Actividad 5.2 Componentes Principales

Integrantes

A01068244 - Jared Andrés Silva Villa
A00227869 - Paola Félix Torres

Fecha: 27 de Octubre del 2024

ÍNDICE

Análisis de Regresión Lineal Múltiple

Ecuación de regresión

gdpp = -41934 + 66.6 child_mort + 28.5 exports + 1549 health - 28.1 imports + 0.7856 income - 100.5 inflation + 389 life_expec + 615 total_fer

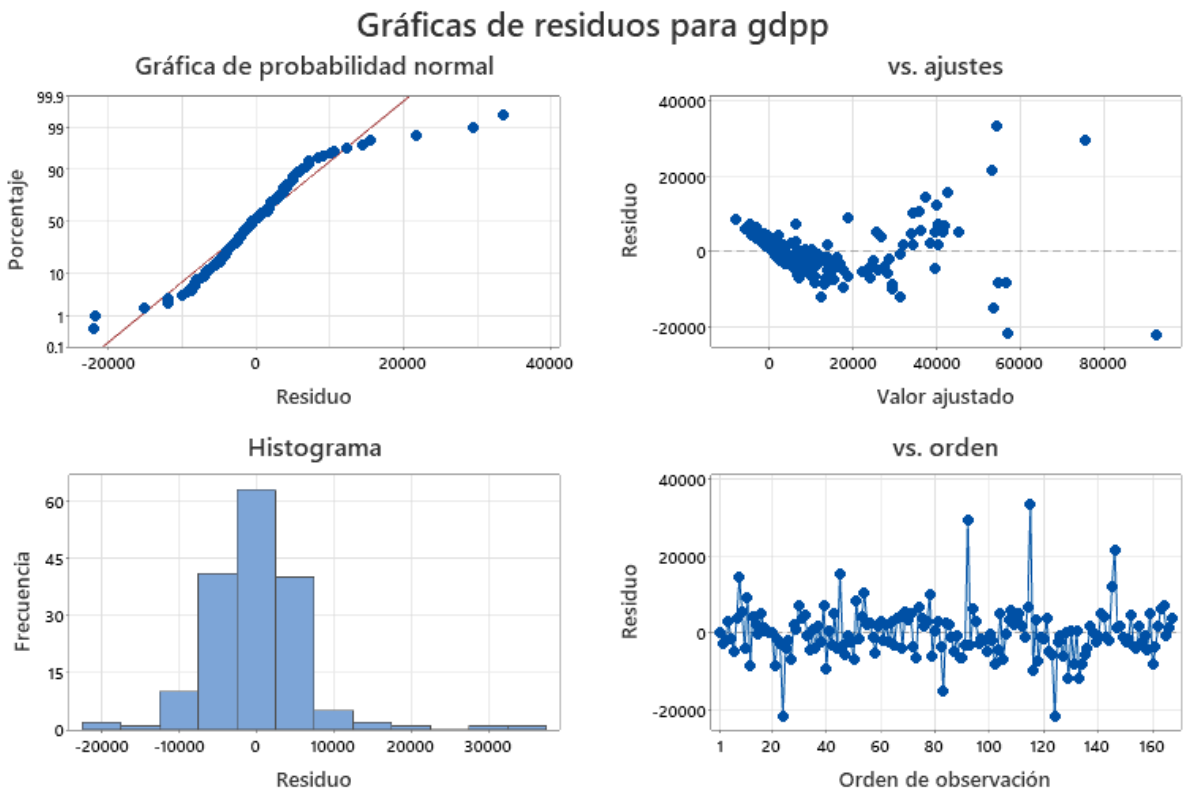
Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-41934	11130	-3.77	0.000	
child_mort	66.6	35.5	1.87	0.063	7.21
exports	28.5	43.2	0.66	0.511	4.93
health	1549	227	6.82	0.000	1.37
imports	-28.1	42.5	-0.66	0.509	3.72
income	0.7856	0.0437	17.99	0.000	2.49
inflation	-100.5	56.7	-1.77	0.078	1.26
life_expec	389	143	2.72	0.007	5.68
total_fer	615	680	0.90	0.367	3.72

Resumen del modelo

S	R-cuadrado	R-cuadrado(ajustado)	R-cuadrado (pred)
6875.57	86.61%	85.93%	82.70%

Analizando los p-values se puede concluir que las variables de health, income y life_expec son predictores significativos en el modelo, mientras que los demás tienen una relación débil con gdpp. Asimismo, al analizar los FIV, se puede ver que child_mort y life_expec tienen valores un tanto elevados, lo cual puede significar que tienen un grado de correlación con otras variables.



