Taller 1: Análisis Multivariado

Introducción y graficas para el análisis mulivariado

Julián Camilo Riaño Moreno

jueves, junio 18, 2020

Contents

Actividad #1

> Busque un conjunto de datos multivariados relacionados con un área de conocimiento de su dominio que tenga al menos 50 observaciones, 5 variables numéricas y 3 categóricas.

Especificaciones sobre la base de datos utilizada.

Para la elaboración de las actividades propuestas en el curso de análisis multivariado, se decidió utiliza una base de datos que se desea utilizar para un trabajo de investigación en marco de la *especialización en estadística aplicada*.

Los datos para estos análisis son obtendios de la iniciativa *Data from the ECDC Surveillance Atlas* para seguimiento de la resistencia a antibióticos en la unión europea llevado acabo por el *European Centre for Disease Prevention and Control*. Esta información se puede descargar desde el siguiente link: (https://atlas.ecdc.europa.eu/public/index.aspx?Dataset=27&HealthTopic=4).

La base de datos original cuenta con 67542 observaciones y 9 variables (*HealthTopic; Population; Indicator; Unit; Time; RegionCode; RegionName; NumValue; TxtValue*).

A través de una limpieza exhaustiva y reorganización de esta base de datos, se llegó a obtener una base de datos de 3478 observaciones y 72 variables, estas 72 variables corresponde a al tipo de *Indicador* por *Antibiótico*, por ejemplo: nombre de la columna o variable *Vancomycin_r_percentage*, esto corresponde al porcentaje de resistencia a la vancomicina. Estas variables son para cara cada una de las 8 bacterias reportadas en el informe (*Acinetobacter spp; Enterococcus faecalis; Enterococcus faecium; Escherichia coli; Klebsiella pneumoniae; Pseudomonas aeruginosa; Staphylococcus aureus; Streptococcus pneumoniae*) y para los 30 países involucrados (*Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta,*

Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, United Kingdom); esta información ha sido recolectada en los años 2000 - 2018.

Se realizó un reanálisis de la base de datos, se eliminaron las variables que no eran informativas como por ejemplo las que se referían al número de aislados en cada una de estas; se eliminaron sin información (*NAs*); únicamente se dejaron las variables que correspondían al % de resistencia para cada uno de los antibióticos reportados para cada una de las bacterias[1].

A continuación, se decidió seleccionar únicamente dos bacterias en las que fueron evaluados los perfiles de resistencia a los mismos antibióticos, a saber: Escherichia coli; Klebsiella pneumoniae. Las cuales además registran los mayores números de observaciones (después de la remoción de *NAs*), 460 y 384 respectivamente. La variables de esta nueva base de datos son : Bacteria, Year, Code_C, Country, Aminoglycosides, Carbapenems, Fluoroquinolones, cephalos_3er_gen.

Adicionalmente se realizón un *web scrapping* de la página: ('https://www.worldatlas.com/articles/the-four-european-regions-as-defined-by-the-united-nations-geoscheme-for-europe.html'), para obtener las regiones del continente europeo al que corresponde del que se obtuvieron observaciones. Se realizó un *match* por la variable `Country` y se adicionó una nueva variable llamada `Region` a la base de datos final. Las especificaciones de las variables de la base datos final se encuentran en la tabla 1.

Table 1: Organizacion de las variables de la base de datos

Nombre de variable	Definicion	Descripción	Unidad	Tipo de variable
Bacteria	Especie de la bacteria	Escherichia coli o Klebsiella pneumoniae	—	Categorica nominal
Year	Ano	2000-2018	Entero	Categorica nominal
Code_C	Codigo del pais	—	—	Categorica nominal
Country	Nombre del pais	—	—	Categorica nominal
Region	Nombre de la region	Eastern o Northern o Southern o Western	—	Categorica nominal
Aminoglycosides	Porcentaje de resistencia	—	%	Cuantitativa continua
Carbapenems	Porcentaje de resistencia	—	%	Cuantitativa continua
Fluoroquinolones	Porcentaje de resistencia	—	%	Cuantitativa continua
cephalos_3er_gen	Porcentaje de resistencia	—	%	Cuantitativa continua

Primer pregunta:

> Encuentre el vector de medias $\bar{x}$, la matriz de covarianzas S y la matriz de correlaciones R de las variables numéricas. Interprete los resultados.

La tabla 2. muestra el vector de medias para las variables de estudio.

[1]para tratar cada bacteria se utilizan diferentes antibióticos, de manera que los % de resistencia reportados en esta base de datosn son algunos diferentes (algunos son los mismos) entre bacterias.

