .805 | 0.90 | 50.2 | 6.2 | 0.5347 |  |
| **8** | **CL16** | **CL11** | 53 | 0.0754 | .740 | .785 | -3.1 | 37.4 | 38.3 | 0.5386 |  |
| **7** | **CL13** | **OB12** | 3 | 0.0039 | .736 | .761 | -1.7 | 43.2 | 1.7 | 0.5409 |  |
| **6** | **CL8** | **CL9** | 59 | 0.0479 | .688 | .731 | -2.6 | 41.5 | 14.4 | 0.6635 |  |
| **5** | **CL6** | **CL15** | 66 | 0.0934 | .595 | .691 | -5.2 | 34.8 | 24.3 | 0.8594 |  |
| **4** | **CL12** | **CL14** | 30 | 0.0418 | .553 | .636 | -4.3 | 39.6 | 19.7 | 0.8751 |  |
| **3** | **CL4** | **CL5** | 96 | 0.3815 | .171 | .529 | -12 | 10.0 | 81.3 | 0.9568 |  |
| **2** | **CL7** | **OB29** | 4 | 0.0163 | .155 | .379 | -6.9 | 18.0 | 5.2 | 1.0385 |  |
| **1** | **CL3** | **CL2** | 100 | 0.1551 | .000 | .000 | 0.00 | . | 18.0 | 1.4141 |  |

Para determinar con cuantos clusters se debe quedar hay que mirar el máximo relativo de Pseudo F-Statistic, Pseudo t-Squared y el R-Squared mas grande que 70%.

Por mirar el Pseudo t-Squared se ve el valor 81,3 que es el máximo relativo, por esto me quedo con 4 clusters, pues el 81,3 es en el cluster 3 y es necesario añadir más 1, el resultado es 4.

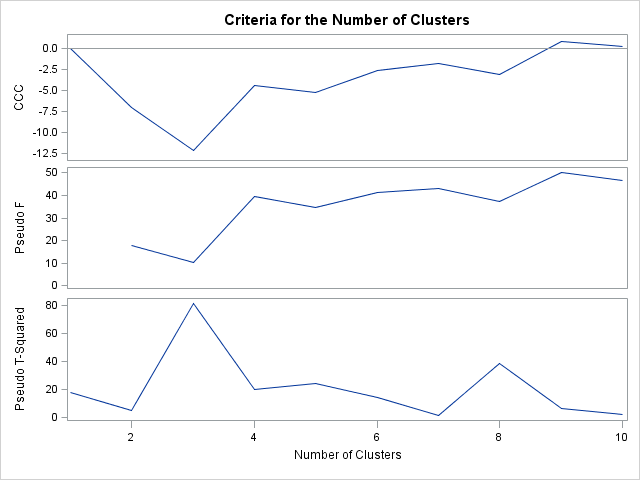
En el método Pseudo F-Statistic donde se elige el cluster por el máximo relativo entonces nos quedamos con 4.

Pero utilizando el R-Squared, dado que es mayor que 70% nos quedaríamos con 7 clusters.

Para el criterio ccc (**Cubic Clustering Criterion)** se puede ver valores negativos, esto es un indicador que hay dados atípicos. Valores de CCC más grandes de 2 a 3 indican buenos clusters.

He encontrado en el enlace abajo más detalles de como se puede interpretar el CCC.

<https://support.sas.com/documentation/onlinedoc/v82/techreport_a108.pdf>

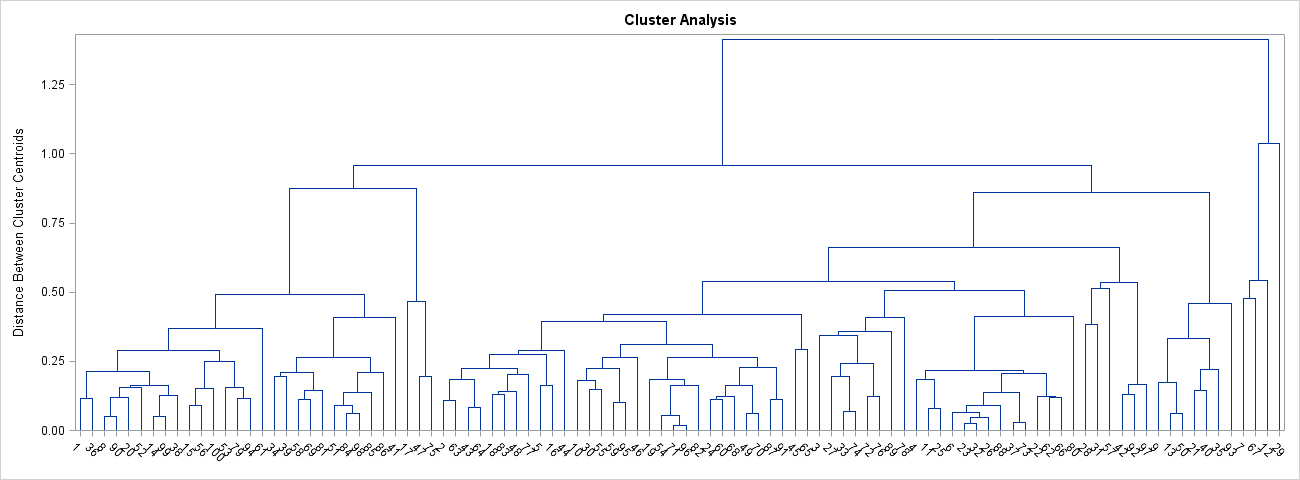


En los tres gráficos arriba se puede ver que lo mejor es quedar con 4 clusters.

Porque el Pseudo T-Squared tiene un pico en 3 clusters y se debe añadir más 1, entonces quedamos con 4.

Lo mismo con el Pseudo F que se puede concluir que se debe quedar con 4.

Dendograma del método de centroide.



Ahora probaremos con el Ward.

Metodo de ward.

**proc** **cluster** data=sample\_paisesnorm method=ward pseudo ccc RSQUARE

outtree=sample\_paisesnormW

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Los resultados son:

The CLUSTER Procedure

Ward's Minimum Variance Cluster Analysis

| **Eigenvalues of the Covariance Matrix** | | | | |
| --- | --- | --- | --- | --- |
|  | **Eigenvalue** | **Difference** | **Proportion** | **Cumulative** |
| **1** | 1.96555872 | 0.94160487 | 0.4914 | 0.4914 |
| **2** | 1.02395385 | 0.08070507 | 0.2560 | 0.7474 |
| **3** | 0.94324878 | 0.87601013 | 0.2358 | 0.9832 |
| **4** | 0.06723865 |  | 0.0168 | 1.0000 |

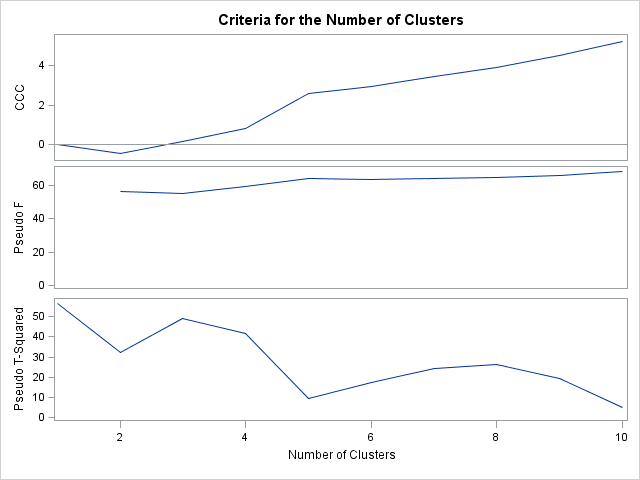
|  |  |
| --- | --- |
| **Root-Mean-Square Total-Sample Standard Deviation** | 1 |

|  |  |
| --- | --- |
| **Root-Mean-Square Distance Between Observations** | 2.828427 |

| **Cluster History** | | | | | | | | | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Number of Clusters** | **Clusters Joined** | | **Freq** | **Semipartial R-Square** | **R-Square** | **Approximate Expected R-Square** | **Cubic Clustering Criterion** | **Pseudo F Statistic** | **Pseudo t-Squared** | **Tie** |
| **10** | **CL19** | **OB29** | 4 | 0.0163 | .872 | .821 | 5.22 | 68.2 | 5.2 |  |
| **9** | **CL15** | **CL13** | 28 | 0.0194 | .853 | .805 | 4.50 | 65.9 | 19.2 |  |
| **8** | **CL22** | **CL20** | 18 | 0.0220 | .831 | .785 | 3.92 | 64.5 | 26.1 |  |
| **7** | **CL14** | **CL24** | 19 | 0.0258 | .805 | .761 | 3.45 | 64.0 | 24.3 |  |
| **6** | **CL8** | **CL18** | 25 | 0.0334 | .772 | .731 | 2.93 | 63.5 | 17.6 |  |
| **5** | **CL10** | **CL12** | 11 | 0.0414 | .730 | .691 | 2.57 | 64.3 | 9.4 |  |
| **4** | **CL7** | **CL11** | 36 | 0.0802 | .650 | .636 | 0.81 | 59.4 | 41.4 |  |
| **3** | **CL9** | **CL6** | 53 | 0.1177 | .532 | .529 | 0.16 | 55.2 | 48.9 |  |
| **2** | **CL3** | **CL5** | 64 | 0.1664 | .366 | .379 | -.46 | 56.5 | 32.1 |  |
| **1** | **CL4** | **CL2** | 100 | 0.3659 | .000 | .000 | 0.00 | . | 56.5 |  |

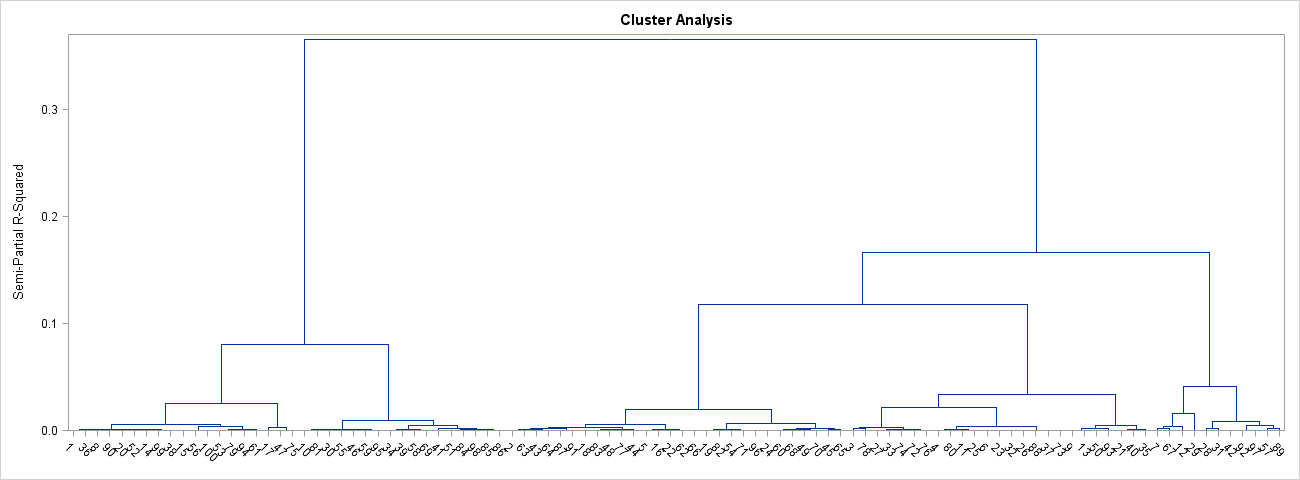
Por el Pseudo t-Squared se debe quedar con 4 clusters. Por el R-Square con 5.

Con el CCC quedo con 5 clusteres porque esta entre 2 y 3, el valor es 2.57.



Con el grafico arriba lo mismo 4 para el Pseudo T-Squared y 5 para el Pseudo F y CCC.

El dendograma con el metodo de Ward.



Dejo abajo el codigo para probar con el metodo average si se desea, pero no lo que utilizado.

**proc** **cluster** data=sample\_paisesnorm method=average pseudo RSQUARE ccc outtree=sample\_paisesnormA

print=**10** plots=den(VERTICAL);

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

**Analisis de clusteres no jerárquico**

Abajo se puede ver los codigos SAS para analisis de clusters no jerárquico, donde se ha probado con grupos (clusters) de 4, 5 y 7, esto por que han sido las cantidades de grupos (clusters) que han salido mejor en la prueba de analisis de clusteres jerarquicos pero no se puede llegar a una conclusion exacta de la mejor cuantidad para se quedar.

Probando ambos jerárquico y no jerárquico la conclusión es que el número de grupos (clusters) que se debe quedar es 4.

Abajo se explicará como se ha llegado a esta conclusion, bien como los codigos y valores analisados.

**Prueba con grupos (clusteres) no jerárquico para 4 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**4** MEAN=MEDIAS2

DRIFT OUT=cluster4 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Resultados para la prueba con grupos (clusteres) no jerárquico para 4 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 68.64 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.48331 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 14.195 |

Se ha puesto en amarillo los valores de Pseudo F Statistic y Cubic Clustering Criterion por que con 4 grupos (clusters) se ha obtenido los mejores valores comparados con 5 y 7 grupos.

