Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Guadalajara

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 501)

Actividad 5.2 Componentes Principales

Samuel Padilla Esqueda | A01641383

Octubre 2024

Realizar una regresión lineal múltiple utilizando todas las variables (sin interacciones ni términos de orden superior). Utilizar gdpp como la variable de respuesta.

- **Incluir la interpretación de los p-value, VIF, supuestos, residuales, etc...**

Regression Equation

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

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-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

child_mort, **exports**, y **life_expec** tienen valores de VIF relativamente altos, lo que sugiere multicolinealidad. Esto significa que hay redundancia en la información aportada por estas variables.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	8	48297074740	6037134343	127,71	0,000
child_mort	1	166006903	166006903	3,51	0,063
exports	1	20545258	20545258	0,43	0,511
health	1	2197072010	2197072010	46,48	0,000
imports	1	20677954	20677954	0,44	0,509
income	1	15300085769	15300085769	323,65	0,000
inflation	1	148654080	148654080	3,14	0,078
life_expec	1	349885273	349885273	7,40	0,007
total_fer	1	38688549	38688549	0,82	0,367
Error	158	7469200973	47273424		
Total	166	55766275714			

Las variables **health** (p = 0,000), **income** (p = 0,000), y **life_expec** (p = 0,007) son significativas y tienen un impacto importante en la variable dependiente **gdpp**.

child_mort (p = 0,063) e **inflation** (p = 0,078) son cercanas al nivel de significancia de 0,05, por lo que podrían ser relevantes, aunque con menos importancia.

Las variables **exports**, **imports**, y **total_fer** no son significativas y tienen un valor (p > 0,05) muy alejado a lo necesario para ser significativas.

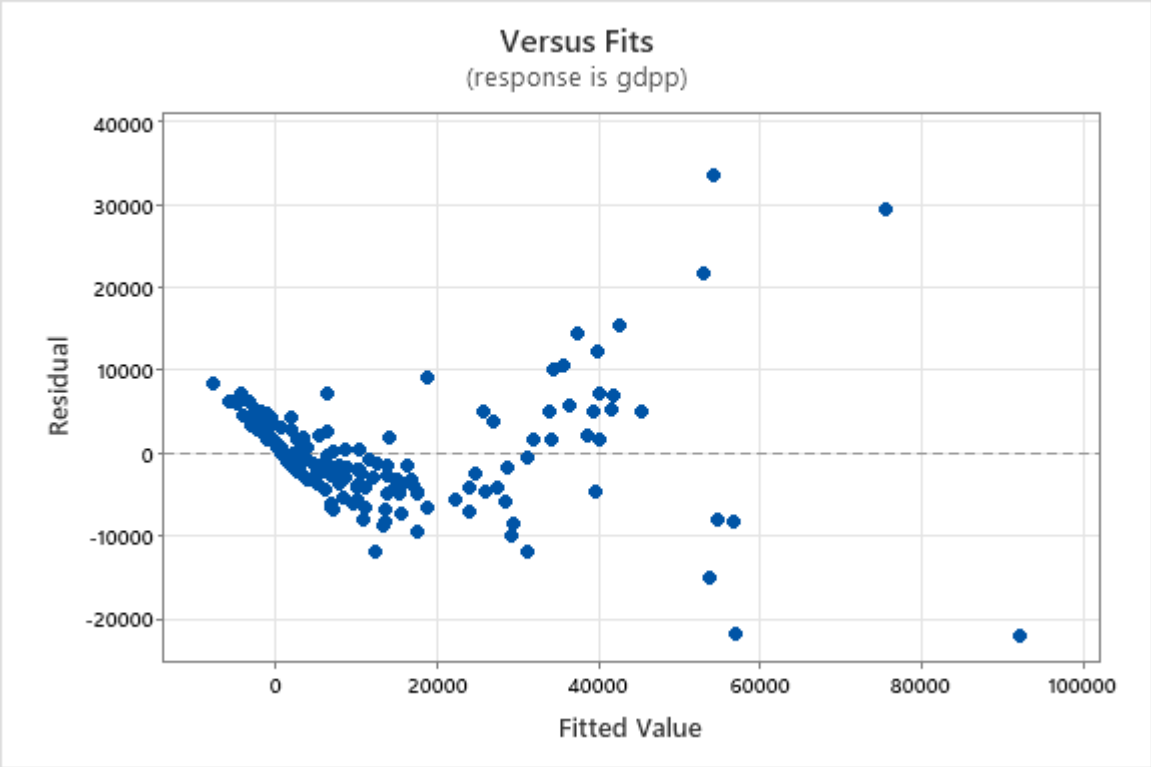
La prueba F (**F-Value** = 127,71, p < 0,000) indica que al menos una de las variables predictoras es significativamente diferente de cero