Table 2: Vector de medias

Aminoglycosides	Carbapenems	Fluoroquinolones	cephalos_3er_gen
0.1662	0.01995	0.2377	0.1916

Table 3: Matriz de correlación entre las variables

	Aminoglycosides	Carbapenems	Fluoroquinolones	cephalos_3er_gen
Aminoglycosides	1	0.435	0.837	0.962
Carbapenems	0.435	1	0.489	0.498
Fluoroquinolones	0.837	0.489	1	0.859
cephalos_3er_gen	0.962	0.498	0.859	1

Table 4: Matriz de covarianza entre las variables

	Aminoglycosides	Carbapenems	Fluoroquinolones	cephalos_3er_gen
Aminoglycosides	0.027	0.006	0.022	0.03
Carbapenems	0.006	0.006	0.006	0.008
Fluoroquinolones	0.022	0.006	0.025	0.026
cephalos_3er_gen	0.03	0.008	0.026	0.037

Segunda pregunta:

> Encuentre la varianza muestral generalizada $|S|$ y la varianza muestral total $tr(S)$. Comente sus resultados.

La varianza muestral generalizada se obtiene a través de la aplicación de la función `det`(determinante) a la matriz de covarianzas, con le siguiente resultado:

$$|S| = -1e^{-08}$$

El resultado de $|S|$ es muy cercano a 0, lo que indica que cada variable es una combinación lineal de las otras, por lo tanto, las variables son linealmente dependientes, es decir la independencia entre las variables es casi nula (esto hace que los datos puedan ser estudiados a través de análisis como PCA).

Seguidamente, Se obtiene la varianza muestral total a través de la sumatoria de la diagonal de la matriz de varianzas, obteniendo el siguiente resultado:

$$tr|S| = 0.95$$

Esto indica que la variabilidad en los datos es mínima ya que el valor de $tr|S|$ es cercano 0, es decir, los datos son cercanos a la media.

Tercera pregunta:

> Construya las siguientes gráficas para los datos:

(a) una matriz de diagramas de dispersión. La figura 1. corresponde a una gráfica de dispersión y correlación para las variables dadas.