**Prueba con grupos (clusteres) no jerárquico para 5 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**5** MEAN=MEDIAS2

DRIFT OUT=cluster5 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Prueba con grupos (clusteres) no jerárquico para 5 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 58.69 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.58643 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 9.680 |

**Prueba con grupos (clusteres) no jerárquico para 7 grupos.**

**PROC** **FASTCLUS** DATA=sample\_paisesnorm MAXCLUSTERS=**7** MEAN=MEDIAS2

DRIFT OUT=cluster7 maxiter=**30**;

var POBL NATALIDA ESPERANZ MORTALID;

**run**;

Prueba con grupos (clusteres) no jerárquico para 7 grupos.

|  |  |
| --- | --- |
| **Pseudo F Statistic =** | 59.70 |

|  |  |
| --- | --- |
| **Approximate Expected Over-All R-Squared =** | 0.66546 |

|  |  |
| --- | --- |
| **Cubic Clustering Criterion =** | 11.147 |

Hasta el momento la conclusion es quedar con 4 grupos (clusters) porque para comparar los grupos (clusteres) hay que comparar los resultados de los valores Pseudo F Statistic y también el Cubic Clustering Criterion. Se sabe que cuanto más grande son el Pseudo F Statistic y el CCC (Cubic Clustering Criterion) mejor es el cluster.

Entonces como se puede ver en los resultados arriba los mas altos son del cluster 4.

**Ahora probaremos el teste de beale (contraste F de Beale):**

El codigo SAS abajo se puede probar el test de beale con grupos (clusters) de 4, 5 y 7.

**proc** **means** data=cluster4 ; var distance; output out=sumacuad4 uss=w4 ;

**run**;

**proc** **means** data=cluster5 ; var distance; output out=sumacuad5 uss=w5 ;

**run**;

**proc** **means** data=cluster7 ; var distance; output out=sumacuad7 uss=w7 ;

**run**;

**data** beale;

merge sumacuad4 sumacuad5 sumacuad7;

k1=(\_freq\_-**4**)\*(**4**\*\*(-**2**/**8**));

k2=(\_freq\_-**5**)\*(**5**\*\*(-**2**/**8**));

k3=(\_freq\_-**7**)\*(**7**\*\*(-**2**/**8**));

fbeale1=(w4-w5)\*k2/(w5\*(k1-k2));

pvalor=**1**-probf(fbeale1,(k1-k2),k2);

fbeale2=(w4-w7)\*k3/(w7\*(k1-k3));

pvalor2=**1**-probf(fbeale2,(k1-k3),k3);

fbeale3=(w5-w7)\*k3/(w7\*(k2-k3));

pvalor3=**1**-probf(fbeale3,(k2-k3),k3);

**run**;

**proc** **print** data=beale;**run**;

Resultados del teste de Beale (contraste F de Beale):

| **Obs** | **\_TYPE\_** | **\_FREQ\_** | **w4** | **w5** | **w7** | **k1** | **k2** | **k3** | **fbeale1** | **pvalor** | **fbeale2** | **pvalor2** | **fbeale3** | **pvalor3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 0 | 100 | 125.909 | 114.090 | 81.6182 | 67.8823 | 63.5303 | 57.1753 | 1.51231 | 0.20544 | 2.89781 | .004545197 | 3.57938 | .003807254 |

Ahora haremos la interpretacion del resultado del teste de beale (contraste F de Beale).

**Comparativos de p-valores.**

fbeale1=(w4-w5)\*k2/(w5\*(k1-k2));

pvalor=**1**-probf(fbeale1,(k1-k2),k2);

fbeale2=(w4-w7)\*k3/(w7\*(k1-k3));

pvalor2=**1**-probf(fbeale2,(k1-k3),k3);

fbeale3=(w5-w7)\*k3/(w7\*(k2-k3));

pvalor3=**1**-probf(fbeale3,(k2-k3),k3);

**Valores de p-valores:**

Pvalor1 = 0.20544

Pvalor2 = .004545197

Pvalor3 = .003807254

El p-valor 1 (comparación de 4 clusters con 5) es 0.20544 es alto, entonces la comparación de 4 clusters con 5 clusters se dice que es muy difícil que 4 clusters sea peor que 5 clusters.

El p-valor2 es bajo siendo .004545197 donde se hace la comparación de 4 frente a 7, se entiende que si 4 clusters son mejores que 7.

El p-valor3 es bajo donde se compara 5 frente a 7, donde si 5 clusters son mejores que 7.

**Conclusión Final**

La conclusión final de la cantidad de grupos (clusters) baseado en el resultado de todos las pruebas es quedar con 4 grupos (clusters).

**Grupos (Clusters)**

El código SAS abajo se puede utilizar para estudiar los grupos.

**proc** **sort** data=cluster4 out=cluster4s;

by cluster;

**proc** **Freq** data=cluster4s;

by cluster; tables PAIS;

**run**;

**proc** **sort** data=cluster5 out=cluster5s;

by cluster;

**proc** **Freq** data=cluster5s;

by cluster; tables PAIS;

**run**;

**proc** **sort** data=cluster7 out=cluster7s;

by cluster;

**proc** **Freq** data=cluster7s;

by cluster; tables PAIS;

**run**;

**Pruebas con R**

He hecho algunas pruebas con R.

En el ejemplo abajo se utiliza el paquete NbClust para determinar el número correcto de grupos (clusters).

Enlace para el tutorial:

<http://www.inside-r.org/packages/cran/NbClust/docs/NbClust>

**Código R:**

setwd("/Users/caiomsouza/git/Bitbucket/ucm/COMPLEMENTOS\_DE\_FORMACION\_EN\_TECNICAS\_DE\_MINERIA\_DE\_DATOS/tareas-entregar/trabajo-31enero16")

paises <- read.csv(file="DatosPaises.csv",head=TRUE,sep=",")

#head(paises, 10)

#Dejar solo POBL NATALIDA ESPERANZ MORTALID

paises.valores <- paises

# Remove la columna Paises

paises.valores$Pais <- NULL

# Remove la columna BalanzaComercial

paises.valores$BalanzaComercial <- NULL

# Remove la columna PIB

paises.valores$PIB <- NULL

# Remove la columna ProdCereales

paises.valores$ProdCereales <- NULL

#head(paises.valores,10)

# Normaliza las variables

paises.valores.normalizar <- scale(paises.valores)

#head(paises.valores.normalizar, 10)

## Prueba el mejor cluster

data1 <- paises.valores

data2 <- paises.valores.normalizar

#data<-iris[,-c(5)]

#data <- data1

data <- data2

res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,

method = "ward.D2", index = "kl")

res$All.index

res$Best.nc

res$Best.partition

res<-NbClust(data, diss=NULL, distance = "euclidean", min.nc=2, max.nc=6,

method = "kmeans", index = "hubert")

res$All.index

res<-NbClust(data, diss=NULL, distance = "manhattan", min.nc=2, max.nc=6,

method = "complete", index = "all")

res$All.index

res$Best.nc

res$All.CriticalValues

res$Best.partition

He probado los datos normalizados y no normalizados por curiosidad, pero si lo sé que por las unidades de medidas no tener las mismas es obligatorio normalizar los datos.

Con el método ward se recomienda 6 grupos/clusters, pero con kmeans y complete se recomienda 2 con los datos normalizados.

Cuando los datos no son normalizados se recomienda 5 grupos con ward y después con los otros métodos 2.

La conclusión es que el mejor número de grupos son 2.

Pero los valores son distintos cuando los valores están normalizados o no.

Con los datos normalizados tenemos las siguientes recomendaciones para el numero de clusters(grupos):

PseudoT2: 2

Test de Beale: 2

CCC: 2

Silhouette: 2

**Conclusiones con R:**

Se puede concluir que con R el número de grupos es de 2 a 6 grupos, pero se recomienda 2 grupos, lo que hemos visto con SAS que no es la mejor recomendación.

No se puede decir que R se equivoca porque no conozco este paquete **NbClust** o suficiente para quizás hacer un trabajo con la calidad exigida.

Quizás con más tiempo y conocimiento de R y del paquete **NbClust** se pudiera llegar a números y resultados más próximos de los conseguidos con SAS.

Pero si se puede decir que con R se recomienda de 2 a 6. Dejo esta análisis con R solo como un complemento al trabajo, pero tengo conocimiento que no ha sido pedido hacer nada con R, solo con SAS y con lo que hemos visto en clase.

**Explicando los grupos (Clusters)**

El código SAS abajo se imprime los grupos para que se pueda interpretar sus resultados y características de cada grupo.

Lo analizáremos solo los resultados con 4 grupos (clusters).

**proc** **sort** data=cluster4 out=cluster4s;

by cluster;

**proc** **Freq** data=cluster4s;

by cluster; tables PAIS;

**run**;

**proc** **print** data=cluster4;**run**;

Abajo los países que pertenecen a cada grupo (cluster).

**Grupo 1 (Cluster 1)**

Cluster=1

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Afganistán** | 1 | 3.45 | 1 | 3.45 |
| **Benin** | 1 | 3.45 | 2 | 6.90 |
| **Burkina Faso** | 1 | 3.45 | 3 | 10.34 |
| **Camerún** | 1 | 3.45 | 4 | 13.79 |
| **Chad** | 1 | 3.45 | 5 | 17.24 |
| **Congo** | 1 | 3.45 | 6 | 20.69 |
| **Etiopía** | 1 | 3.45 | 7 | 24.14 |
| **Gabón** | 1 | 3.45 | 8 | 27.59 |
| **Ghana** | 1 | 3.45 | 9 | 31.03 |
| **Guinea** | 1 | 3.45 | 10 | 34.48 |
| **Haití** | 1 | 3.45 | 11 | 37.93 |
| **Lesoto** | 1 | 3.45 | 12 | 41.38 |
| **Madagascar** | 1 | 3.45 | 13 | 44.83 |
| **Malawi** | 1 | 3.45 | 14 | 48.28 |
| **Malí** | 1 | 3.45 | 15 | 51.72 |
| **Mozambique** | 1 | 3.45 | 16 | 55.17 |
| **Namibia** | 1 | 3.45 | 17 | 58.62 |
| **Níger** | 1 | 3.45 | 18 | 62.07 |
| **Papúa Nueva Guinea** | 1 | 3.45 | 19 | 65.52 |
| **Republica Checa** | 1 | 3.45 | 20 | 68.97 |
| **República Democrática Popular de Corea** | 1 | 3.45 | 21 | 72.41 |
| **Rumania** | 1 | 3.45 | 22 | 75.86 |
| **Senegal** | 1 | 3.45 | 23 | 79.31 |
| **Serbia** | 1 | 3.45 | 24 | 82.76 |
| **Togo** | 1 | 3.45 | 25 | 86.21 |
| **Uganda** | 1 | 3.45 | 26 | 89.66 |
| **Yemen** | 1 | 3.45 | 27 | 93.10 |
| **Zambia** | 1 | 3.45 | 28 | 96.55 |
| **Zimbabue** | 1 | 3.45 | 29 | 100.00 |

**Grupo 2 (Cluster 2)**

The FREQ Procedure

Cluster=2

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Bangladesh** | 1 | 20.00 | 1 | 20.00 |
| **Brasil** | 1 | 20.00 | 2 | 40.00 |
| **Federación Rusa** | 1 | 20.00 | 3 | 60.00 |
| **Filipinas** | 1 | 20.00 | 4 | 80.00 |
| **Pakistán** | 1 | 20.00 | 5 | 100.00 |

**Grupo 3 (Cluster 3)**

The FREQ Procedure

Cluster=3

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Albania** | 1 | 2.56 | 1 | 2.56 |
| **Australia** | 1 | 2.56 | 2 | 5.13 |
| **Bolivia (Estado Plurinacional de)** | 1 | 2.56 | 3 | 7.69 |
| **Canadá** | 1 | 2.56 | 4 | 10.26 |
| **Chile** | 1 | 2.56 | 5 | 12.82 |
| **Colombia** | 1 | 2.56 | 6 | 15.38 |
| **Cuba** | 1 | 2.56 | 7 | 17.95 |
| **Ecuador** | 1 | 2.56 | 8 | 20.51 |
| **Fiji** | 1 | 2.56 | 9 | 23.08 |
| **Irak** | 1 | 2.56 | 10 | 25.64 |
| **Irlanda** | 1 | 2.56 | 11 | 28.21 |
| **Irán (República islámica de)** | 1 | 2.56 | 12 | 30.77 |
| **Israel** | 1 | 2.56 | 13 | 33.33 |
| **Jordania** | 1 | 2.56 | 14 | 35.90 |
| **Kazajstán** | 1 | 2.56 | 15 | 38.46 |
| **Libia** | 1 | 2.56 | 16 | 41.03 |
| **Líbano** | 1 | 2.56 | 17 | 43.59 |
| **Marruecos** | 1 | 2.56 | 18 | 46.15 |
| **Mongolia** | 1 | 2.56 | 19 | 48.72 |
| **Myanmar** | 1 | 2.56 | 20 | 51.28 |
| **Nepal** | 1 | 2.56 | 21 | 53.85 |
| **Nicaragua** | 1 | 2.56 | 22 | 56.41 |
| **Nueva Caledonia** | 1 | 2.56 | 23 | 58.97 |
| **Nueva Zelanda** | 1 | 2.56 | 24 | 61.54 |
| **Omán** | 1 | 2.56 | 25 | 64.10 |
| **Panamá** | 1 | 2.56 | 26 | 66.67 |
| **Paraguay** | 1 | 2.56 | 27 | 69.23 |
| **Perú** | 1 | 2.56 | 28 | 71.79 |
| **República Dominicana** | 1 | 2.56 | 29 | 74.36 |
| **República Unida de Tanzanía** | 1 | 2.56 | 30 | 76.92 |
| **República de Moldova** | 1 | 2.56 | 31 | 79.49 |
| **Ruanda** | 1 | 2.56 | 32 | 82.05 |
| **Sri Lanka** | 1 | 2.56 | 33 | 84.62 |
| **Tailandia** | 1 | 2.56 | 34 | 87.18 |
| **Turquía** | 1 | 2.56 | 35 | 89.74 |
| **Túnez** | 1 | 2.56 | 36 | 92.31 |
| **Uzbekistán** | 1 | 2.56 | 37 | 94.87 |
| **Venezuela (República Bolivariana de)** | 1 | 2.56 | 38 | 97.44 |
| **Viet Nam** | 1 | 2.56 | 39 | 100.00 |

