ACTIVIDAD 2. Componentes Principales

Instituto Tecnológico y de Estudios Superiores de Mty

Grace Aviance Silva Aróstegui A01285158

Concentración TC3006C:

Carlos Alberto Sánchez Villanueva A01640495

Inteligencia Artificial Avanada para Ciencia de Datos

Fecha, 27 de octubre del 2024.

Módulo: Estadística Prof: Ramiro Zermeño Díaz

Campus Guadalajara, Zapopan.

Dataset: Country

Column	Description
country	Name of the country
child_mort	Death of children under 5 years of age per 1000 live births
exports	Exports of goods and services. Given as %age of the Total GDP
health	Total health spending as %age of Total GDP
imports	Imports of goods and services. Given as %age of the Total GDP
Income	Net income per person
Inflation	The measurement of the annual growth rate of the Total GDP
life_expec	Average of years a new born child would live if current mortality patterns are to remain the same
total_fer	Children that would be born to each woman if the current age-fertility rates remain the same.
gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population.

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

Regression Equation

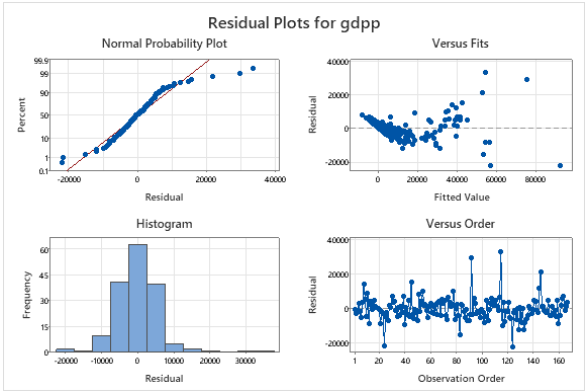
gdpp = -41934 + 66.6 child_mort + 28.5 exports + 1549 health - 28.1 imports + 0.7856 income - 100.5 inflation + 389 life_expec + 615 total_fer

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
6875.57	86.61%	85.93%	82.70%

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



Interpretación de los p-value, VIF, supuestos, residuales, etc...

Los p-values de cada variable independiente indican la significancia estadística de cada predictor en relación con la variable dependiente. Un valor p menor a 0.05 sugiere que la variable independiente tiene un efecto estadísticamente significativo sobre la variable dependiente. Observamos que las variables significativas son: health, income, life_expec.

El VIF mide la multicolinealidad entre las variables independientes. Valores de VIF mayores a 5 indican que la variable está altamente correlacionada con otras variables independientes. Si alguna variable tiene un VIF alto, significa que tiene una relación fuerte con otra variable independiente, lo cual puede afectar la estabilidad de las estimaciones de los coeficientes. Se suelen eliminar esas variables de VIF alto, que en este caso son: child_mort y lfe_expc.

Los supuestos de una regresión linal son:

- Linealidad: La relación entre las variables independientes y la variable dependiente debe ser lineal.
  - En el gráfico *Versus Fits* (Residuals vs Fitted Values) los residuales muestran una dispersión en forma de curva, lo que indica una posible falta de linealidad en el modelo.
- Independencia de los errores: Los residuales deben ser independientes unos de otros.
  - Se cumple. En el gráfico *Versus Order* observamos que no hay tendencia en los residuos
- Normalidad de los residuales: Los errores deben estar distribuidos normalmente.
  - En *Normal Probability Plot*(Q-Q Plot) con un modelo ideal, los puntos deberían alinearse en la línea diagonal. En nuestro gráfico siguen la línea en la mayoría de los casos, pero hay valores en los extremos que se desvían significativamente. Los residuales no siguen perfectamente una distribución normal. lo que podría influir en la precisión de las predicciones.
  - En el *Histogram* de residuales si tiende a distribuirse con normalidad pero no perfectamente.
- Homoscedasticidad: Los residuales deben tener varianza constante.
  - En el gráfico *Versus Fits* se observa una expansión de los puntos conforme los valores ajustados aumentan, sugiriendo heteroscedasticidad (varianza no constante de los residuales).

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 vecotres propios, direcciones de los componentes y si existen o no agrupaciones de los datos.

Pasos:

1. Estandarizar los datos

   En Minitab en Calc >Standardize elegimos todas las columnas menos 'country'.

2. Calculamos la matriz de covarianza

   En Stat >Multivariate >Principal Components obtenemos este punto 2 y el 3.