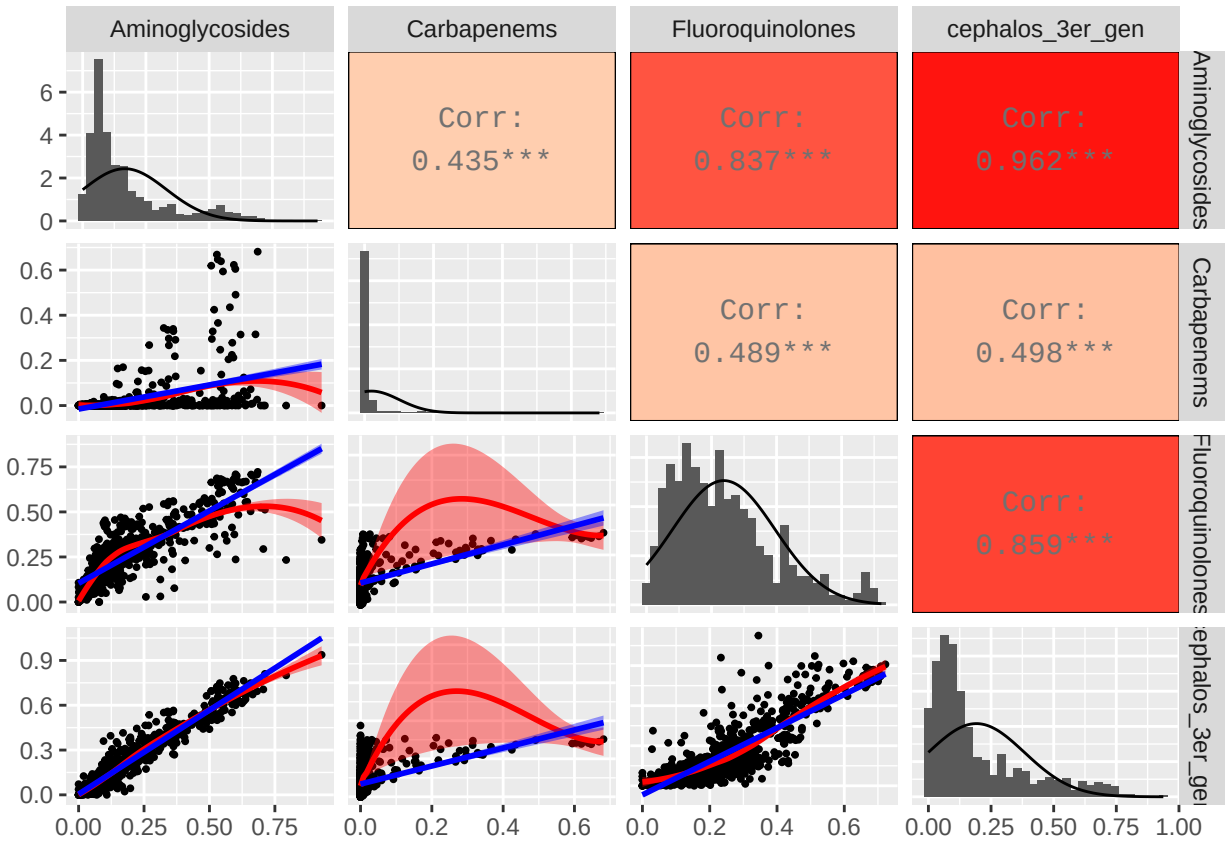

Figure 1: Gráfica de dispersión y correlaciones

El resultado de la figura 1. da información valiosa para el desarrollo del resto de las actividades. En primera medida como se puede observar y anticipar las variables de estudio no siguen una distribución normal, e incluso es posible concluir que la variable `Carbapenems` es la menos informativa ya que exhibe la menor correlación con las demás; en las regresiones lineales realizadas entre variables (líneas azules: método `lm` y rojas: metodo local `loess`, en las dispersiones de la cuadricula inferior) se observa una pobre relación lineal. Inclusive se evidencia poca dispersión de los datos de esta variable. Esto indica que las bacterias estudiadas cuando son resistentes a los Carbapenems, necesariamente no son resistentes a otros antibióticos.

Diferente a lo observado para las cefalosporinas de 3er generación. Como se puede observar en la figura 1. existe una alta correlación y relación lineal entre la variables `Cephalos_3er_gen` y `Aminoglycosides` (Corr = 0.962) y `Cephalos_3er_gen` y `Fluoroquinolones` (Corr = 0.837). Lo que indica que cuando hay resistencia a Cefalosporia de 3er generación también se espera que se de resistencia a Aminoglicosidos y Fluoroquinolonas.

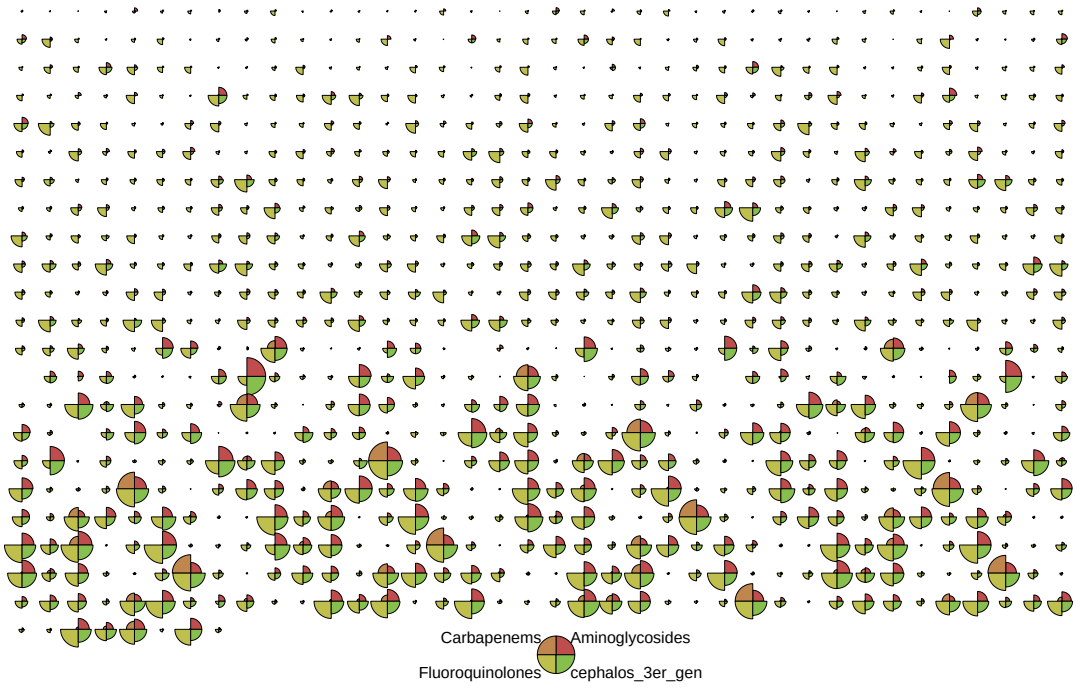

Figure 2: Diagrama de estrellas para las variables estudiadas

(b) diagrama de estrellas La figura 2. muestra por cada observación el peso que tiene cada variable a través de las proporciones mostradas en los segmentos que conforman la estrella. Como se puede observar, aunque con la cantidad de datos que se están estudiando es difícil visualizar el diagrama adecuadamente, las observaciones donde el porcentaje de resistencia es mayor (radio de la estrella mayor), se evidencia el poco peso de la variable `Carbapenems` esto muy problemente está dado a la baja resistencia encontrada para la mayor parte de las observaciones en este estudio.