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6875,57	86,61%	85,93%	82,70%

R-sq: 86,61%. Esto indica que el modelo explica una gran parte de la varianza en la variable dependiente, lo cual es muy bueno.

R-sq(adj): 85,93%. Este valor ajustado toma en cuenta el número de variables en el modelo y es también alto.



Algunas variables, como income, parecen tener una relación más fuerte y lineal con la variable dependiente, pero los residuos parecen distribuirse de manera aleatoria alrededor de la línea horizontal en y=0, esto sugiere que el modelo cumple con los supuestos de:

- Linealidad

- Homoscedasticidad
- Normalidad de los residuos

Aplicar la técnica de componentes principales a los datos para reducir la dimensionalidad de las variables predictoras.

- **Incluir la explicación del procedimiento y la interpretación de los resultados, valores y vectores propios, direcciones de los componentes y si existen o no agrupaciones de los datos.**
- **Elegir los componentes principales que expliquen al menos el 80% de la varianza total.**

Para realizar el análisis de componentes principales hay que seguir los siguientes pasos:

1.Estandarizar los datos.
2. Calcular matriz de covarianza o correlación.
3. Calcular valores y vectores propios.
4. Elegir los componentes principales.

Usando Minitab podemos llevar a cabo el análisis mediante la herramienta en Stat -> Multivariate -> Principal Components.

Aquí debemos elegir entre usar la matriz de covarianza o correlación, si usamos la matriz de correlación, Minitab va a estandariza las variables transformándolas para que tengan media cero y varianza unitaria. También decidimos el numero de componentes que se van a generar, en mi caso fueron 8.

Los resultados son los siguientes:

Eigenanalysis of the Correlation Matrix

Eigenvalue	3,5746	1,5439	1,1634	0,7388	0,5622	0,2235	0,1085	0,0850
Proportion	0,447	0,193	0,145	0,092	0,070	0,028	0,014	0,011
Cumulative	0,447	0,640	0,785	0,878	0,948	0,976	0,989	1,000

- PC1 explica 44.7% de la varianza total.
- PC2 explica 19.3%, lo que hace que la varianza acumulada sea 64.0%.
- PC3 agrega otro 14.5%, lo que hace que la varianza acumulada sea 78.5%.

Para tomar por lo menos el 80% exacto, vamos a tomar el PC4 también.

- Con PC4 alcanzamos 87.8% de la varianza explicada.

Eigenvectors

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

PC1: Está fuertemente influenciado por child_mort, life_expec y total_fer, sugiriendo que este componente podría estar relacionado con indicadores de salud y fertilidad.

PC2: exports y imports tienen las cargas más significativas, lo que indica que este componente está relacionado con el comercio.

PC3: La variable health tiene la carga más alta, lo que indica que este componente se relaciona principalmente con el gasto en salud.

PC4: Las cargas en inflation y health indican que este componente podría estar relacionado con el crecimiento económico y el gasto.

Obtener las ecuaciones de Transformación Lineal de cada componente en función de las variables más importantes.

Usando la tabla de Eigenvectores, los valores en cada fila representan los coeficientes para cada variable en cada componente.