**Grupo 4 (Cluster 4)**

The FREQ Procedure

Cluster=4

| **PAIS** | | | | |
| --- | --- | --- | --- | --- |
| **PAIS** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **Alemania** | 1 | 3.70 | 1 | 3.70 |
| **Armenia** | 1 | 3.70 | 2 | 7.41 |
| **Austria** | 1 | 3.70 | 3 | 11.11 |
| **Bielorrusia** | 1 | 3.70 | 4 | 14.81 |
| **Bosnia y Herzegovina** | 1 | 3.70 | 5 | 18.52 |
| **Bulgaria** | 1 | 3.70 | 6 | 22.22 |
| **Croacia** | 1 | 3.70 | 7 | 25.93 |
| **Dinamarca** | 1 | 3.70 | 8 | 29.63 |
| **Eslovaquia** | 1 | 3.70 | 9 | 33.33 |
| **Eslovenia** | 1 | 3.70 | 10 | 37.04 |
| **España** | 1 | 3.70 | 11 | 40.74 |
| **Finlandia** | 1 | 3.70 | 12 | 44.44 |
| **Francia** | 1 | 3.70 | 13 | 48.15 |
| **Georgia** | 1 | 3.70 | 14 | 51.85 |
| **Grecia** | 1 | 3.70 | 15 | 55.56 |
| **Hungría** | 1 | 3.70 | 16 | 59.26 |
| **Lituania** | 1 | 3.70 | 17 | 62.96 |
| **Noruega** | 1 | 3.70 | 18 | 66.67 |
| **Países Bajos** | 1 | 3.70 | 19 | 70.37 |
| **Polonia** | 1 | 3.70 | 20 | 74.07 |
| **Portugal** | 1 | 3.70 | 21 | 77.78 |
| **Reino Unido** | 1 | 3.70 | 22 | 81.48 |
| **República Democrática Popular Lao** | 1 | 3.70 | 23 | 85.19 |
| **República Democrática del Congo** | 1 | 3.70 | 24 | 88.89 |
| **República de Corea** | 1 | 3.70 | 25 | 92.59 |
| **Suecia** | 1 | 3.70 | 26 | 96.30 |
| **Ucrania** | 1 | 3.70 | 27 | 100.00 |

**Conclusiones de los 4 grupos.**

Es muy difícil agrupar países, interpretar los resultados es una tarea aún más difícil, llevando en consideración que la tentativa ha sido agrupar 100 países de una amuestra aleatoria de 133 países, por las variables: Población, Natalidad, Esperanza de Vida y Mortalidad.

**Grupo 1:**

Son 29% de la amuestra.

Países con **natalidad alta, esperanza de vida muy baja y alta mortalidad**.

**Grupo 2:**

Son 5% de la amuestra.

Es el grupo más pequeño solo con 5 países (Brasil, Federación Rusa, Pakistán, Filipinas y Bangladesh) siendo países con **gran población**.

**Grupo 3:**

Son 39% de la amuestra.

Creo que son los países **que no están ni el grupo 1, 2 y 4**. Quizás países que son parecidos, pero se puede ver países (marcados con color) con gran distancia del grupo.

**Grupo 4:**

Son 27% de la amuestra.

Es un grupo grande con países con **esperanza de vida alta, natalidad baja**.

Resultado final de los 100 paises agrupados en 4 grupos (clusters), las variables estan normalizadas, los datos estan ordenados por la columna cluster.

**proc** **sort** data=cluster4 out=cluster4s;

by cluster;

**proc** **print** data=cluster4s;

**run**;

Resultado de SAS:

|  |
| --- |
| The SAS System |

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **CLUSTER** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Afganistán | 0.0278959 | 1.22957146 | -1.2669588576 | 0.0475754 | -4766 | 566 | 157.13532 | 1 | 0.50611 |
| **2** | Benin | -0.481223 | 1.31949186 | -1.3361915821 | 0.4052853 | -1057 | 690 | 393.304324 | 1 | 0.21309 |
| **3** | Burkina Faso | -0.312298 | 1.69715751 | -1.4631182436 | 0.5483693 | -837 | 579 | 377.841176 | 1 | 0.41393 |
| **4** | Camerún | -0.175514 | 1.40042021 | -1.8323594407 | 1.2280182 | -952 | 1145 | 287.733333 | 1 | 0.95192 |
| **5** | Chad | -0.415342 | 2.15575152 | -2.2708333623 | 2.158064 | 903 | 736 | 43.048546 | 1 | 2.13877 |
| **6** | Congo | -0.631387 | 1.38243613 | -1.082338259 | 0.1906594 | 5213 | 2987 | 2.35377358 | 1 | 0.49654 |
| **7** | Etiopía | 1.6721449 | 1.01376252 | -0.8861788731 | -0.238593 | -6216 | 302 | 424.364723 | 1 | 2.12680 |
| **8** | Gabón | -0.701021 | 0.79795357 | -0.8169461486 | 0.1906594 | 5706 | 8278 | 7.90697674 | 1 | 0.87950 |
| **9** | Ghana | -0.072664 | 1.04073864 | -1.1284934087 | 0.2622014 | -3077 | 1326 | 185.14121 | 1 | 0.43431 |
| **10** | Guinea | -0.43973 | 1.40941225 | -1.474657031 | 0.6914533 | 66 | 481 | 190.483333 | 1 | 0.35217 |
| **11** | Haití | -0.467705 | 0.32137549 | -0.9784891724 | 0.1548884 | -2568 | 621 | 246.949153 | 1 | 1.11663 |
| **12** | Lesoto | -0.688082 | 0.62710483 | -2.4554539608 | 2.3011479 | -1404 | 1083 | 75.6450766 | 1 | 2.33178 |
| **13** | Madagascar | -0.161996 | 1.15763515 | -0.7246358493 | -0.560531 | -1464 | 415 | 124.711605 | 1 | 1.17805 |
| **14** | Malawi | -0.336107 | 1.58925304 | -1.1284934087 | 0.0475754 | -1044 | 464 | 629.517698 | 1 | 0.50912 |
| **15** | Malí | -0.325127 | 2.02087093 | -1.5669673303 | 0.9060792 | -1434 | 672 | 154.096516 | 1 | 0.87357 |
| **16** | Mozambique | -0.072582 | 1.6252212 | -1.8669758029 | 1.1922472 | -1600 | 424 | 50.1621622 | 1 | 1.00064 |
| **17** | Namibia | -0.683061 | 0.74400134 | -0.7477134242 | -0.417447 | -1276 | 5113 | 2.97611379 | 1 | 1.25097 |
| **18** | Níger | -0.29409 | 2.50644105 | -1.1631097709 | 0.4052853 | -1537 | 360 | 116.190233 | 1 | 1.18714 |
| **19** | Papúa Nueva Guinea | -0.554636 | 0.6450889 | -0.9784891724 | -0.274364 | 1792 | 1415 | 11.8907563 | 1 | 1.08138 |
| **20** | Republica Checa | -0.453717 | 1.11267495 | -2.4554539608 | 2.4084609 | 6420 | 19616 | 1629.30154 | 1 | 2.29393 |
| **21** | República Democrática Popular de Corea | -0.067616 | 1.85901422 | -1.4631182436 | 0.7987663 | -1075 | 570 | 1727.26236 | 1 | 0.68851 |
| **22** | Rumania | -0.183543 | 0.9867864 | -0.8861788731 | -0.345905 | -12528 | 7685 | 1216.98726 | 1 | 0.96139 |
| **23** | Senegal | -0.386098 | 1.16662719 | -0.5284764634 | -0.488989 | -2383 | 998 | 196.097837 | 1 | 1.23146 |
| **24** | Serbia | -0.493638 | 1.52630876 | -0.574631613 | -0.667844 | -6920 | 5412 | 1838.67445 | 1 | 1.36336 |
| **25** | Togo | -0.567244 | 1.29251574 | -1.3592691569 | 0.2979724 | -356 | 503 | 68.0041558 | 1 | 0.33247 |
| **26** | Uganda | 0.17097 | 1.95792665 | -1.5554285428 | 0.6199113 | -1594 | 636 | 226.798485 | 1 | 0.82660 |
| **27** | Yemen | -0.092694 | 1.01376252 | -0.8400237234 | -0.488989 | -1249 | 1358 | 43.0289707 | 1 | 1.10351 |
| **28** | Zambia | -0.359585 | 1.67917343 | -1.3823467317 | 0.4410563 | 1888 | 1225 | 129.955571 | 1 | 0.36717 |
| **29** | Zimbabue | -0.35804 | 1.27453166 | -1.8438982281 | 0.9418502 | -601 | 721 | 88.2230864 | 1 | 0.73819 |
| **30** | Bangladesh | 3.4393285 | -0.1372185 | 0.0253853323 | -1.061325 | -11877 | 758 | 5684.91934 | 2 | 0.88451 |
| **31** | Brasil | 4.7359345 | -0.6137966 | 0.38308774201 | -0.846699 | 10378 | 10978 | 272.824753 | 2 | 1.61724 |
| **32** | Federación Rusa | 3.2059571 | -0.8296056 | -0.1130801166 | 1.943438 | 168013 | 10618 | 278.162132 | 2 | 2.25278 |
| **33** | Filipinas | 1.8232473 | 0.1864949 | -0.32077829 | -0.632073 | -6992 | 2136 | 1786.91037 | 2 | 1.71051 |
| **34** | Pakistán | 3.9477025 | 0.70803318 | -0.5630928256 | -0.345905 | -16373 | 1008 | 1287.39645 | 2 | 1.09780 |
| **35** | Albania | -0.6635 | -0.7936374 | 0.77540651394 | -0.453218 | -2861 | 3786 | 577.685262 | 3 | 0.90394 |
| **36** | Australia | -0.132118 | -0.7576692 | 1.30619073479 | -0.632073 | 10724 | 57593 | 82.6334118 | 3 | 1.00514 |
| **37** | Bolivia (Estado Plurinacional de) | -0.469939 | 0.22246305 | -0.3553946522 | -0.310135 | 998 | 1935 | 49.2931894 | 3 | 1.14646 |
| **38** | Canadá | 0.1979516 | -0.9914623 | 1.27157437256 | -0.417447 | -4638 | 47297 | 694.911701 | 3 | 1.21972 |
| **39** | Chile | -0.274143 | -0.7576692 | 1.20234164811 | -1.204409 | 11068 | 12685 | 225.751787 | 3 | 0.96594 |
| **40** | Colombia | 0.523248 | -0.5148842 | 0.34847137978 | -0.954012 | -973 | 6180 | 95.7396875 | 3 | 0.71587 |
| **41** | Cuba | -0.431591 | -1.0274304 | 0.97156589991 | -0.274364 | -5975 | 5702 | 121.623946 | 3 | 1.12461 |
| **42** | Ecuador | -0.331555 | -0.0652822 | 0.54463076574 | -1.168638 | -3176 | 4637 | 363.654456 | 3 | 0.43166 |
| **43** | Fiji | -0.719836 | -0.1102424 | -0.124618904 | -0.596302 | -975 | 3649 | 20.2988235 | 3 | 0.87932 |
| **44** | Irak | 0.1080406 | 1.18461127 | -0.1823128411 | -1.132867 | 8567 | 3783 | 563.086587 | 3 | 1.63235 |
| **45** | Irán (República islámica de) | 1.3049967 | -0.3440354 | 0.49847561611 | -1.347493 | 35912 | 5663 | 453.322342 | 3 | 1.52314 |
| **46** | Irlanda | -0.616186 | -0.5868205 | 1.13310892365 | -0.739386 | 58265 | 48893 | 451.758218 | 3 | 0.89662 |
| **47** | Israel | -0.538855 | -0.0383061 | 1.30619073479 | -1.132867 | -2817 | 31222 | 456.133843 | 3 | 1.00890 |
| **48** | Jordania | -0.56374 | 0.53718443 | 0.34847137978 | -1.633661 | -8062 | 4094 | 83.7485605 | 3 | 1.20547 |
| **49** | Kazajstán | -0.293566 | 0.05161431 | -0.1938516285 | 0.1548884 | 33220 | 9299 | 58.2563289 | 3 | 1.25579 |
| **50** | Líbano | -0.623911 | -0.6227886 | 0.93694953768 | -1.383264 | -13439 | 8850 | 242.71487 | 3 | 0.93278 |
| **51** | Libia | -0.57072 | -0.020322 | 0.08307926935 | -1.132867 | 35510 | 13400 | 14.1908173 | 3 | 0.68401 |
| **52** | Marruecos | 0.1422503 | -0.0562902 | 0.32539380496 | -0.989783 | -17620 | 2869 | 257.684768 | 3 | 0.39731 |
| **53** | Mongolia | -0.668715 | 0.24044713 | -0.2169292033 | -0.810928 | -379 | 2286 | 3.13064835 | 3 | 1.00113 |
| **54** | Myanmar | 0.6836753 | -0.3350434 | -0.5977091878 | -0.095509 | 3901 | 799 | 2783.14937 | 3 | 1.53739 |
| **55** | Nepal | -0.002093 | -0.0832663 | -0.2053904159 | -0.703615 | -4550 | 607 | 1883.68115 | 3 | 0.71671 |
| **56** | Nicaragua | -0.585259 | -0.0832663 | 0.42924289165 | -1.311722 | -2384 | 1535 | 191.232499 | 3 | 0.66748 |
| **57** | Nueva Caledonia | -0.736776 | -0.5598444 | 0.62540227761 | -0.560531 | -1820 | 36789 | 18.5054348 | 3 | 0.74975 |
| **58** | Nueva Zelanda | -0.623028 | -0.7396852 | 1.24849679774 | -0.596302 | 466 | 33260 | 88.7674645 | 3 | 1.07428 |
| **59** | Omán | -0.662342 | -0.1012504 | 0.63694106502 | -2.062913 | 16827 | 20923 | 39.7148676 | 3 | 1.33908 |
| **60** | Panamá | -0.643664 | -0.1911708 | 0.75232893912 | -1.24018 | -8313 | 7355 | 161.716998 | 3 | 0.70715 |
| **61** | Paraguay | -0.572265 | -0.020322 | 0.22154471827 | -1.025554 | -3524 | 3103 | 231.421674 | 3 | 0.58119 |
| **62** | Perú | 0.0668233 | -0.1372185 | 0.39462652942 | -1.025554 | 6747 | 5026 | 200.118104 | 3 | 0.30138 |
| **63** | República de Moldova | -0.630863 | -1.1443269 | 1.22541922293 | -1.061325 | -2314 | 1627 | 969.746341 | 3 | 1.28086 |
| **64** | República Dominicana | -0.470491 | 0.47424016 | -0.6092479752 | -0.52476 | -10349 | 5089 | 242.448138 | 3 | 1.38364 |
| **65** | República Unida de Tanzanía | 0.5158267 | -0.0472981 | 0.27923865532 | -0.88247 | -4186 | 686 | 515.276753 | 3 | 0.71112 |
| **66** | Ruanda | -0.459566 | -0.5598444 | 1.00618226214 | -1.061325 | -1146 | 526 | 401.4965 | 3 | 0.73003 |
| **67** | Sri Lanka | -0.186246 | -0.4969001 | 0.44078167906 | -0.667844 | -5205 | 2388 | 1661.61338 | 3 | 0.29871 |
| **68** | Tailandia | 1.0963722 | -0.9644861 | 0.38308774201 | -0.274364 | 10252 | 5102 | 319.596523 | 3 | 1.52974 |
| **69** | Túnez | -0.450048 | -0.3170593 | 0.44078167906 | -0.667844 | -5791 | 4143 | 0.59529715 | 3 | 0.37095 |
| **70** | Turquía | 1.2513922 | -0.4159717 | 0.46385925388 | -0.989783 | -71661 | 10135 | 28.8853594 | 3 | 1.39989 |
| **71** | Uzbekistán | 0.021716 | 0.12355062 | -0.2861619278 | -0.52476 | 3201 | 1423 | 277.871862 | 3 | 0.91760 |
| **72** | Venezuela (República Bolivariana de) | 0.0563948 | -0.1731867 | 0.3715489546 | -1.061325 | 31930 | 13559 | 179.166667 | 3 | 0.30808 |
| **73** | Viet Nam | 1.6941054 | -0.4069797 | 0.55616955315 | -0.954012 | -12121 | 1302 | 4113.40159 | 3 | 1.83916 |
| **74** | Alemania | 1.4755215 | -1.2252553 | 1.14464771106 | 0.8345373 | 205408 | 41100 | 2659.28619 | 4 | 1.67949 |
| **75** | Armenia | -0.661817 | -0.7756533 | 0.44078167906 | 0.1906594 | -2771 | 3125 | 191.307784 | 4 | 0.89579 |
| **76** | Austria | -0.512039 | -1.1173508 | 1.1908028607 | 0.3337434 | -5712 | 46377 | 1593.5443 | 4 | 0.73291 |
| **77** | Bielorrusia | -0.481692 | -0.9195259 | 0.03692411971 | 2.050751 | -9601 | 5818 | 764.607435 | 4 | 1.49102 |
| **78** | Bosnia y Herzegovina | -0.637733 | -1.153319 | 0.63694106502 | 0.6556823 | -4402 | 4380 | 511.767948 | 4 | 0.54853 |
| **79** | Bulgaria | -0.539214 | -1.1263429 | 0.3715489546 | 2.3726899 | -4902 | 6587 | 1371.64747 | 4 | 1.67262 |
| **80** | Croacia | -0.624463 | -1.0903747 | 0.7292513643 | 1.4068731 | -8244 | 13754 | 2265.72666 | 4 | 0.77617 |
| **81** | Dinamarca | -0.590418 | -1.0364225 | 1.06387619919 | 0.4410563 | 12589 | 57614 | 3342.91159 | 4 | 0.64090 |
| **82** | Eslovaquia | -0.594391 | -1.0274304 | 0.60232470279 | 0.4768273 | -2098 | 16381 | 1366.35703 | 4 | 0.57839 |
| **83** | Eslovenia | -0.686923 | -1.0274304 | 1.0754149866 | 0.3337434 | -1588 | 23352 | 1190.30417 | 4 | 0.77509 |
| **84** | España | 0.542091 | -1.1353349 | 1.32926830961 | 0.1191174 | -69274 | 30999 | 717.158754 | 4 | 1.12523 |
| **85** | Finlandia | -0.595467 | -1.0094463 | 1.12157013624 | 0.4052853 | 719 | 46165 | 1309.54048 | 4 | 0.69163 |
| **86** | Francia | 0.9934394 | -0.8565817 | 1.28311315997 | 0.1548884 | -91697 | 40617 | 2367.79351 | 4 | 1.41530 |
| **87** | Georgia | -0.626311 | -0.7396852 | 0.44078167906 | 1.0849342 | -3580 | 2652 | 89.3115619 | 4 | 0.70186 |
| **88** | Grecia | -0.435178 | -1.171303 | 1.13310892365 | 0.7272243 | -39337 | 26967 | 493.14951 | 4 | 0.49365 |
| **89** | Hungría | -0.467263 | -1.1353349 | 0.4869368287 | 1.6930411 | 7147 | 12939 | 2305.02979 | 4 | 0.99765 |
| **90** | Lituania | -0.657403 | -1.0544065 | 0.26769986791 | 2.4084609 | -2658 | 12089 | 980.225194 | 4 | 1.76080 |
| **91** | Noruega | -0.608627 | -0.9195259 | 1.21388043552 | -0.023967 | 53344 | 87611 | 1154.23387 | 4 | 1.04659 |
| **92** | Países Bajos | -0.284737 | -1.0184384 | 1.21388043552 | -0.023967 | 52718 | 50339 | 978.908795 | 4 | 0.94545 |
| **93** | Polonia | 0.3206654 | -1.0364225 | 0.7292513643 | 0.4768273 | -18320 | 12479 | 1866.63914 | 4 | 0.57231 |
| **94** | Portugal | -0.451538 | -1.2072712 | 1.13310892365 | 0.6556823 | -26838 | 22503 | 301.231573 | 4 | 0.53111 |
| **95** | Reino Unido | 0.9866802 | -0.8385976 | 1.12157013624 | 0.2622014 | -152487 | 38796 | 0.00232802 | 4 | 1.31444 |
| **96** | República de Corea | 0.6107588 | -1.0544065 | 0.87925560063 | 0.5841403 | 41172 | 22588 | 3548.76957 | 4 | 0.79739 |
| **97** | República Democrática del Congo | 1.0755981 | -0.9914623 | 0.06000169453 | 0.9776212 | 800 | 347 | 58.7388697 | 4 | 1.42737 |
| **98** | República Democrática Popular Lao | -0.570858 | -0.6767409 | -0.1015413292 | 0.2622014 | -314 | 1054 | 1657.15674 | 4 | 1.16019 |
| **99** | Suecia | -0.484699 | -0.8925498 | 1.28311315997 | 0.3337434 | 9616 | 52053 | 1421.08888 | 4 | 0.77797 |
| **100** | Ucrania | 0.5157991 | -1.0004543 | 0.00230775748 | 2.4442319 | -9337 | 3066 | 71.7391336 | 4 | 1.93243 |