(c) caras de Chernoff Para elaborar este diagrama se realizó una selección al azar de 100 observaciones, esto con el fin de poder visualizar y especificar mejor cada uno de los rasgos de la caras. Esta estratégia es poco eficiente cuando la cantidad de datos es alta.

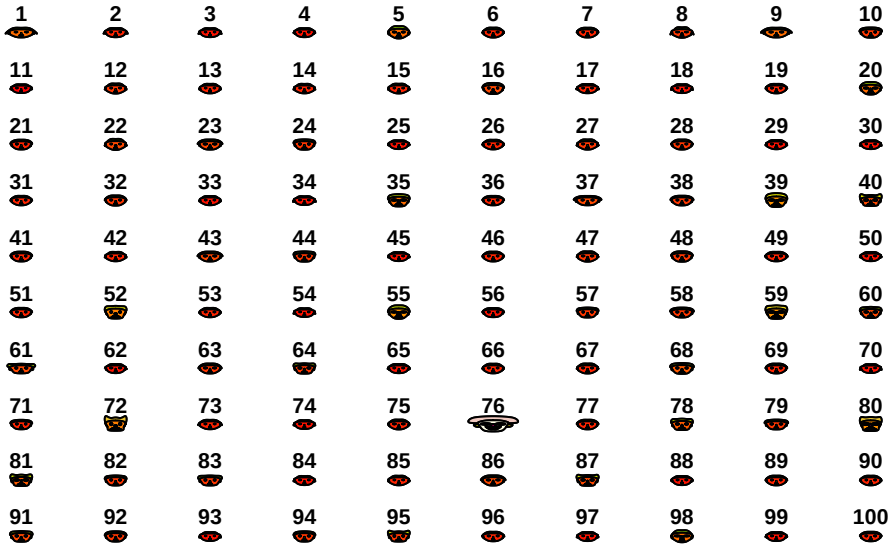

Figure 3: Diagrama de caras de Chernoff

```
## effect of variables:
##  modified item       Var
##  "height of face   " "Aminoglycosides"
##  "width of face    " "Carbapenems"
##  "structure of face" "Fluoroquinolones"
##  "height of mouth  " "cephalos_3er_gen"
##  "width of mouth   " "Aminoglycosides"
##  "smiling          " "Carbapenems"
##  "height of eyes   " "Fluoroquinolones"
##  "width of eyes    " "cephalos_3er_gen"
##  "height of hair   " "Aminoglycosides"
##  "width of hair    " "Carbapenems"
##  "style of hair    " "Fluoroquinolones"
##  "height of nose   " "cephalos_3er_gen"
##  "width of nose    " "Aminoglycosides"
##  "width of ear     " "Carbapenems"
##  "height of ear    " "Fluoroquinolones"
```

Como se puede ver en la figura 3. y la leyenda que la acompaña, a cada variable se le asigna un conjunto de rasgos, y se evalua para cada una de las observaciones el impacto que tiene cada variable. En este análisis llama la atención la cara de la observación número 76, esta cara es ancha y el pelo es ancho y de estilo particular; esto indica las variables `Carbapenems` y `Fluoroquinolonas` ('width of face' lo que corresponde

a `Carbapenems`, 'width of hair' que corresponde también a `Carbapenems` y 'style of hair' que corresponde a `Fluoroquinolones`) son las más representativas en esta observación.

(d) curvas de Andrews Para este gráfico se definió evaluar la consistencia y correlaciones de las variables númericas dadas con relación a una variable categórica `Year`. Se definió esta variable como `factor` y se incluyo en el análisis.