PC1=−0.473*child_mort+0.308*exports+0.145*health+0.195*imports+0.387*income−0.220*inflation + 0.464*life_expec − 0.457*total_fer

PC2=−0.214*child_mort−0.608*exports+0.242*health−0.661*imports−0.031*income−0.006*inflation+0.237*life_expec−0.177*total_fer

PC3=0.100*child_mort−0.146*exports+0.647*health+0.285*imports−0.248*income−0.616*inflation−0.158*life_expec+0.051*total_fer

PC4=−0.115*child_mort−0.102*exports−0.680*health−0.056*imports−0.315*income−0.621*inflation−0.004*life_expec−0.159*total_fer

Ahora reducimos las ecuaciones de cada componente en función a las variables más importantes:

PC1:

- Cargas significativas: child_mort, exports, life_expec.
- PC1=−0.473*child_mort+0.308*exports+0.464*life_expec

PC2:

- Cargas significativas: exports, imports, health.
- PC2=−0.608*exports−0.661*imports+0.242*health

PC3:

- Cargas significativas: health, imports, inflation.
- PC3=0.647*health+0.285*imports−0.616*inflation

PC4:

- Cargas significativas: health, income, inflation.
- PC4=−0.680*health−0.315*income−0.621*inflation

Dar un nombre a cada componente principal con base en las variables que lo conforman.

PC1 = Calidad de Vida

PC2 = Dependencia Comercial y Salud

PC3 = Estabilidad Económica

PC4 = Presiones Económicas

Realizar nuevamente la regresión con los componentes principales seleccionados.

- **Realizar la interpretación adecuada.**

Regression Equation

gdpp = 12964 + 6726 PC1 + 618 PC2 - 883 PC3
- 7516 PC4

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12964	897	14,46	0,000	
PC1	6726	476	14,14	0,000	1,00
PC2	618	724	0,85	0,394	1,00
PC3	-883	834	-1,06	0,291	1,00
PC4	-7516	1046	-7,18	0,000	1,00

Solo **PC1** (p = 0,000) y **PC4** (p = 0,000) son estadísticamente significativas.

PC2 (p = 0,394) y **PC3** (p = 0,291) no son significativas, lo que sugiere que no aportan información útil para predecir la variable dependiente.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
11586,1	61,00%	60,04%	54,84%

R-sq: 61,00%. Esto indica que el modelo ahora explica una menor cantidad de la varianza en la variable dependiente en comparación con el modelo anterior.

R-sq(adj): 60,04%, también refleja una reducción en la capacidad del modelo.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	34019924428	8504981107	63,36	0,000
PC1	1	26843079811	26843079811	199,97	0,000
PC2	1	97885223	97885223	0,73	0,394
PC3	1	150587861	150587861	1,12	0,291
PC4	1	6928371533	6928371533	51,61	0,000
Error	162	21746351286	134236736		
Total	166	55766275714			

La prueba F (F-Value = 63,36, p < 0,000) sugiere que el modelo es significativo en general, aunque no se explica tanta varianza como en el modelo anterior.

Comparar y comentar las diferencias entre ambos modelos de regresión (antes de aplicar la técnica de componentes principales y después de aplicarla).

Antes del PCA:

- El modelo tiene un buen ajuste y las variables independientes muestran significancia, indicando que proporcionan información útil sobre la variable dependiente.

Después del PCA:

- La capacidad explicativa del modelo se reduce significativamente. Esto puede ser indicativo de que, aunque los componentes principales pueden capturar la varianza, no siempre son tan informativos como las variables originales. Esto es especialmente relevante si los componentes que se están utilizando no representan adecuadamente la información contenida en las variables originales.

Realizar un análisis de conglomerados (clusters) utilizando los componentes principales y presentar una visualización de los países en cada uno de los grupos.