El código SAS abajo ayuda a mirar los grupos de otra forma.

**proc** **Freq** data=cluster4 ;

tables cluster;**run**;

**proc** **sort**;by cluster;

**proc** **print**;

by cluster;

**run**;

The FREQ Procedure

| **Cluster** | | | | |
| --- | --- | --- | --- | --- |
| **CLUSTER** | **Frequency** | **Percent** | **Cumulative Frequency** | **Cumulative Percent** |
| **1** | 29 | 29.00 | 29 | 29.00 |
| **2** | 5 | 5.00 | 34 | 34.00 |
| **3** | 39 | 39.00 | 73 | 73.00 |
| **4** | 27 | 27.00 | 100 | 100.00 |

El grupo 3 es el mas grande grupo y el grupo 2 el mas pequeño.

|  |
| --- |
| The SAS System |

**Grupo 1: Natalidad Alta, esperanza de vida muy baja y alta mortalidad.**

Cluster=1

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Afganistán | 0.0278959 | 1.22957146 | -1.2669588576 | 0.0475754 | -4766 | 566 | 157.13532 | 0.50611 |
| **2** | Benin | -0.481223 | 1.31949186 | -1.3361915821 | 0.4052853 | -1057 | 690 | 393.304324 | 0.21309 |
| **3** | Burkina Faso | -0.312298 | 1.69715751 | -1.4631182436 | 0.5483693 | -837 | 579 | 377.841176 | 0.41393 |
| **4** | Camerún | -0.175514 | 1.40042021 | -1.8323594407 | 1.2280182 | -952 | 1145 | 287.733333 | 0.95192 |
| **5** | Chad | -0.415342 | 2.15575152 | -2.2708333623 | 2.158064 | 903 | 736 | 43.048546 | 2.13877 |
| **6** | Congo | -0.631387 | 1.38243613 | -1.082338259 | 0.1906594 | 5213 | 2987 | 2.35377358 | 0.49654 |
| **7** | Etiopía | 1.6721449 | 1.01376252 | -0.8861788731 | -0.238593 | -6216 | 302 | 424.364723 | 2.12680 |
| **8** | Gabón | -0.701021 | 0.79795357 | -0.8169461486 | 0.1906594 | 5706 | 8278 | 7.90697674 | 0.87950 |
| **9** | Ghana | -0.072664 | 1.04073864 | -1.1284934087 | 0.2622014 | -3077 | 1326 | 185.14121 | 0.43431 |
| **10** | Guinea | -0.43973 | 1.40941225 | -1.474657031 | 0.6914533 | 66 | 481 | 190.483333 | 0.35217 |
| **11** | Haití | -0.467705 | 0.32137549 | -0.9784891724 | 0.1548884 | -2568 | 621 | 246.949153 | 1.11663 |
| **12** | Lesoto | -0.688082 | 0.62710483 | -2.4554539608 | 2.3011479 | -1404 | 1083 | 75.6450766 | 2.33178 |
| **13** | Madagascar | -0.161996 | 1.15763515 | -0.7246358493 | -0.560531 | -1464 | 415 | 124.711605 | 1.17805 |
| **14** | Malawi | -0.336107 | 1.58925304 | -1.1284934087 | 0.0475754 | -1044 | 464 | 629.517698 | 0.50912 |
| **15** | Malí | -0.325127 | 2.02087093 | -1.5669673303 | 0.9060792 | -1434 | 672 | 154.096516 | 0.87357 |
| **16** | Mozambique | -0.072582 | 1.6252212 | -1.8669758029 | 1.1922472 | -1600 | 424 | 50.1621622 | 1.00064 |
| **17** | Namibia | -0.683061 | 0.74400134 | -0.7477134242 | -0.417447 | -1276 | 5113 | 2.97611379 | 1.25097 |
| **18** | Níger | -0.29409 | 2.50644105 | -1.1631097709 | 0.4052853 | -1537 | 360 | 116.190233 | 1.18714 |
| **19** | Papúa Nueva Guinea | -0.554636 | 0.6450889 | -0.9784891724 | -0.274364 | 1792 | 1415 | 11.8907563 | 1.08138 |
| **20** | Republica Checa | -0.453717 | 1.11267495 | -2.4554539608 | 2.4084609 | 6420 | 19616 | 1629.30154 | 2.29393 |
| **21** | República Democrática Popular de Corea | -0.067616 | 1.85901422 | -1.4631182436 | 0.7987663 | -1075 | 570 | 1727.26236 | 0.68851 |
| **22** | Rumania | -0.183543 | 0.9867864 | -0.8861788731 | -0.345905 | -12528 | 7685 | 1216.98726 | 0.96139 |
| **23** | Senegal | -0.386098 | 1.16662719 | -0.5284764634 | -0.488989 | -2383 | 998 | 196.097837 | 1.23146 |
| **24** | Serbia | -0.493638 | 1.52630876 | -0.574631613 | -0.667844 | -6920 | 5412 | 1838.67445 | 1.36336 |
| **25** | Togo | -0.567244 | 1.29251574 | -1.3592691569 | 0.2979724 | -356 | 503 | 68.0041558 | 0.33247 |
| **26** | Uganda | 0.17097 | 1.95792665 | -1.5554285428 | 0.6199113 | -1594 | 636 | 226.798485 | 0.82660 |
| **27** | Yemen | -0.092694 | 1.01376252 | -0.8400237234 | -0.488989 | -1249 | 1358 | 43.0289707 | 1.10351 |
| **28** | Zambia | -0.359585 | 1.67917343 | -1.3823467317 | 0.4410563 | 1888 | 1225 | 129.955571 | 0.36717 |
| **29** | Zimbabue | -0.35804 | 1.27453166 | -1.8438982281 | 0.9418502 | -601 | 721 | 88.2230864 | 0.73819 |