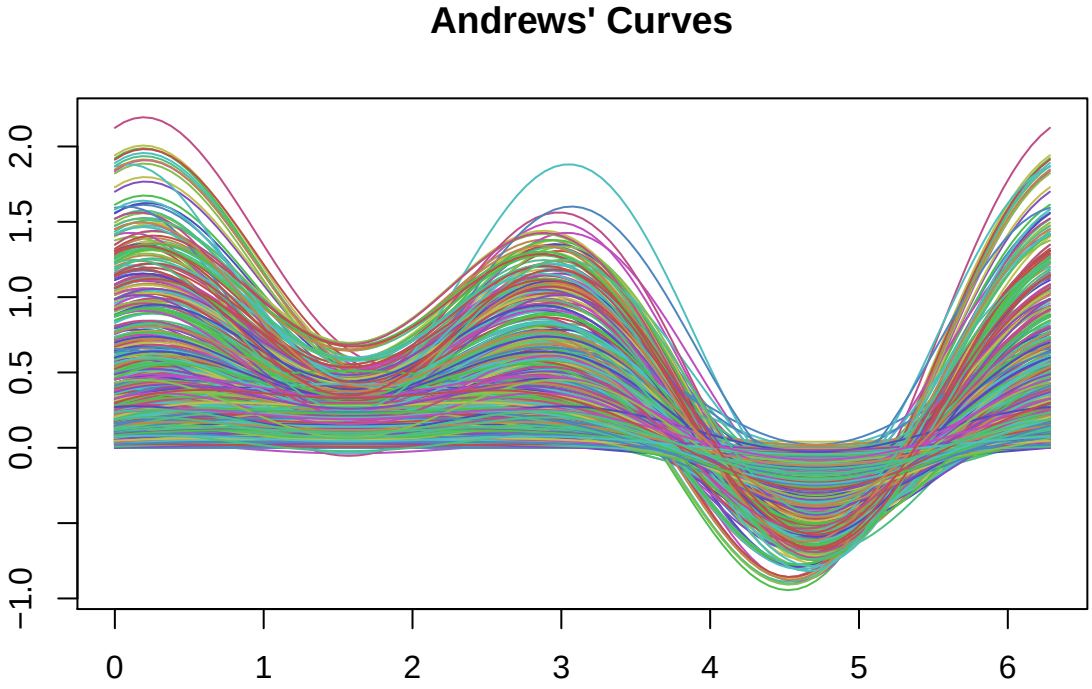

Figure 4: Curvas de Andrew

En la figura 4. se observa que la dimesionalidad de las curvas esparcialmente constante. Algunas se observaciones se comportan más planas y otras más onduladas lo que corresponde a observaciones que tiene dimensionalidades con relación al los años distintas al conjunto de datos.

(e) otros gráficos apropiados Además se eleaboraron gráficos tipo "boxplots" para cada una de las variables que corresponden a los porcentajes de resistencia a determinado antibiótico; esto se analizó por año, tipo de bacteria y region.

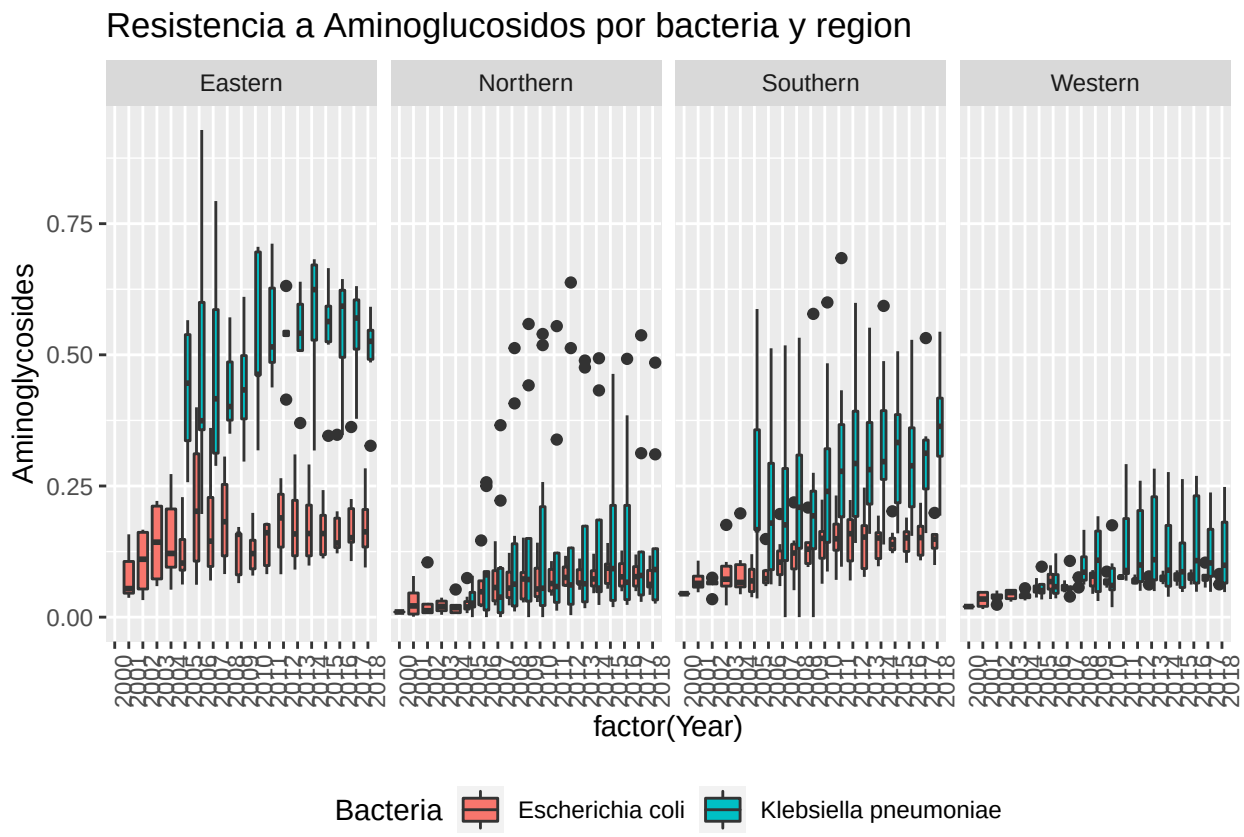

Figure 5: Boxplot 1: resistencia a Aminoglucosidos por bacteria y region en tiempo