Los residuos del modelo muestran un comportamiento cercano a la normalidad, con una distribución simétrica alrededor de cero en el histograma y alineación en la gráfica Q-Q, aunque con algunas desviaciones en los extremos que podrían indicar valores atípicos. La gráfica de residuos vs. ajustes sugiere posible heterocedasticidad, ya que los residuos no se distribuyen de forma constante, lo cual podría afectar la precisión de los intervalos de confianza. Finalmente, los residuos en el orden de observación no muestran patrones, lo que sugiere que se cumple el supuesto de independencia.

Análisis de Componentes Principales (PCA)

Análisis de los valores y vectores propios de la matriz de correlación

Valor propio	3.5746	1.5439	1.1634	0.7388	0.5622	0.2235	0.1085	0.0850
Proporción	0.447	0.193	0.145	0.092	0.070	0.028	0.014	0.011
Acumulada	0.447	0.640	0.785	0.878	0.948	0.976	0.989	1.000

Vectores propios

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
child_mort	-0.473	-0.214	0.100	-0.115	0.297	-0.203	0.135	0.748
exports	0.308	-0.608	-0.146	-0.102	0.058	0.053	0.696	-0.109
health	0.145	0.242	0.647	-0.680	-0.059	-0.014	0.183	-0.044
imports	0.195	-0.661	0.285	-0.056	-0.315	0.037	-0.569	0.125
income	0.387	-0.031	-0.248	-0.315	0.728	-0.179	-0.351	-0.054
inflation	-0.220	-0.006	-0.616	-0.621	-0.418	-0.064	-0.086	0.010
life_expec	0.464	0.237	-0.158	-0.004	-0.091	0.600	0.020	0.578
total_fer	-0.457	-0.177	0.051	-0.159	0.304	0.747	-0.090	-0.272

Los **valores propios** nos indican cuánta varianza es explicada por cada componente principal, podemos observar que nuestro componente PC1 tiene el valor de 3.57 por lo que significa que este explica **3.57 veces** más varianza que una variable original promedio.La **proporción** se puede explicar como el porcentaje de la varianza total que explica cada componente, nuestro componente PC1 explica el **44.7%** de la varianza total de nuestros datos.

La **varianza acumulada** nos muestra cuánta de la varianza total de los datos ha sido explicada al sumar los componentes, viendo nuestros datos podemos decir que PC1 + PC2 explican el 64.0% de la varianza.

Los vectores propios nos muestran cómo cada variable original contribuye a cada componente principal. Con nuestros resultados podemos observar que **life_expec** tiene un valor de **0.464** lo que nos indica que esta variable tiene un fuerte influencia en el primer componente.

Estos coeficientes determinan la **dirección del componente**, un coeficiente más alto significa que esa variable está más alineada con la dirección principal de la varianza

capturada por el componente.Podemos observar que **life_expec** y **icome** son las variables que más definen la dirección de ese componente PC1, lo que podría estar asociado con el bienestar y el desarrollo de los países.

Componente que explican al menos el 80% de la varianza total

Podemos darnos cuenta que **PC1,PC2 y PC3** llegan a obtener el 78.5%, por lo que tenemos que incluir **PC4** y nos daría un 87.7% y elegiremos los 4 primeros componentes.

Las ecuaciones los componentes principales con **4 variables** más importantes.

PC1 = −0.473 × child_mort + 0.464 × life_expec −0.457 × total_fer + 0.387 × income
PC2 = −0.661 × imports −0.608 × exports +0.242 × health +0.237 × life_expec
PC3 = 0.647 × health −0.616 × inflation +0.285 × imports −0.248 × income
PC4 = −0.680 × health −0.621 × inflation −0.315 × income −0.056 × imports

PC1 = Salud y Desarrollo Social.
PC2 = Comercio Exterior.
PC3 = PC3: Inflación y Salud.
PC4 = Economía y Salud.

Regresión con los componentes principales seleccionados.