**Grupo 2: Población muy grande.**

Cluster=2

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **30** | Bangladesh | 3.4393285 | -0.1372185 | 0.0253853323 | -1.061325 | -11877 | 758 | 5684.91934 | 0.88451 |
| **31** | Brasil | 4.7359345 | -0.6137966 | 0.38308774201 | -0.846699 | 10378 | 10978 | 272.824753 | 1.61724 |
| **32** | Federación Rusa | 3.2059571 | -0.8296056 | -0.1130801166 | 1.943438 | 168013 | 10618 | 278.162132 | 2.25278 |
| **33** | Filipinas | 1.8232473 | 0.1864949 | -0.32077829 | -0.632073 | -6992 | 2136 | 1786.91037 | 1.71051 |
| **34** | Pakistán | 3.9477025 | 0.70803318 | -0.5630928256 | -0.345905 | -16373 | 1008 | 1287.39645 | 1.09780 |

**Grupo 3: Creo que son los países que no están ni el grupo 1, 2 y 4. Quizás países que son parecidos, pero se puede ver países (marcados con color) con gran distancia del grupo.**

Cluster=3

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **35** | Albania | -0.6635 | -0.7936374 | 0.77540651394 | -0.453218 | -2861 | 3786 | 577.685262 | 0.90394 |
| **36** | Australia | -0.132118 | -0.7576692 | 1.30619073479 | -0.632073 | 10724 | 57593 | 82.6334118 | 1.00514 |
| **37** | Bolivia (Estado Plurinacional de) | -0.469939 | 0.22246305 | -0.3553946522 | -0.310135 | 998 | 1935 | 49.2931894 | 1.14646 |
| **38** | Canadá | 0.1979516 | -0.9914623 | 1.27157437256 | -0.417447 | -4638 | 47297 | 694.911701 | 1.21972 |
| **39** | Chile | -0.274143 | -0.7576692 | 1.20234164811 | -1.204409 | 11068 | 12685 | 225.751787 | 0.96594 |
| **40** | Colombia | 0.523248 | -0.5148842 | 0.34847137978 | -0.954012 | -973 | 6180 | 95.7396875 | 0.71587 |
| **41** | Cuba | -0.431591 | -1.0274304 | 0.97156589991 | -0.274364 | -5975 | 5702 | 121.623946 | 1.12461 |
| **42** | Ecuador | -0.331555 | -0.0652822 | 0.54463076574 | -1.168638 | -3176 | 4637 | 363.654456 | 0.43166 |
| **43** | Fiji | -0.719836 | -0.1102424 | -0.124618904 | -0.596302 | -975 | 3649 | 20.2988235 | 0.87932 |
| **44** | Irak | 0.1080406 | 1.18461127 | -0.1823128411 | -1.132867 | 8567 | 3783 | 563.086587 | 1.63235 |
| **45** | Irán (República islámica de) | 1.3049967 | -0.3440354 | 0.49847561611 | -1.347493 | 35912 | 5663 | 453.322342 | 1.52314 |
| **46** | Irlanda | -0.616186 | -0.5868205 | 1.13310892365 | -0.739386 | 58265 | 48893 | 451.758218 | 0.89662 |
| **47** | Israel | -0.538855 | -0.0383061 | 1.30619073479 | -1.132867 | -2817 | 31222 | 456.133843 | 1.00890 |
| **48** | Jordania | -0.56374 | 0.53718443 | 0.34847137978 | -1.633661 | -8062 | 4094 | 83.7485605 | 1.20547 |
| **49** | Kazajstán | -0.293566 | 0.05161431 | -0.1938516285 | 0.1548884 | 33220 | 9299 | 58.2563289 | 1.25579 |
| **50** | Líbano | -0.623911 | -0.6227886 | 0.93694953768 | -1.383264 | -13439 | 8850 | 242.71487 | 0.93278 |
| **51** | Libia | -0.57072 | -0.020322 | 0.08307926935 | -1.132867 | 35510 | 13400 | 14.1908173 | 0.68401 |
| **52** | Marruecos | 0.1422503 | -0.0562902 | 0.32539380496 | -0.989783 | -17620 | 2869 | 257.684768 | 0.39731 |
| **53** | Mongolia | -0.668715 | 0.24044713 | -0.2169292033 | -0.810928 | -379 | 2286 | 3.13064835 | 1.00113 |
| **54** | Myanmar | 0.6836753 | -0.3350434 | -0.5977091878 | -0.095509 | 3901 | 799 | 2783.14937 | 1.53739 |
| **55** | Nepal | -0.002093 | -0.0832663 | -0.2053904159 | -0.703615 | -4550 | 607 | 1883.68115 | 0.71671 |
| **56** | Nicaragua | -0.585259 | -0.0832663 | 0.42924289165 | -1.311722 | -2384 | 1535 | 191.232499 | 0.66748 |
| **57** | Nueva Caledonia | -0.736776 | -0.5598444 | 0.62540227761 | -0.560531 | -1820 | 36789 | 18.5054348 | 0.74975 |
| **58** | Nueva Zelanda | -0.623028 | -0.7396852 | 1.24849679774 | -0.596302 | 466 | 33260 | 88.7674645 | 1.07428 |
| **59** | Omán | -0.662342 | -0.1012504 | 0.63694106502 | -2.062913 | 16827 | 20923 | 39.7148676 | 1.33908 |
| **60** | Panamá | -0.643664 | -0.1911708 | 0.75232893912 | -1.24018 | -8313 | 7355 | 161.716998 | 0.70715 |
| **61** | Paraguay | -0.572265 | -0.020322 | 0.22154471827 | -1.025554 | -3524 | 3103 | 231.421674 | 0.58119 |
| **62** | Perú | 0.0668233 | -0.1372185 | 0.39462652942 | -1.025554 | 6747 | 5026 | 200.118104 | 0.30138 |
| **63** | República de Moldova | -0.630863 | -1.1443269 | 1.22541922293 | -1.061325 | -2314 | 1627 | 969.746341 | 1.28086 |
| **64** | República Dominicana | -0.470491 | 0.47424016 | -0.6092479752 | -0.52476 | -10349 | 5089 | 242.448138 | 1.38364 |
| **65** | República Unida de Tanzanía | 0.5158267 | -0.0472981 | 0.27923865532 | -0.88247 | -4186 | 686 | 515.276753 | 0.71112 |
| **66** | Ruanda | -0.459566 | -0.5598444 | 1.00618226214 | -1.061325 | -1146 | 526 | 401.4965 | 0.73003 |
| **67** | Sri Lanka | -0.186246 | -0.4969001 | 0.44078167906 | -0.667844 | -5205 | 2388 | 1661.61338 | 0.29871 |
| **68** | Tailandia | 1.0963722 | -0.9644861 | 0.38308774201 | -0.274364 | 10252 | 5102 | 319.596523 | 1.52974 |
| **69** | Túnez | -0.450048 | -0.3170593 | 0.44078167906 | -0.667844 | -5791 | 4143 | 0.59529715 | 0.37095 |
| **70** | Turquía | 1.2513922 | -0.4159717 | 0.46385925388 | -0.989783 | -71661 | 10135 | 28.8853594 | 1.39989 |
| **71** | Uzbekistán | 0.021716 | 0.12355062 | -0.2861619278 | -0.52476 | 3201 | 1423 | 277.871862 | 0.91760 |
| **72** | Venezuela (República Bolivariana de) | 0.0563948 | -0.1731867 | 0.3715489546 | -1.061325 | 31930 | 13559 | 179.166667 | 0.30808 |
| **73** | Viet Nam | 1.6941054 | -0.4069797 | 0.55616955315 | -0.954012 | -12121 | 1302 | 4113.40159 | 1.83916 |

**Grupo 4: Natalidad Baja, esperanza de vida Alta.**

Cluster=4

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **74** | Alemania | 1.4755215 | -1.2252553 | 1.14464771106 | 0.8345373 | 205408 | 41100 | 2659.28619 | 1.67949 |
| **75** | Armenia | -0.661817 | -0.7756533 | 0.44078167906 | 0.1906594 | -2771 | 3125 | 191.307784 | 0.89579 |
| **76** | Austria | -0.512039 | -1.1173508 | 1.1908028607 | 0.3337434 | -5712 | 46377 | 1593.5443 | 0.73291 |
| **77** | Bielorrusia | -0.481692 | -0.9195259 | 0.03692411971 | 2.050751 | -9601 | 5818 | 764.607435 | 1.49102 |
| **78** | Bosnia y Herzegovina | -0.637733 | -1.153319 | 0.63694106502 | 0.6556823 | -4402 | 4380 | 511.767948 | 0.54853 |
| **79** | Bulgaria | -0.539214 | -1.1263429 | 0.3715489546 | 2.3726899 | -4902 | 6587 | 1371.64747 | 1.67262 |
| **80** | Croacia | -0.624463 | -1.0903747 | 0.7292513643 | 1.4068731 | -8244 | 13754 | 2265.72666 | 0.77617 |
| **81** | Dinamarca | -0.590418 | -1.0364225 | 1.06387619919 | 0.4410563 | 12589 | 57614 | 3342.91159 | 0.64090 |
| **82** | Eslovaquia | -0.594391 | -1.0274304 | 0.60232470279 | 0.4768273 | -2098 | 16381 | 1366.35703 | 0.57839 |
| **83** | Eslovenia | -0.686923 | -1.0274304 | 1.0754149866 | 0.3337434 | -1588 | 23352 | 1190.30417 | 0.77509 |
| **84** | España | 0.542091 | -1.1353349 | 1.32926830961 | 0.1191174 | -69274 | 30999 | 717.158754 | 1.12523 |
| **85** | Finlandia | -0.595467 | -1.0094463 | 1.12157013624 | 0.4052853 | 719 | 46165 | 1309.54048 | 0.69163 |
| **86** | Francia | 0.9934394 | -0.8565817 | 1.28311315997 | 0.1548884 | -91697 | 40617 | 2367.79351 | 1.41530 |
| **87** | Georgia | -0.626311 | -0.7396852 | 0.44078167906 | 1.0849342 | -3580 | 2652 | 89.3115619 | 0.70186 |
| **88** | Grecia | -0.435178 | -1.171303 | 1.13310892365 | 0.7272243 | -39337 | 26967 | 493.14951 | 0.49365 |
| **89** | Hungría | -0.467263 | -1.1353349 | 0.4869368287 | 1.6930411 | 7147 | 12939 | 2305.02979 | 0.99765 |
| **90** | Lituania | -0.657403 | -1.0544065 | 0.26769986791 | 2.4084609 | -2658 | 12089 | 980.225194 | 1.76080 |
| **91** | Noruega | -0.608627 | -0.9195259 | 1.21388043552 | -0.023967 | 53344 | 87611 | 1154.23387 | 1.04659 |
| **92** | Países Bajos | -0.284737 | -1.0184384 | 1.21388043552 | -0.023967 | 52718 | 50339 | 978.908795 | 0.94545 |
| **93** | Polonia | 0.3206654 | -1.0364225 | 0.7292513643 | 0.4768273 | -18320 | 12479 | 1866.63914 | 0.57231 |
| **94** | Portugal | -0.451538 | -1.2072712 | 1.13310892365 | 0.6556823 | -26838 | 22503 | 301.231573 | 0.53111 |
| **95** | Reino Unido | 0.9866802 | -0.8385976 | 1.12157013624 | 0.2622014 | -152487 | 38796 | 0.00232802 | 1.31444 |
| **96** | República de Corea | 0.6107588 | -1.0544065 | 0.87925560063 | 0.5841403 | 41172 | 22588 | 3548.76957 | 0.79739 |
| **97** | República Democrática del Congo | 1.0755981 | -0.9914623 | 0.06000169453 | 0.9776212 | 800 | 347 | 58.7388697 | 1.42737 |
| **98** | República Democrática Popular Lao | -0.570858 | -0.6767409 | -0.1015413292 | 0.2622014 | -314 | 1054 | 1657.15674 | 1.16019 |
| **99** | Suecia | -0.484699 | -0.8925498 | 1.28311315997 | 0.3337434 | 9616 | 52053 | 1421.08888 | 0.77797 |
| **100** | Ucrania | 0.5157991 | -1.0004543 | 0.00230775748 | 2.4442319 | -9337 | 3066 | 71.7391336 | 1.93243 |

El código SAS abajo nos ayuda a estudiar un poco más los grupos.

**proc** **means** data=cluster4 noprint;

by cluster;

var POBL NATALIDA ESPERANZ MORTALID;

output out=centinic mean=POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=centinic;**run**;

**PROC** **STANDARD** DATA=cluster4 MEAN=**0** STD=**1** OUT=cluster4out;

var POBL NATALIDA ESPERANZ MORTALID;

**RUN**;

**PROC** **PRINT** DATA=cluster4out (OBS=**7**);

**RUN**;

|  |
| --- |
| The SAS System |

| **Obs** | **CLUSTER** | **\_TYPE\_** | **\_FREQ\_** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | 1 | 0 | 1 | 0.0278959 | 1.22957146 | -1.2669588576 | 0.0475754 |
| **2** | 3 | 0 | 1 | -0.6635 | -0.7936374 | 0.77540651394 | -0.453218 |

|  |
| --- |
| The SAS System |

| **Obs** | **PAIS** | **POBL** | **NATALIDA** | **ESPERANZ** | **MORTALID** | **BALANZAC** | **PIB** | **PRODCERE** | **CLUSTER** | **DISTANCE** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | Afganistán | 0.0278959 | 1.22957146 | -1.2669588576 | 0.0475754 | -4766 | 566 | 157.13532 | 1 | 0.50611 |
| **2** | Albania | -0.6635 | -0.7936374 | 0.77540651394 | -0.453218 | -2861 | 3786 | 577.685262 | 3 | 0.90394 |
| **3** | Alemania | 1.4755215 | -1.2252553 | 1.14464771106 | 0.8345373 | 205408 | 41100 | 2659.28619 | 4 | 1.67949 |
| **4** | Armenia | -0.661817 | -0.7756533 | 0.44078167906 | 0.1906594 | -2771 | 3125 | 191.307784 | 4 | 0.89579 |
| **5** | Australia | -0.132118 | -0.7576692 | 1.30619073479 | -0.632073 | 10724 | 57593 | 82.6334118 | 3 | 1.00514 |
| **6** | Austria | -0.512039 | -1.1173508 | 1.1908028607 | 0.3337434 | -5712 | 46377 | 1593.5443 | 4 | 0.73291 |
| **7** | Bangladesh | 3.4393285 | -0.1372185 | 0.0253853323 | -1.061325 | -11877 | 758 | 5684.91934 | 2 | 0.88451 |

**Conclusiones:**

Se puede mejorar la clasificación de los grupos. Pienso que se puede bajar para 2 grupos o subir para más grupos para poder clasificar los países de una forma mejor porque en los grupos se puede ver distancias altas entre los países.