El Boxplot 1. corresponde a al análisis de la resistencia de las dos bacterias seleccionadas para este ejercicio (*Escherichia coli; Klebsiella pneumoniae*) al grupo de antibióticos conocido como "aminoglucosidos". Para este grupos evidencia un incremento leve de la resistencia a los antbióticos de los dos tipos de bacteria en el tiempo. La bacteria que exhibe mayores porcentajes y varianza en la resistencia a los antibióticos es la *Klebsiella pneumoniae*, esto es particularmente relevante para las regiones del este y sur de europa (Èastern,Southern'). Llama la atención que en estas regiones el incremento de los resportes de resistencia a aminoglucosidos incremento en el 2005, esto puede estar correlacionado con mayor reporte de niveles de resistencia y mayor conciencia de la problemática mundial con respecto a la resistencia a antibióticos.

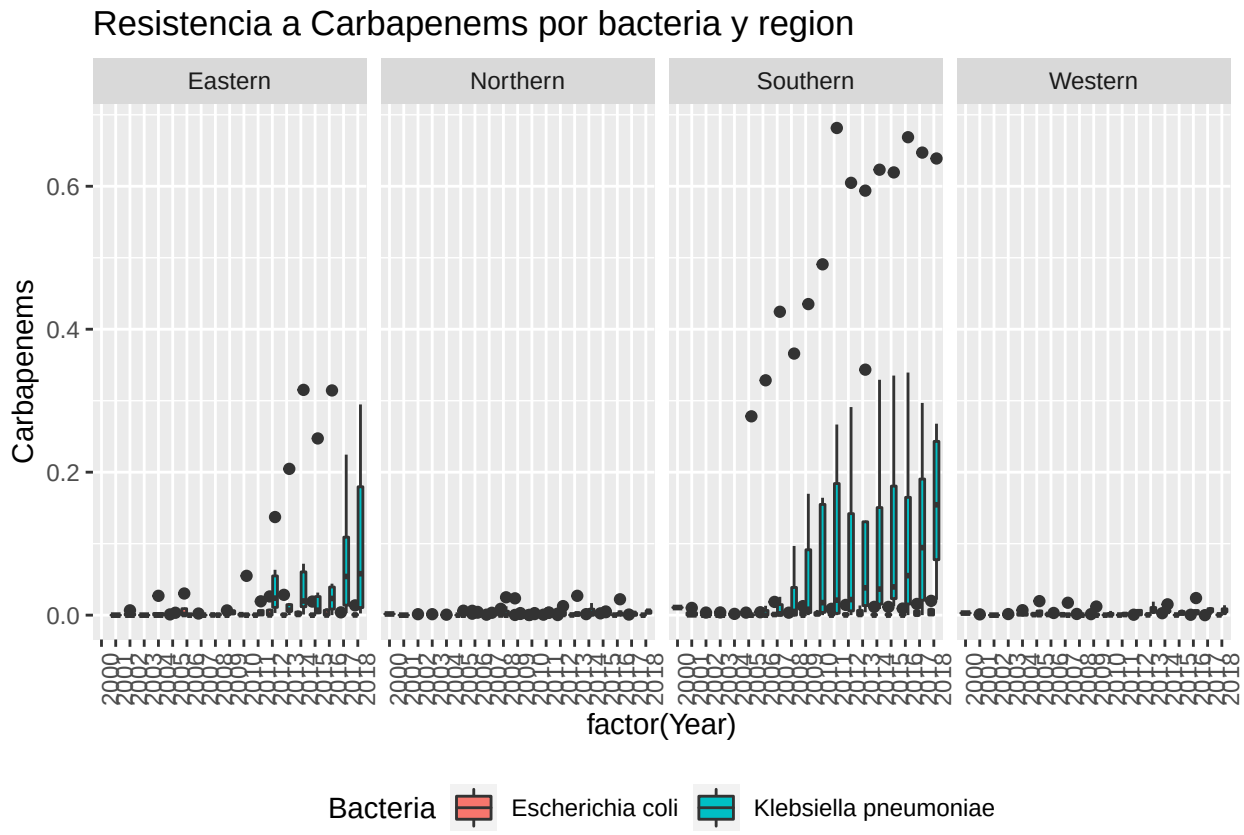

Figure 6: Boxplot 2: resistencia a Carbapenems por bacteria y region en tiempo

El Boxplot 2. que se encuentra gráficado en la figura 6. resulta particularmente interesante, dado que muestra la baja frecuencia de resistencia al grupo de antibióticos conocido como `Carbapenems`. Al igual que con los Animoglucosidos, la bacteria que reporta mayor resistencia es la *Klebsiella pneumoniae* para las regiones de este y sur de europa. En este caso el incremento de resistencia incia desde el año 2009. Esta información, asociada a las correlaciones mostradas en la figura 1. muestran el poco peso que tienen los Carbapenemicos en la resistencia total a los antibióticos y la resistencia combinada.