Ecuación de regresión

gdpp = 12964 + 6726 PC_1 + 618 PC_2 - 883 PC_3 - 7516 PC_4

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	12964	897	14.46	0.000	
PC_1	6726	476	14.14	0.000	1.00
PC_2	618	724	0.85	0.394	1.00
PC_3	-883	834	-1.06	0.291	1.00
PC_4	-7516	1046	-7.18	0.000	1.00

Resumen del modelo

S	R-cuadrado	R-cuadrado(ajustado)	R-cuadrado (pred)
11586.1	61.00%	60.04%	54.84%

Podemos observar que tanto **PC_1** como **PC_4** son componentes relevantes, PC1 es positivamente correlacionado y PC4 negativamente correlacionado con GDPP, el modelo explica el 61% de la variabilidad en GDPP. PC3 y PC3 no tiene un impacto significativo en el modelo, por lo que sugiere que no son relevantes para predecir GDPP.

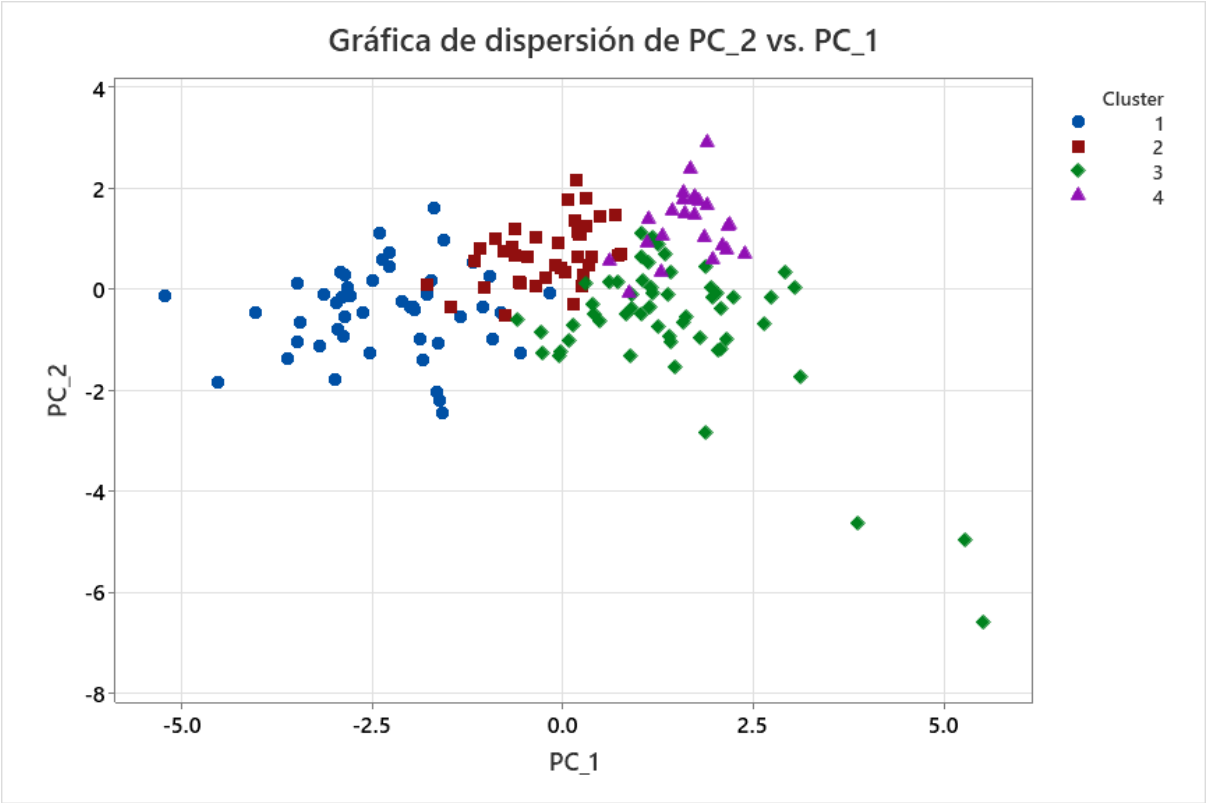
Comparación del modelo original contra el modelo de componentes.