# 3. Ejecución – Análisis Discriminante

En primer lugar hemos de comprobar si se cumple la hipótesis de normalidad. Para ello utilizamos el “proc univariate”.

1. **Hacemos el test de la hipótesis de la normalidad con todas las variables.**

**proc** **univariate** data=cluster4 normal plot;

VAR POBL NATALIDA ESPERANZ MORTALID CLUSTER;

**run**;

La hipótesis es que la variable es normal.

Que es una variable normal?

<https://pt.wikipedia.org/wiki/Distribui%C3%A7%C3%A3o_normal>

Un p-valor de 0.01% significa que hay la confianza de 99% que se puede rechazar la hipótesis nula o sea se piensa que la variable es normal, pero no existe una certeza 100%.

Tabla para ayudar la interpretación

|  |  |  |
| --- | --- | --- |
| P-valor Bajo | Menor que 0.01% | No es normal |
| P-valor Alto | Más grande que 0.01% | Normal |

**Los resultados con las variables POBL NATALIDA ESPERANZ MORTALID CLUSTER**

|  |  |
| --- | --- |
| POBL | Bajo = No es normal |
| NATALIDA | Bajo = No es normal |
| ESPERANZ | Bajo = No es normal |
| MORTALID | Alto = Normal |
| CLUSTER | Bajo = No es normal |

1. **Comprobación de la existencia de datos atípicos**

**Los resultados con las variables POBL NATALIDA ESPERANZ MORTALID CLUSTER**

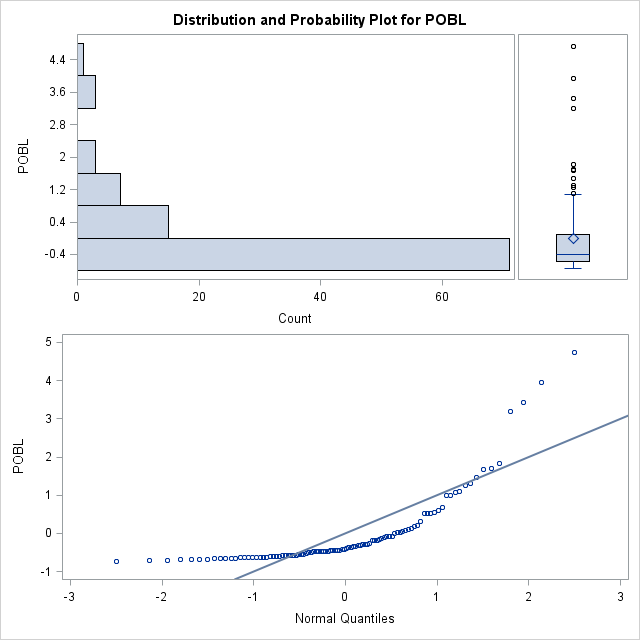
|  |  |
| --- | --- |
| POBL | Tiene datos atípicos |
| NATALIDA | No tiene datos atípicos |
| ESPERANZ | No tiene datos atípicos |
| MORTALID | Tiene datos atípicos |
| CLUSTER | No tiene datos atípicos |

En el gráfico de distribución y probabilidad de la variable POBL se puede ver los datos atípicos, ha sido con el BOXPLOT que yo he mirado para cada variable la existencia de datos atípicos.

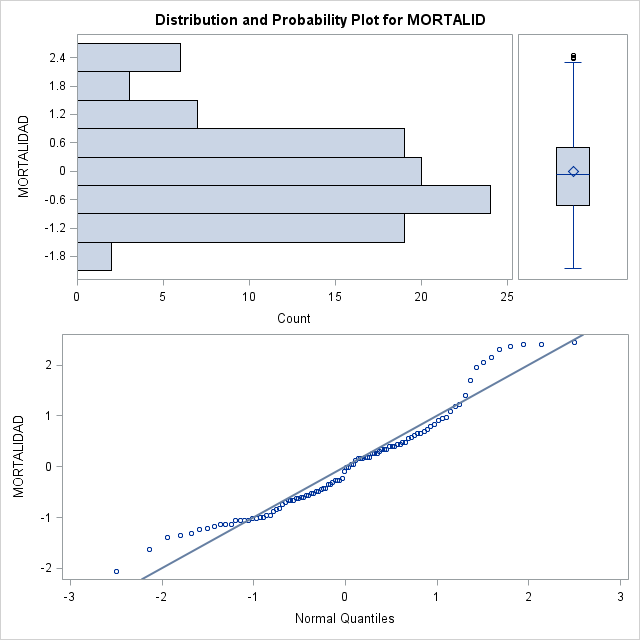
**Quitar atípicos**

Pienso que para la variable POBL hay que quitar todos los países donde POBL sea más grande que 1.2 (datos normalizados). Pero quizás lo correcto es quitar los que tienen 4.4 y 3.6, quizás quitando los países del grupo con muy grande población (grupo 3).

Otra cosa el grupo 3 tiene solo 5 observaciones y se pide como mínimo por lo menos 20 observaciones para cada grupo.



Hay países donde la mortalidad es de 2.4 y quizás lo mejor son quitar estos datos.



**Quitado atípicos con SAS – Tentativa 1**

No estoy haciendo bien el trabajo de quitar datos atípicos.

He pensado quitar todo el grupo 3, pero en la variable POBL aún me sale unos datos atípicos, mismo quitando los países con una gran población.

/\*

Quitando datos atípicos – Tentativa 1

\*/

**data** cluster4id;

set cluster4;

id=\_N\_;

**run**;

**proc** **print** data=cluster4id;**run**;

**proc** **print** data=cluster4id;

**run**;

**data** cluster4sinatipicos;

set cluster4id;

if cluster = **3** then delete;

**run**;

**proc** **print** data=cluster4sinatipicos;

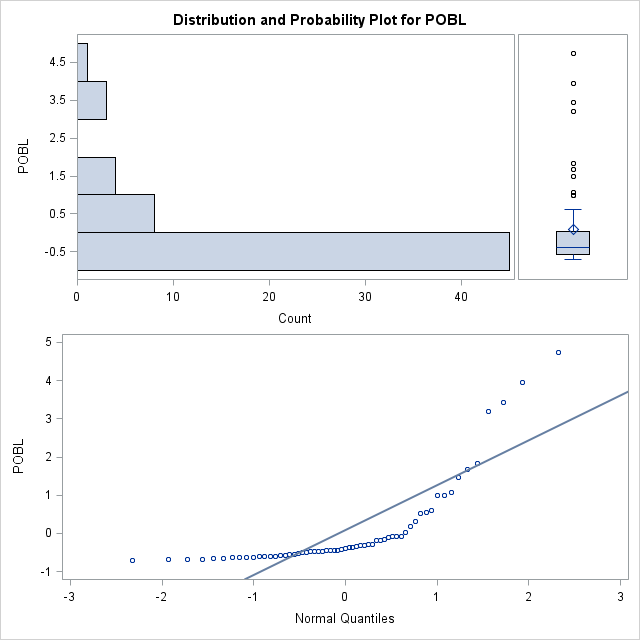
**run**;

**proc** **univariate** data=cluster4sinatipicos normal plot;

VAR POBL NATALIDA ESPERANZ MORTALID CLUSTER;

id id;

**run**;



**Quitando atípicos tentativa 2**

Ahora vamos probar otra forma para quitar los datos atípicos mirando la tabla de Extreme Observations de SAS y quitar de forma individual los datos.

/\*

Quitando datos atipicos - Tentativa 2

\*/

**data** cluster4id;

set cluster4;

id=\_N\_;

**run**;

**proc** **discrim** data=cluster4id method=normal pool=test wcov pcov list crossvalidate crosslist outstat=salida;

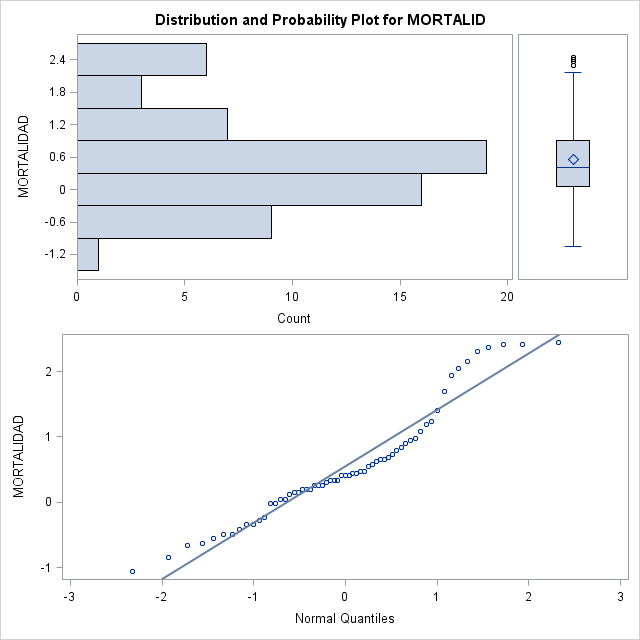
class CLUSTER;

VAR POBL NATALIDA ESPERANZ MORTALID;

**run**;

MORTALID

| **Extreme Observations** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **Lowest** | | | **Highest** | | |
| **Value** | **id** | **Obs** | **Value** | **id** | **Obs** |
| -1.061325 | 7 | 5 | 2.30115 | 47 | 32 |
| -0.846699 | 12 | 9 | 2.37269 | 13 | 10 |
| -0.667844 | 86 | 54 | 2.40846 | 50 | 33 |
| -0.632073 | 31 | 22 | 2.40846 | 75 | 47 |
| -0.560531 | 51 | 34 | 2.44423 | 93 | 57 |



Con el código SAS abajo se quita algunos países atípicos por la variable de Mortalidad.

**data** cluster4sinatipicos2;

set cluster4id;

if \_N\_ in (**7**,**12**,**47**,**13**) then delete;

**run**;

**proc** **print** data=cluster4sinatipicos2;

**run**;

**proc** **discrim** data=cluster4sinatipicos2 method=normal pool=test wcov pcov list crossvalidate crosslist outstat=salida;

class CLUSTER;

VAR POBL NATALIDA ESPERANZ MORTALID;

**run**;

1. **Transformaciones de las variables que no son normales**

He intentado transformar las variables que no son normales utilizando el log pero me ha dado un error:

NOTE: Invalid argument to function LOG(-0.63138734) at line 639 column 11.

WARNING: Limit set by ERRORS= option reached. Further errors of this type will not be printed.

PAIS=Congo POBL=-0.631387 NATALIDA=1.38243613 ESPERANZ=-1.082338259 MORTALID=0.1906594

BALANZAC=5213 PIB=2987 PRODCERE=2.35377358 CLUSTER=1 DISTANCE=0.4965423476 logESPERANZ=.

logNATALIDA=0.3238472567 logPOBL=. \_ERROR\_=1 \_N\_=20

NOTE: Mathematical operations could not be performed at the following places. The results of

the operations have been set to missing values.

Each place is given by: (Number of times) at (Line):(Column).

42 at 637:15 62 at 638:15 71 at 639:11

NOTE: There were 100 observations read from the data set WORK.CLUSTER4.

NOTE: The data set WORK.LOGCLUSTER has 100 observations and 13 variables.

NOTE: DATA statement used (Total process time):

real time 0.03 seconds

cpu time 0.03 seconds

He conseguido arreglar utilizando las variables originales, pero solo en el final de todo hay la parte que yo he hecho con variables logarítmicas. Donde se quita los atípicos utilizando las variables logarítmicas.

Código SAS:

**data** logcluster;

set cluster4;

logESPERANZ = log(ESPERANZ);

logNATALIDA = log(NATALIDA);

logPOBL = log(POBL);

**run**;

**proc** **univariate** data=logcluster4 normal plot;

var logESPERANZ logNATALIDA logNATALIDA logPOBL;

by CLUSTER;

**run**;

En el final he conseguido poner el logaritmo.

Creo que esto pasa porque las variables están normalizadas y por esto he tenido que hacer una transformación para agrupar los clusters con los valores originales de las variables.