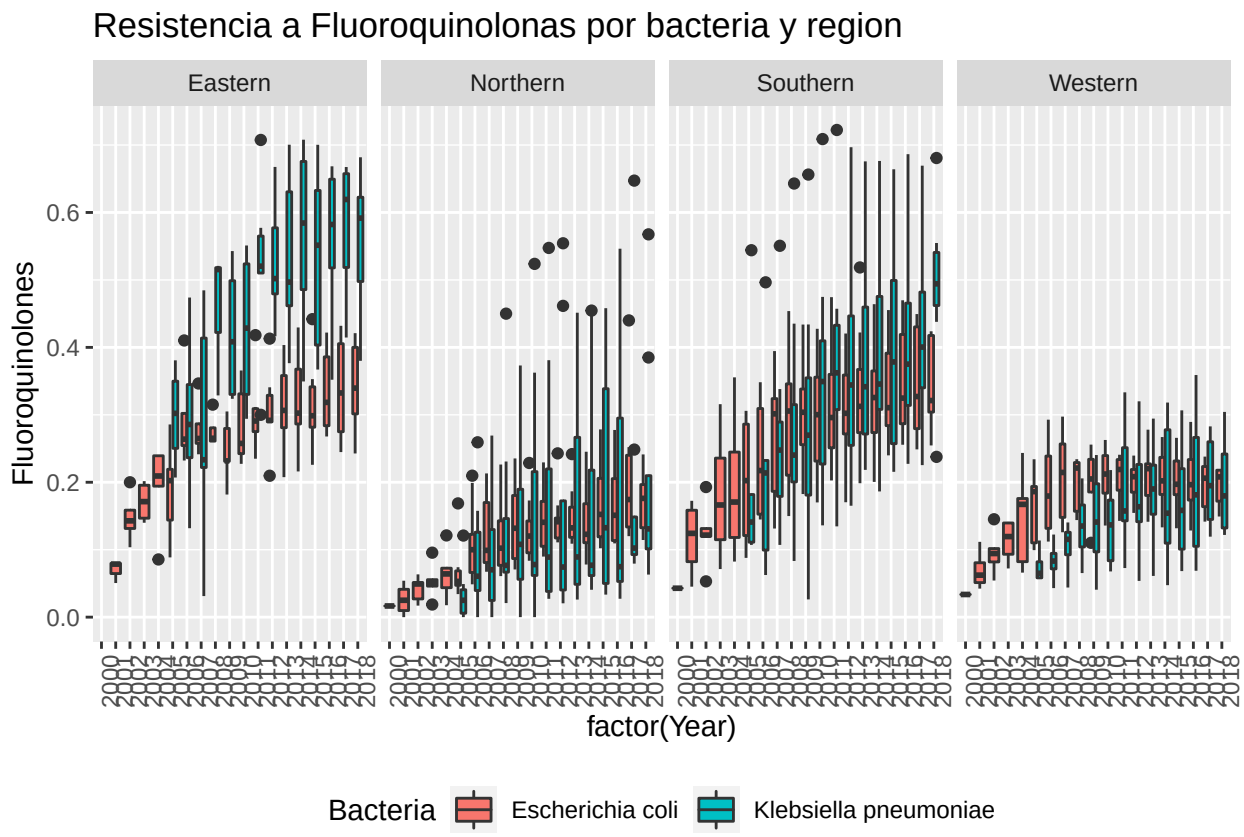

Figure 7: Boxplots 3: resistencia a Fluoroquinolonas por bacteria y region en tiempo

El boxplot número 4. representa el porcentaje de resistencia al grupo de antibióticos "Fluoroquinolonas". Respecto a este grupo, se observa un incremento en el tiempo de la resistencia en todas las regiones y para las dos bacterias. Las regiones del occidente de europa `western` se evidencia un incremento inicial y una estabilización de este comportamiento desde el 2009. Las regiones de sur y este de europa se ve el mayor incremento y variabilidad en la resistencia especialmente en la bacteria *Klebsiella pneumoniae*, desde el 2005. Que al igual a los casos anteriores puede estar dada al incremento y frecuencia del reporte.