Method

Number of clusters	3
Standardized variables	No

Final Partition

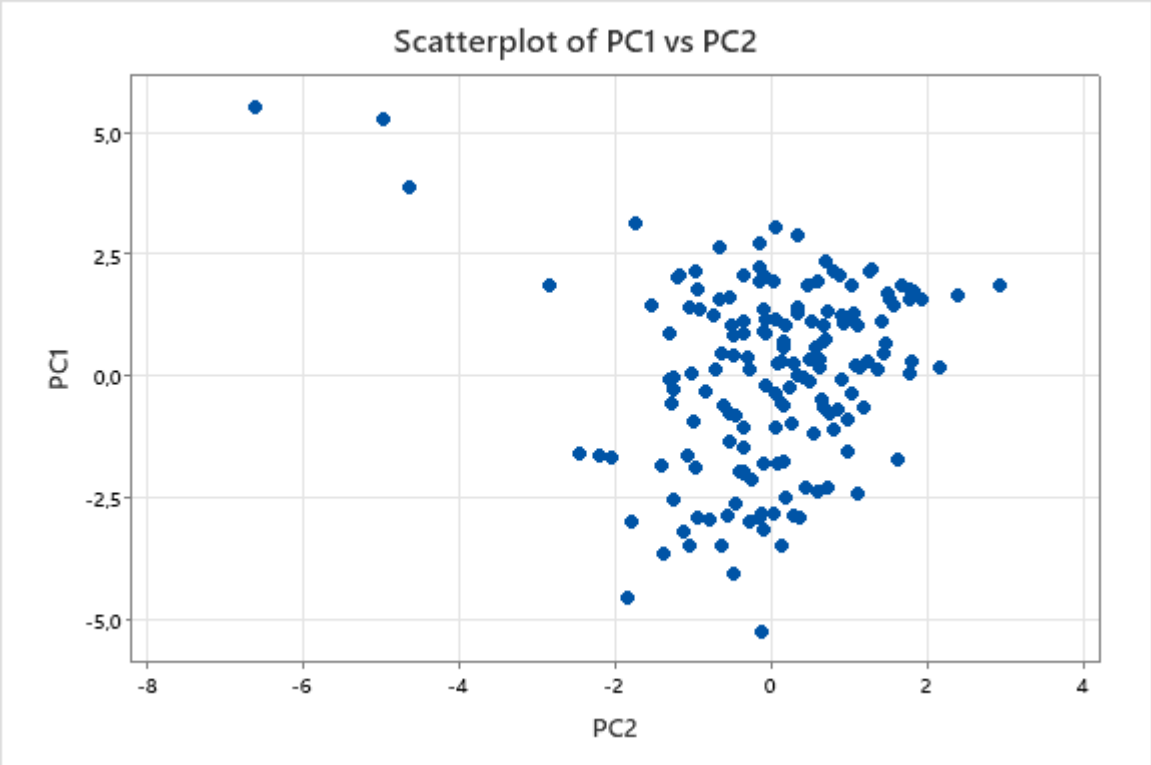
	Number of observations	Within cluster sum of squares	Average distance from centroid	Maximum distance from centroid
Cluster1	46	206,992	1,754	8,396
Cluster2	61	290,009	1,770	7,704
Cluster3	60	136,602	1,386	3,370

Cluster Centroids

Variable	Cluster1	Cluster2	Cluster3	Grand centroid
PC1	-2,4152	1,8009	0,0208	-0,0000
PC2	-0,4980	0,0652	0,3156	-0,0000
PC3	0,3045	0,3393	-0,5784	-0,0000
PC4	-0,2354	-0,2373	0,4217	0,0000

Distances Between Cluster Centroids

	Cluster1	Cluster2	Cluster3
Cluster1	0,0000	4,2536	2,7942
Cluster2	4,2536	0,0000	2,1232
Cluster3	2,7942	2,1232	0,0000



Por mas que intente cambiar el color a los closters no fue posible, esta es la grafica con PC1 y PC2, si se pudieran ver los colores deberían verse separados los grupos de datos que pertenecen a cada closter.