**data** cluster\_pequeno;

set cluster4id;

DROP POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=cluster\_pequeno;

**run**;

**data** sample\_paises\_con\_clusters;

merge sample\_paises cluster\_pequeno;

**run**;

**proc** **print** data=sample\_paises\_con\_clusters;

**run**;

**proc** **print** data=sample\_paises;

**run**;

He hecho el Box-Cox

**data** sample\_paises\_con\_clusterstransf;

set sample\_paises\_con\_clusters;

z=**0**;

**run**;

**proc** **transreg** data=sample\_paises\_con\_clusterstransf maxiter=**0** nozeroconstant detail

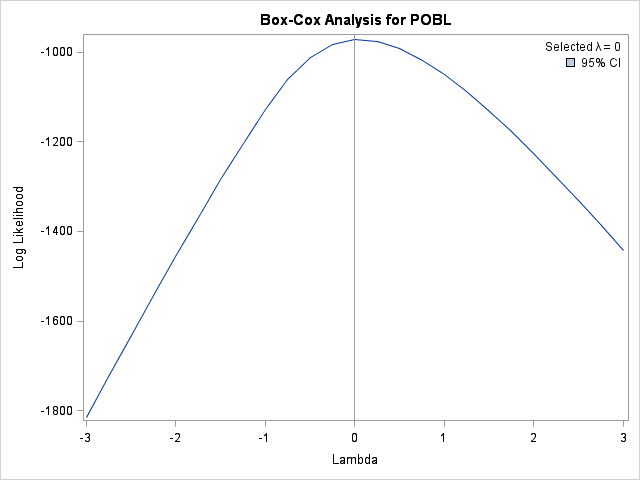
plots=(transformation(dependent)scatter);

model boxcox(POBL)=identity(z);

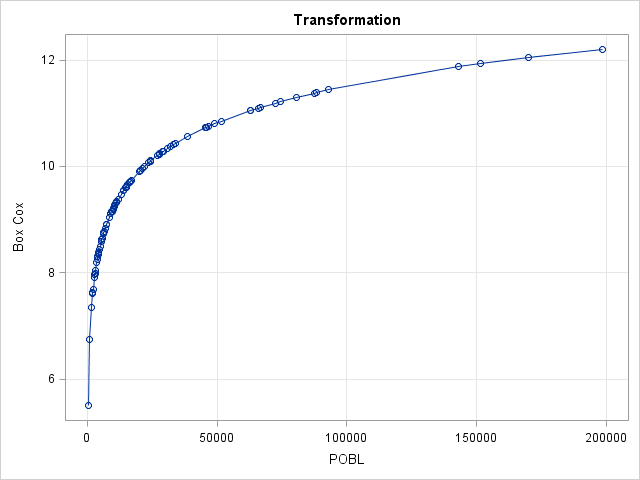
output out=tdatos;

**run**;

Resultados Box-Cox:



| **Model Statement Specification Details** | | | | |
| --- | --- | --- | --- | --- |
| **Type** | **DF** | **Variable** | **Description** | **Value** |
| **Dep** | **1** | **BoxCox(POBL)** | Lambda Used | 0 |
|  |  |  | Lambda | 0 |
|  |  |  | Log Likelihood | -971.4 |
|  |  |  | Conv. Lambda | 0 |
|  |  |  | Conv. Lambda LL | -971.4 |
|  |  |  | CI Limit | -973.3 |
|  |  |  | Alpha | 0.05 |
|  |  |  | Label | POBL |
| **Ind** | **0** | **Identity(z)** | Options | All Zero |



**proc** **print** data=tdatos;

**run**;

|  |
| --- |
| The SAS System |

| **Obs** | **\_TYPE\_** | **\_NAME\_** | **POBL** | **TPOBL** | **Intercept** | **z** | **TIntercept** | **Tz** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **1** | SCORE | ROW1 | 27963 | 10.2386 | 1 | 0 | 1 | 0 |
| **2** | SCORE | ROW2 | 2902 | 7.9732 | 1 | 0 | 1 | 0 |
| **3** | SCORE | ROW3 | 80435 | 11.2952 | 1 | 0 | 1 | 0 |
| **4** | SCORE | ROW4 | 2963 | 7.9940 | 1 | 0 | 1 | 0 |
| **5** | SCORE | ROW5 | 22163 | 10.0062 | 1 | 0 | 1 | 0 |
| **6** | SCORE | ROW6 | 8392 | 9.0350 | 1 | 0 | 1 | 0 |
| **7** | SCORE | ROW7 | 151617 | 11.9291 | 1 | 0 | 1 | 0 |
| **8** | SCORE | ROW8 | 9509 | 9.1600 | 1 | 0 | 1 | 0 |
| **9** | SCORE | ROW9 | 9492 | 9.1582 | 1 | 0 | 1 | 0 |
| **10** | SCORE | ROW10 | 9918 | 9.2021 | 1 | 0 | 1 | 0 |
| **11** | SCORE | ROW11 | 3836 | 8.2522 | 1 | 0 | 1 | 0 |
| **12** | SCORE | ROW12 | 198615 | 12.1991 | 1 | 0 | 1 | 0 |
| **13** | SCORE | ROW13 | 7407 | 8.9102 | 1 | 0 | 1 | 0 |
| **14** | SCORE | ROW14 | 15632 | 9.6571 | 1 | 0 | 1 | 0 |
| **15** | SCORE | ROW15 | 20590 | 9.9326 | 1 | 0 | 1 | 0 |
| **16** | SCORE | ROW16 | 34127 | 10.4378 | 1 | 0 | 1 | 0 |
| **17** | SCORE | ROW17 | 11897 | 9.3840 | 1 | 0 | 1 | 0 |
| **18** | SCORE | ROW18 | 17015 | 9.7419 | 1 | 0 | 1 | 0 |
| **19** | SCORE | ROW19 | 45918 | 10.7346 | 1 | 0 | 1 | 0 |
| **20** | SCORE | ROW20 | 4066 | 8.3104 | 1 | 0 | 1 | 0 |
| **21** | SCORE | ROW21 | 4317 | 8.3703 | 1 | 0 | 1 | 0 |
| **22** | SCORE | ROW22 | 11308 | 9.3333 | 1 | 0 | 1 | 0 |
| **23** | SCORE | ROW23 | 5551 | 8.6217 | 1 | 0 | 1 | 0 |
| **24** | SCORE | ROW24 | 14934 | 9.6114 | 1 | 0 | 1 | 0 |
| **25** | SCORE | ROW25 | 5407 | 8.5954 | 1 | 0 | 1 | 0 |
| **26** | SCORE | ROW26 | 2053 | 7.6271 | 1 | 0 | 1 | 0 |
| **27** | SCORE | ROW27 | 46601 | 10.7494 | 1 | 0 | 1 | 0 |
| **28** | SCORE | ROW28 | 87562 | 11.3801 | 1 | 0 | 1 | 0 |
| **29** | SCORE | ROW29 | 143158 | 11.8717 | 1 | 0 | 1 | 0 |
| **30** | SCORE | ROW30 | 860 | 6.7569 | 1 | 0 | 1 | 0 |
| **31** | SCORE | ROW31 | 93039 | 11.4408 | 1 | 0 | 1 | 0 |
| **32** | SCORE | ROW32 | 5368 | 8.5882 | 1 | 0 | 1 | 0 |
| **33** | SCORE | ROW33 | 62961 | 11.0503 | 1 | 0 | 1 | 0 |
| **34** | SCORE | ROW34 | 1542 | 7.3408 | 1 | 0 | 1 | 0 |
| **35** | SCORE | ROW35 | 4250 | 8.3547 | 1 | 0 | 1 | 0 |
| **36** | SCORE | ROW36 | 24318 | 10.0990 | 1 | 0 | 1 | 0 |
| **37** | SCORE | ROW37 | 11178 | 9.3217 | 1 | 0 | 1 | 0 |
| **38** | SCORE | ROW38 | 11013 | 9.3068 | 1 | 0 | 1 | 0 |
| **39** | SCORE | ROW39 | 9999 | 9.2102 | 1 | 0 | 1 | 0 |
| **40** | SCORE | ROW40 | 10015 | 9.2118 | 1 | 0 | 1 | 0 |
| **41** | SCORE | ROW41 | 30868 | 10.3375 | 1 | 0 | 1 | 0 |
| **42** | SCORE | ROW42 | 74254 | 11.2152 | 1 | 0 | 1 | 0 |
| **43** | SCORE | ROW43 | 4617 | 8.4375 | 1 | 0 | 1 | 0 |
| **44** | SCORE | ROW44 | 7420 | 8.9119 | 1 | 0 | 1 | 0 |
| **45** | SCORE | ROW45 | 6518 | 8.7823 | 1 | 0 | 1 | 0 |
| **46** | SCORE | ROW46 | 16311 | 9.6996 | 1 | 0 | 1 | 0 |
| **47** | SCORE | ROW47 | 2011 | 7.6064 | 1 | 0 | 1 | 0 |
| **48** | SCORE | ROW48 | 4337 | 8.3749 | 1 | 0 | 1 | 0 |
| **49** | SCORE | ROW49 | 6265 | 8.7427 | 1 | 0 | 1 | 0 |
| **50** | SCORE | ROW50 | 3123 | 8.0465 | 1 | 0 | 1 | 0 |
| **51** | SCORE | ROW51 | 21080 | 9.9561 | 1 | 0 | 1 | 0 |
| **52** | SCORE | ROW52 | 14769 | 9.6003 | 1 | 0 | 1 | 0 |
| **53** | SCORE | ROW53 | 15167 | 9.6269 | 1 | 0 | 1 | 0 |
| **54** | SCORE | ROW54 | 32108 | 10.3769 | 1 | 0 | 1 | 0 |
| **55** | SCORE | ROW55 | 2713 | 7.9058 | 1 | 0 | 1 | 0 |
| **56** | SCORE | ROW56 | 24321 | 10.0991 | 1 | 0 | 1 | 0 |
| **57** | SCORE | ROW57 | 51733 | 10.8539 | 1 | 0 | 1 | 0 |
| **58** | SCORE | ROW58 | 2193 | 7.6930 | 1 | 0 | 1 | 0 |
| **59** | SCORE | ROW59 | 26876 | 10.1990 | 1 | 0 | 1 | 0 |
| **60** | SCORE | ROW60 | 5738 | 8.6549 | 1 | 0 | 1 | 0 |
| **61** | SCORE | ROW61 | 16292 | 9.6984 | 1 | 0 | 1 | 0 |
| **62** | SCORE | ROW62 | 4891 | 8.4952 | 1 | 0 | 1 | 0 |
| **63** | SCORE | ROW63 | 246 | 5.5053 | 1 | 0 | 1 | 0 |
| **64** | SCORE | ROW64 | 4369 | 8.3823 | 1 | 0 | 1 | 0 |
| **65** | SCORE | ROW65 | 2944 | 7.9875 | 1 | 0 | 1 | 0 |
| **66** | SCORE | ROW66 | 16631 | 9.7190 | 1 | 0 | 1 | 0 |
| **67** | SCORE | ROW67 | 170044 | 12.0438 | 1 | 0 | 1 | 0 |
| **68** | SCORE | ROW68 | 3621 | 8.1945 | 1 | 0 | 1 | 0 |
| **69** | SCORE | ROW69 | 6848 | 8.8317 | 1 | 0 | 1 | 0 |
| **70** | SCORE | ROW70 | 6209 | 8.7338 | 1 | 0 | 1 | 0 |
| **71** | SCORE | ROW71 | 29374 | 10.2879 | 1 | 0 | 1 | 0 |
| **72** | SCORE | ROW72 | 38575 | 10.5604 | 1 | 0 | 1 | 0 |
| **73** | SCORE | ROW73 | 10585 | 9.2672 | 1 | 0 | 1 | 0 |
| **74** | SCORE | ROW74 | 62716 | 11.0464 | 1 | 0 | 1 | 0 |
| **75** | SCORE | ROW75 | 10506 | 9.2597 | 1 | 0 | 1 | 0 |
| **76** | SCORE | ROW76 | 49090 | 10.8014 | 1 | 0 | 1 | 0 |
| **77** | SCORE | ROW77 | 4085 | 8.3151 | 1 | 0 | 1 | 0 |
| **78** | SCORE | ROW78 | 65939 | 11.0965 | 1 | 0 | 1 | 0 |
| **79** | SCORE | ROW79 | 24501 | 10.1065 | 1 | 0 | 1 | 0 |
| **80** | SCORE | ROW80 | 6260 | 8.7419 | 1 | 0 | 1 | 0 |
| **81** | SCORE | ROW81 | 9898 | 9.2001 | 1 | 0 | 1 | 0 |
| **82** | SCORE | ROW82 | 45649 | 10.7287 | 1 | 0 | 1 | 0 |
| **83** | SCORE | ROW83 | 10294 | 9.2393 | 1 | 0 | 1 | 0 |
| **84** | SCORE | ROW84 | 20299 | 9.9183 | 1 | 0 | 1 | 0 |
| **85** | SCORE | ROW85 | 12957 | 9.4694 | 1 | 0 | 1 | 0 |
| **86** | SCORE | ROW86 | 9059 | 9.1115 | 1 | 0 | 1 | 0 |
| **87** | SCORE | ROW87 | 20201 | 9.9135 | 1 | 0 | 1 | 0 |
| **88** | SCORE | ROW88 | 9383 | 9.1467 | 1 | 0 | 1 | 0 |
| **89** | SCORE | ROW89 | 66692 | 11.1078 | 1 | 0 | 1 | 0 |
| **90** | SCORE | ROW90 | 6391 | 8.7626 | 1 | 0 | 1 | 0 |
| **91** | SCORE | ROW91 | 10639 | 9.2723 | 1 | 0 | 1 | 0 |
| **92** | SCORE | ROW92 | 72311 | 11.1887 | 1 | 0 | 1 | 0 |
| **93** | SCORE | ROW93 | 45648 | 10.7287 | 1 | 0 | 1 | 0 |
| **94** | SCORE | ROW94 | 33149 | 10.4088 | 1 | 0 | 1 | 0 |
| **95** | SCORE | ROW95 | 27739 | 10.2306 | 1 | 0 | 1 | 0 |
| **96** | SCORE | ROW96 | 28996 | 10.2749 | 1 | 0 | 1 | 0 |
| **97** | SCORE | ROW97 | 88358 | 11.3892 | 1 | 0 | 1 | 0 |
| **98** | SCORE | ROW98 | 23592 | 10.0687 | 1 | 0 | 1 | 0 |
| **99** | SCORE | ROW99 | 13918 | 9.5409 | 1 | 0 | 1 | 0 |
| **100** | SCORE | ROW100 | 13974 | 9.5450 | 1 | 0 | 1 | 0 |

No he conseguido utilizer el tdatos, me sale error.