3. Calculamos los valores propios y vectores propios

Eigenvectors

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
child_mort_1	-0.473	-0.214	0.100	-0.115	0.297	-0.203	0.135	0.748
exports_1	0.308	-0.608	-0.146	-0.102	0.058	0.053	0.696	-0.109
health_1	0.145	0.242	0.647	-0.680	-0.059	-0.014	0.183	-0.044
imports_1	0.195	-0.661	0.285	-0.056	-0.315	0.037	-0.569	0.125
income_1	0.387	-0.031	-0.248	-0.315	0.728	-0.179	-0.351	-0.054
inflation_1	-0.220	-0.006	-0.616	-0.621	-0.418	-0.064	-0.086	0.010
life_expec_1	0.464	0.237	-0.158	-0.004	-0.091	0.600	0.020	0.578
total_fer_1	-0.457	-0.177	0.051	-0.159	0.304	0.747	-0.090	-0.272

Figura 3: Vectores Propios de cada componente respecto cada variable

	C11	C12	C13	C14	C15	C16	C17	C18	C19
	child_mort_1	exports_1	health_1	imports_1	income_1	inflation_1	life_expec_1	total_fer_1	gdpp_1
1	1.28766	-1.13487	0.27825	-0.08221	-0.80582	0.15686	-1.61424	1.89718	-0.67714
2	-0.53733	-0.47822	-0.09673	0.07062	-0.37424	-0.31141	0.64592	-0.85739	-0.48417
3	-0.27201	-0.09882	-0.96318	-0.63984	-0.22018	0.78691	0.66841	-0.03829	-0.46398
4	2.00179	0.77306	-1.44373	-0.16482	-0.58329	1.38289	-1.17570	2.12177	-0.51472
5	-0.69355	0.16019	-0.28603	0.49608	0.10143	-0.59994	0.70215	-0.54032	-0.04169
6	-0.58940	-0.81019	0.46756	-1.27595	0.08068	1.24099	0.58970	-0.38178	-0.14535
7	-0.50014	-0.74088	-0.87944	-0.06569	-0.54179	-0.00112	0.30859	-0.83097	-0.53163
8	-0.82993	-0.77736	0.69691	-1.07355	1.25818	-0.62643	1.28686	-0.67243	2.12431
9	-0.84232	0.37177	1.52332	0.03758	1.35155	-0.65358	1.11820	-0.99611	1.85151
10	0.02306	0.48121	-0.34064	-1.08181	-0.05938	0.56933	-0.16369	-0.67904	-0.38869
11	-0.60676	-0.22286	0.39111	-0.13177	0.29854	-0.77335	0.36481	-0.71867	0.82034
12	-0.73570	1.03571	-0.67193	0.16563	1.24262	-0.03234	0.61219	-0.52050	0.42206
13	0.27598	-0.91598	-1.19981	-1.03638	-0.76277	-0.06072	-0.01751	-0.40821	-0.66596
14	-0.59684	-0.05870	0.42023	0.07475	-0.09569	-0.70580	0.69090	-0.77152	0.16563
15	-0.81257	0.37542	-0.43894	0.72739	-0.04900	0.69231	-0.01751	-0.96309	-0.37832
16	-0.83737	1.28743	1.41410	1.14871	1.24262	-0.55832	1.06197	-0.71867	1.71512
17	-0.48278	0.62349	-0.58820	0.43825	-0.48058	-0.62832	0.09494	-0.15719	-0.47053
18	1.80342	-0.63144	-0.98866	-0.40026	-0.79493	-0.65245	-0.98454	1.59331	-0.66596
19	0.10985	0.05075	-0.58820	0.98349	-0.55632	-0.16951	0.17365	-0.37518	-0.58838
20	0.20655	0.00332	-0.71926	-0.52005	-0.60871	0.09443	0.11743	0.16649	-0.59929

Figura 1: Se muestran las primeras 20 filas de 167

Eigenanalysis of the Covariance 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

Figura 2: Valores propios de cada variable

4. Elegimos los componentes principales

Elegir los componentes principales que expliquen al menos el 80 % de la varianza total.

Eigenanalysis of the Covariance 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

Figura 4: Selección de Componentes Principales

Los primeros 4 componentes tienen un 87.8 % de explicabilidad.

Obtener las ecuaciones de Transformación Lineal de cada componente en función de las variables más importantes. Para esto, repetimos el PCA pero ahora con 4 componentes

Eigenvectors

Variable	PC1	PC2	PC3	PC4
child_mort_1	-0.473	-0.214	0.100	-0.115
exports_1	0.308	-0.608	-0.146	-0.102
health_1	0.145	0.242	0.647	-0.680
imports_1	0.195	-0.661	0.285	-0.056
income_1	0.387	-0.031	-0.248	-0.315
inflation_1	-0.220	-0.006	-0.616	-0.621
life_expec_1	0.464	0.237	-0.158	-0.004
total_fer_1	-0.457	-0.177	0.051	-0.159

Figura 5: Coeficientes elegidos por componente principal

Ecuaciones de Transformación Lineal:

- PC1 = -0.473 child_mort_1 + 0.464 life_expec_1 - 0.457 total_fer_1
- PC2 = -0.608 exports_1 - 0.661 imports_1
- PC3 = 0.647 health_1 - 0.616 inflation_1
- PC4 = -0.680 health_1 - 0.621 inflation_1

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

- PC1 = Salud y calidad de vida (basado en mortalidad infantil, esperanza de vida y fertilidad).
- PC2 = Actividad comercial (nivel de exportaciones e importaciones).
- PC3 = Inversión en Salud y Estabilidad Económica (mayor gasto en salud con menor inflación).
- PC4 = Perfil Económico de Salud e Inflación.