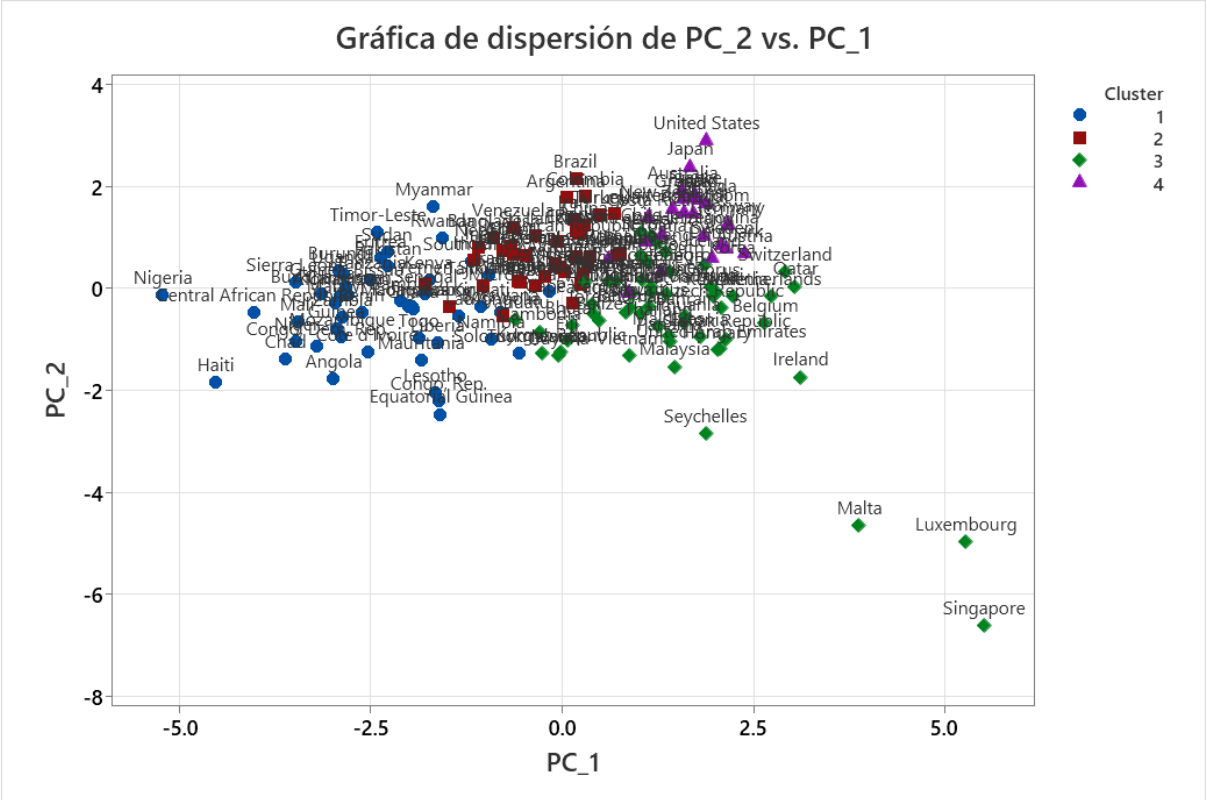
En cuanto al R-Cuadrado podemos darnos cuenta que el modelo original es superior con un **86%**, esto nos indica que explica mejor la varianza de gdp en comparación con el modelo de componente que tiene un r-cuadrado de **61%**.

Lo que podemos destacar del modelo de componentes es la **multicolinealidad**, donde podemos observar que el modelo original presenta problemas, como se refleja en sus **FIV** altos de algunas variables, en cuanto al modelo de componentes principales reduce este problema al **combinar variables correlacionadas** en componentes que tienen menos colinealidad.

Podemos concluir que el modelo original puede ser utilizado si buscas maximizar la explicación de la varianza, sin embargo si deseas un modelo el cual no presente problemas de multicolinealidad el modelo de componentes es el indicado, es importante recordar que la multicolinealidad puede hacer que los resultados sean menos fiables y más difíciles de interpretar.

Gráfica de conglomerados (clusters)





Podemos observar claramente **agrupaciones**, esto indica que las observaciones dentro de cada cluster son similares en sus características con los componentes PC1 Y PC2.

El **espacio o distancia** entre cluster nos indica que tanto las características que los definen son diferentes o similares, entre mayor espacio más diferente son estas.

También el **tamaño de los cluster** puede dar lugar a una hipótesis sobre las características o factores comunes que definen a los países dentro de ese cluster.

A Partir de este gráfico podemos hacer un análisis más exhaustivo sobre los países y sus características con las variables, por ejemplo, si un cluster contiene paises que todos están más desarrollados, podríamos concluir que comparten características económicas y de salud similares.