**proc** **discrim** data=tdatos method=normal pool=test wcov pcov list crossvalidate crosslist outstat=salida;

class \_NAME\_;

**run**;

**proc** **discrim** data=tdatos pool=test testlisterr crossvalidate;

class \_NAME\_;

**run**;

**Observación importante:**

En esta etapa lo correcto era crear las variables logarítmicas y quitar los datos atípicos.

No he conseguido hacer estas dos etapas entonces los próximos pasos voy hacer con las variables normalizadas y con los datos atípicos.

En el final intentare otra vez hacer con variables logarítmicas.

1. **Test de igualdad de varianzas**

Con el código SAS:

**proc** **discrim** data=cluster4 pool=test testlisterr crossvalidate;

class CLUSTER;

VAR POBL NATALIDA ESPERANZ MORTALID;

**run**;

The DISCRIM Procedure

Test of Homogeneity of Within Covariance Matrices

| **Chi-Square** | **DF** | **Pr > ChiSq** |
| --- | --- | --- |
| 142.449019 | 30 | <.0001 |

Se hace el test de igualdad de varianzas y como el p-valor ha dado es < .0001 se rechaza la hipótesis de clasificación lineal. Es regla cuadrática.

|  |
| --- |
| The SAS System |

The DISCRIM Procedure

| **Generalized Squared Distance to CLUSTER** | | | | |
| --- | --- | --- | --- | --- |
| **From CLUSTER** | **1** | **2** | **3** | **4** |
| **1** | -7.83137 | 14.08659 | 21.65840 | 299.00354 |
| **2** | 150.04373 | -9.63065 | 30.23784 | 64.24026 |
| **3** | 66.85171 | 2381 | -6.58306 | 20.30043 |
| **4** | 384.58744 | 4097 | 12.03582 | -7.92040 |

Abajo la tabla de error:

|  |
| --- |
| The SAS System |

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.CLUSTER4

Resubstitution Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **From CLUSTER** | **1** | **2** | **3** | **4** | **Total** |
| **1** | |  | | --- | | 28 | | 96.55 | | |  | | --- | | 1 | | 3.45 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 29 | | 100.00 | |
| **2** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 5 | | 100.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 5 | | 100.00 | |
| **3** | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 35 | | 89.74 | | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 39 | | 100.00 | |
| **4** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 3.70 | | |  | | --- | | 26 | | 96.30 | | |  | | --- | | 27 | | 100.00 | |
| **Total** | |  | | --- | | 30 | | 30.00 | | |  | | --- | | 6 | | 6.00 | | |  | | --- | | 36 | | 36.00 | | |  | | --- | | 28 | | 28.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **Total** |
| **Rate** | 0.0345 | 0.0000 | 0.1026 | 0.0370 | 0.0435 |
| **Priors** | 0.2500 | 0.2500 | 0.2500 | 0.2500 |  |

|  |
| --- |
| The SAS System |

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.CLUSTER4

Cross-validation Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **From CLUSTER** | **1** | **2** | **3** | **4** | **Total** |
| **1** | |  | | --- | | 28 | | 96.55 | | |  | | --- | | 1 | | 3.45 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 29 | | 100.00 | |
| **2** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 4 | | 80.00 | | |  | | --- | | 1 | | 20.00 | | |  | | --- | | 5 | | 100.00 | |
| **3** | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 35 | | 89.74 | | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 39 | | 100.00 | |
| **4** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 3.70 | | |  | | --- | | 1 | | 3.70 | | |  | | --- | | 25 | | 92.59 | | |  | | --- | | 27 | | 100.00 | |
| **Total** | |  | | --- | | 30 | | 30.00 | | |  | | --- | | 2 | | 2.00 | | |  | | --- | | 40 | | 40.00 | | |  | | --- | | 28 | | 28.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **Total** |
| **Rate** | 0.0345 | 1.0000 | 0.1026 | 0.0741 | 0.3028 |
| **Priors** | 0.2500 | 0.2500 | 0.2500 | 0.2500 |  |

1. **Transformaciones de las variables con logaritmos**

Después de mucho pensar y leer, he hecho el merge de los datos originales con los datos clasificados por clusters.

Teniendo las variables originales fue posible hacer el logaritmo.

He probado los dos datasets que he generado con y sin logaritmo y es muy interesante pensar que poner el logaritmo quita muchos de los datos atípicos.

Abajo una parte del código creados:

**data** cluster\_pequeno;

set cluster4id;

DROP POBL NATALIDA ESPERANZ MORTALID;

**run**;

**proc** **print** data=cluster\_pequeno;

**run**;

**data** sample\_paises\_con\_clusters;

merge sample\_paises cluster\_pequeno;

**run**;

**proc** **print** data=sample\_paises\_con\_clusters;

**run**;

**proc** **print** data=sample\_paises;

**run**;

**data** logcluster;

set sample\_paises\_con\_clusters;

logESPERANZ = log(ESPERANZ);

logNATALIDA = log(NATALIDA);

logPOBL = log(POBL);

logMORTALID = log(MORTALID);

**run**;

**proc** **print** data=logcluster;

**run**;

**proc** **univariate** data=sample\_paises\_con\_clusters normal plot;

VAR POBL NATALIDA ESPERANZ MORTALID;

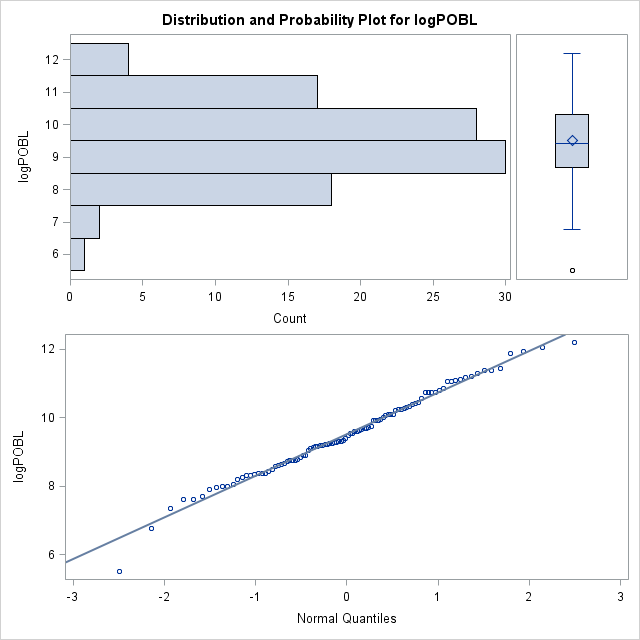
**run**;

**proc** **univariate** data=logcluster normal plot;

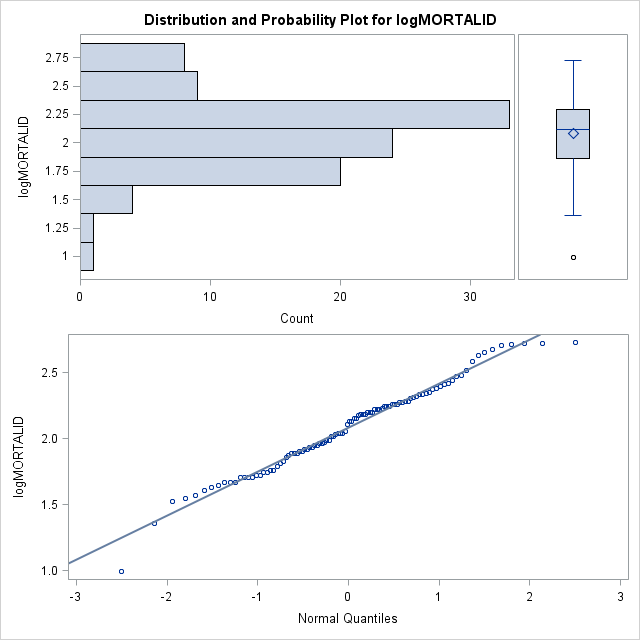
var logESPERANZ logNATALIDA logMORTALID logPOBL;

**run**;

Se puede ver que las variables Mortalidad y población con logaritmo se quitan muhos de los datos atípicos.



| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.989007 | **Pr < W** | 0.5851 |
| **Kolmogorov-Smirnov** | **D** | 0.050504 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.029312 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.221744 | **Pr > A-Sq** | >0.2500 |



| **Tests for Normality** | | | | |
| --- | --- | --- | --- | --- |
| **Test** | **Statistic** | | **p Value** | |
| **Shapiro-Wilk** | **W** | 0.981577 | **Pr < W** | 0.1761 |
| **Kolmogorov-Smirnov** | **D** | 0.072183 | **Pr > D** | >0.1500 |
| **Cramer-von Mises** | **W-Sq** | 0.060298 | **Pr > W-Sq** | >0.2500 |
| **Anderson-Darling** | **A-Sq** | 0.410063 | **Pr > A-Sq** | >0.2500 |

**Hacemos el test de la hipótesis de la normalidad con todas las variables con logaritimos.**

Tabla para ayudar la interpretación

|  |  |  |
| --- | --- | --- |
| P-valor Bajo | Menor que 0.01% | No es normal |
| P-valor Alto | Más grande que 0.01% | Normal |

**Los resultados con las variables POBL NATALIDA ESPERANZ MORTALID CLUSTER con logaritimo.**

|  |  |
| --- | --- |
| logPOBL | Alto = Normal |
| logNATALIDA | Bajo = No es normal |
| logESPERANZ | Bajo = No es normal |
| logMORTALID | Alto = Normal |
| logCLUSTER | Bajo = No es normal |

1. **Comprobación de la existencia de datos atípicos con logaritimo.**

**Los resultados con las variables POBL NATALIDA ESPERANZ MORTALID CLUSTER**

|  |  |
| --- | --- |
| logPOBL | Tiene datos atípicos muy pocos |
| logNATALIDA | No tiene datos atípicos |
| logESPERANZ | No tiene datos atípicos |
| logMORTALID | Tiene datos atípicos muy pocos |
| logCLUSTER | No tiene datos atípicos |

**Test de igualdad de varianzas**

Con el código SAS:

**proc** **discrim** data=logcluster pool=test testlisterr crossvalidate;

var logESPERANZ logNATALIDA logMORTALID logPOBL;

class CLUSTER;

**run**;

Ahora:

The DISCRIM Procedure

Test of Homogeneity of Within Covariance Matrices

| **Chi-Square** | **DF** | **Pr > ChiSq** |
| --- | --- | --- |
| 104.557467 | 30 | <.0001 |

Se hace el test de igualdad de varianzas y como el p-valor ha dado es < .0001 se rechaza la hipótesis de clasificación lineal. Es regla cuadrática.

Anterior:

The DISCRIM Procedure

Test of Homogeneity of Within Covariance Matrices

| **Chi-Square** | **DF** | **Pr > ChiSq** |
| --- | --- | --- |
| 142.449019 | 30 | <.0001 |

Se hace el test de igualdad de varianzas y como el p-valor ha dado es < .0001 se rechaza la hipótesis de clasificación lineal. Es regla cuadrática.

|  |
| --- |
| The SAS System |

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.LOGCLUSTER

Resubstitution Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **From CLUSTER** | **1** | **2** | **3** | **4** | **Total** |
| **1** | |  | | --- | | 29 | | 100.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 29 | | 100.00 | |
| **2** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 5 | | 100.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 5 | | 100.00 | |
| **3** | |  | | --- | | 1 | | 2.56 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 36 | | 92.31 | | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 39 | | 100.00 | |
| **4** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 3.70 | | |  | | --- | | 26 | | 96.30 | | |  | | --- | | 27 | | 100.00 | |
| **Total** | |  | | --- | | 30 | | 30.00 | | |  | | --- | | 5 | | 5.00 | | |  | | --- | | 37 | | 37.00 | | |  | | --- | | 28 | | 28.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **Total** |
| **Rate** | 0.0000 | 0.0000 | 0.0769 | 0.0370 | 0.0285 |
| **Priors** | 0.2500 | 0.2500 | 0.2500 | 0.2500 |  |

|  |
| --- |
| The SAS System |

The DISCRIM Procedure

Classification Summary for Calibration Data: WORK.LOGCLUSTER

Cross-validation Summary using Quadratic Discriminant Function

| **Number of Observations and Percent Classified into CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
| **From CLUSTER** | **1** | **2** | **3** | **4** | **Total** |
| **1** | |  | | --- | | 27 | | 93.10 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 2 | | 6.90 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 29 | | 100.00 | |
| **2** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 4 | | 80.00 | | |  | | --- | | 1 | | 20.00 | | |  | | --- | | 5 | | 100.00 | |
| **3** | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 1 | | 2.56 | | |  | | --- | | 34 | | 87.18 | | |  | | --- | | 2 | | 5.13 | | |  | | --- | | 39 | | 100.00 | |
| **4** | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 0 | | 0.00 | | |  | | --- | | 1 | | 3.70 | | |  | | --- | | 26 | | 96.30 | | |  | | --- | | 27 | | 100.00 | |
| **Total** | |  | | --- | | 29 | | 29.00 | | |  | | --- | | 1 | | 1.00 | | |  | | --- | | 41 | | 41.00 | | |  | | --- | | 29 | | 29.00 | | |  | | --- | | 100 | | 100.00 | |
| **Priors** | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | | 0.25 | |  | | |  | | --- | |  | |  | |

| **Error Count Estimates for CLUSTER** | | | | | |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **Total** |
| **Rate** | 0.0690 | 1.0000 | 0.1282 | 0.0370 | 0.3086 |
| **Priors** | 0.2500 | 0.2500 | 0.2500 | 0.2500 |  |