Realizar nuevamente la regresión con los componentes principales seleccionados. Interpreta

Regression Equation

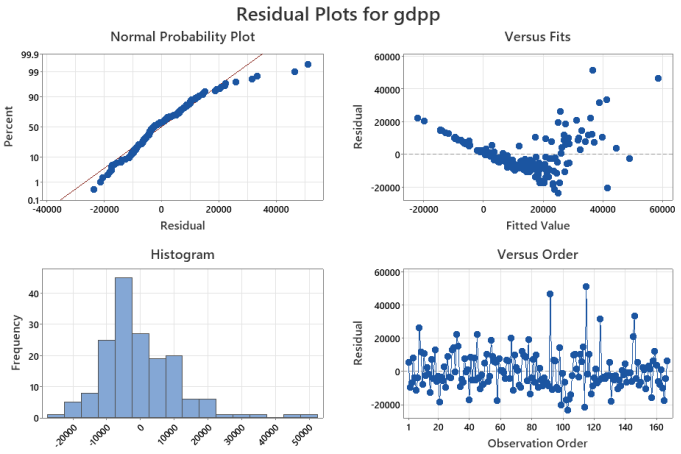
gdpp = 12964 + 6726 PC1 + 618 PC2 - 883 PC3 - 7516 PC4

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

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
11586.1	61.00%	60.04%	54.84%



Interpretación de los p-value, VIF, supuestos, residuales, etc...

Los p-values de cada variable independiente indican la significancia estadística de cada predictor en relación con la variable dependiente.Observamos que las variables significativas son: PC1 (Salud y calidad de vida), PC4 (*Perfil Económico de Salud e Inflación.*)

El VIF mide la multicolinealidad entre las variables independientes. Valores de VIF mayores a 5 indican que la variable está altamente correlacionada con otras variables independientes. No hay alguna variable con un VIF alto, lo que significa que no hay una relaciones fuertes con otras variables.

Los supuestos de una regresión linal son:

- Linealidad: La relación entre las variables independientes y la variable dependiente debe ser lineal.
  - En el gráfico *Versus Fits* (Residuals vs Fitted Values) los residuales muestran una dispersión en forma de curva, lo que indica una posible falta de linealidad en el modelo. Realmenet similar al anterior.
- Independencia de los errores: Los residuales deben ser independientes unos de otros.
  - En el gráfico *Versus Order* observamos que no hay tendencia en los residuos

- Normalidad de los residuales: Los errores deben estar distribuidos normalmente.
  - En *Normal Probability Plot*(Q-Q Plot) con un modelo ideal, los puntos deberían alinearse en la línea diagonal. Los residuales no siguen una distribución normal.
  - En el *Histogram* de residuales no se tienden a distribuirse con normalidad.
- Homoscedasticidad: Los residuales deben tener varianza constante.
  - En el gráfico *Versus Fits* se observa una expansión de los puntos conforme los valores ajustados aumentan, sugiriendo heteroscedasticidad (varianza no constante de los residuales).

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

La versión simplificada del modelo usando PCA reduce la cantidad de predictores, lo cual facilita su aplicación y disminuye la colinealidad. En cambio, el modelo completo emplea todas las variables, lo que aumenta su complejidad.

El modelo completo permite interpretar directamente el impacto de cada variable sobre la variable de interés, facilitando la comprensión de sus efectos. Por su parte, el modelo con PCA, aunque más eficiente, requiere un análisis adicional para interpretar los componentes principales.

Si bien el modelo completo enfrenta problemas de multicolinealidad, el modelo con PCA resuelve este inconveniente al generar predictores independientes entre sí.

En el modelo completo, múltiples variables pueden resultar significativas, mientras que en el modelo con PCA, solo el primer componente (PC1) demostró ser un predictor fuerte para la variable gdpp.

En conclusión, el modelo completo es preferible cuando se busca una interpretación clara de los efectos de cada variable, aunque esté sujeto a multicolinealidad. Por otro lado, el modelo con PCA es más sencillo y estable, libre de problemas de colinealidad, pero con una interpretación menos intuitiva.

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