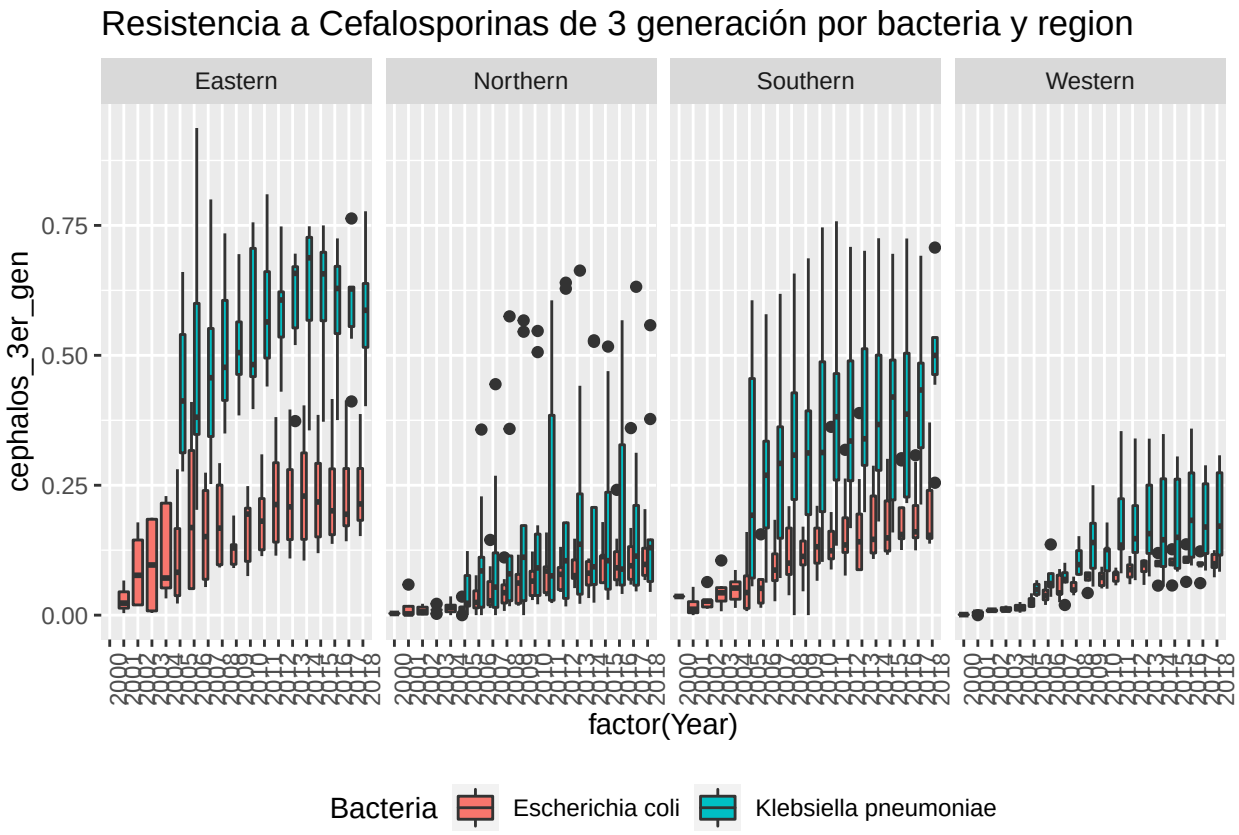

Figure 8: Boxplot 4: resistencia a Cefalosporinas de 3 generación por bacteria y region en tiempo

Finalmente la figura 8 muestra la resistencia a las ceaflosporinas de 3er generación. Para estas el incremento de la resistencia presenta un incremento sostenido en el tiempo. De la misma forma la bacteria *Klebsiella pneumoniae* es la que presenta mayores niveles y variabilidad de la resistencia, especialmente en la región del sur y este de europa, siendo mayor en esta última desde el 2005.

Cuarta pregunta:

Investigue si su conjunto de datos proviene de una distribución normal multivariada.

Para comprar la normalidad de los datos se decidió comprobar a través de un test de Shapiro-Wilk multivariado, el cual permite constatar la normalidad del conjunto de observaciones, y se formula de la siguiente manera (donde N indica una distribución normal:

$$H_o : el\ conjunto\ de\ datos = N$$

$$H_1 : el\ conjunto\ de\ datos \neq N$$

Table 5: Test de shapiro-wilk (w) para los datos

W	$p - value$
0.469	2.489e-44

El resultado obtenido de este estadístico es presentado en la tabla 5. lo que arroja un valor de $W = 0.469$ significativo ($p - value < 0.05$), por tal razón, se rechaza la hipotesis nula, así que se puede aseverar los datos no siguen una distribución normal (N).