Universidad Nacional del Oeste

Licenciatura en Informática

Explotación de Datos

Integrantes Grupo 1:

- Robledo Alan
- Farías Gonzalo
- Ramirez Gonzalo
- Romano Diego

Docentes:

- Perez Silvia
- Mendoza Dante

10 de noviembre de 2023

Índice

1. Introducción

El presente trabajo práctico explora las posibilidades que brindan los modelos de regresión lineal múltiple para el proceso de explotación de datos. El objetivos es generar modelos que nos permitan efectuar predicciones sobre la variable respuesta.
El dataset utilizado para esta ocasión mide una serie de parámetros sociales. La variable que se pretente predecir es la Satisfacción de Vida.
El origen del dataset es World Happiness Report - Data cuyo website es https://worldhappiness.report/data/

El Informe Mundial sobre la Felicidad es una publicación de la Red de Soluciones para el Desarrollo Sostenible, impulsada por los datos de la Encuesta Mundial Gallup. El Informe Mundial sobre la Felicidad refleja una demanda mundial de mayor atención a la felicidad y el bienestar como criterios para las políticas gubernamentales. Revisa el estado de la felicidad en el mundo actual y muestra cómo la ciencia de la felicidad explica las variaciones personales y nacionales en la felicidad.

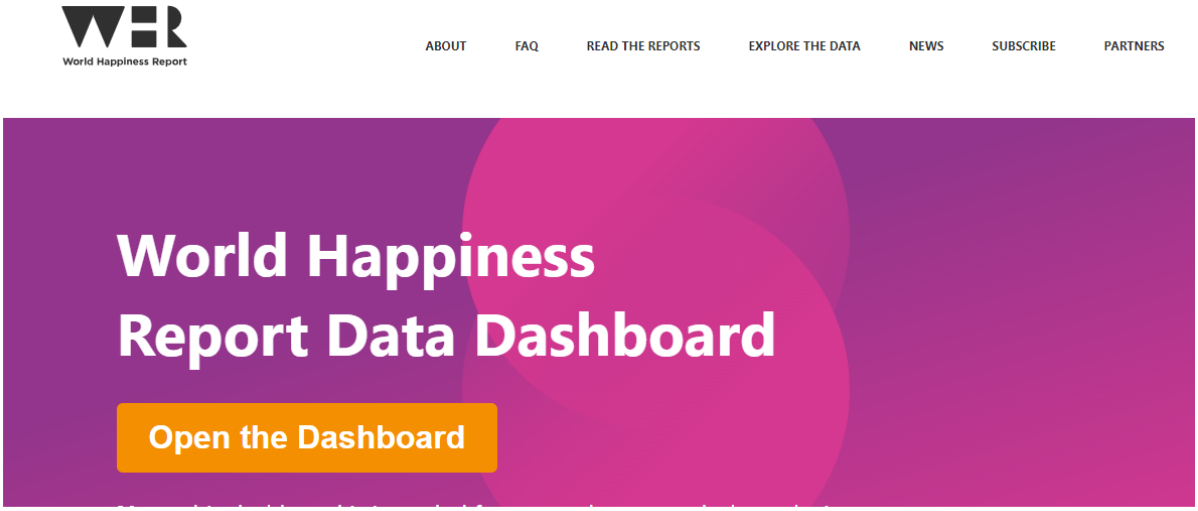

Figura 1

2. Analisis y preparación del dataset

El informe clasifica a los países del mundo en función de varios datos/factores relacionados con la felicidad y el bienestar de sus ciudadanos.

2.1. Descripción de los datos:

- Inidice de felicidad(nuestra variable a predecir): se basa en la encuesta del Gallup World Poll (GWP) lanzada el 26 de febrero de 2021, que abarca los años 2005-2020. A menos que se especifique lo contrario, refleja la respuesta promedio nacional a la pregunta sobre las evaluaciones de la vida. Formulada en inglés como 'Imagina una escalera del 0 al 10, donde el 10 representa la mejor vida posible y el 0 la peor. ¿En qué escalón te encuentras ahora?' También conocida como la escalera de vida de Cantril en nuestro análisis.
- PIB per cápita: es una medida económica que se utiliza para evaluar y comparar el nivel de ingresos promedio de la población en un país.
- Expectativa de Vida Saludable (HLE): Las expectativas de vida saludable al nacer se derivan de datos extraídos del Observatorio Global de la Salud de la Organización Mundial de la Salud (OMS) hasta septiembre de 2020.
- Apoyo social: Representa el promedio nacional de respuestas binarias (0 o 1) a la pregunta del GWP sobre la disponibilidad de parientes o amigos para ayudar en tiempos difíciles.
- Libertad para tomar decisiones de vida: Es el promedio nacional de respuestas a la pregunta del GWP sobre la satisfacción con la libertad para elegir el rumbo de la vida.
- Generosidad: Calculada como el residuo de regresionar las donaciones benéficas (respuesta a la pregunta del GWP) sobre el PIB per cápita a nivel nacional.
- Percepción de la Corrupción: Se basa en respuestas a dos preguntas del GWP sobre la prevalencia de la corrupción en el gobierno y los negocios. La percepción general se obtiene promediando las respuestas binarias (0 o 1).
- Afecto Positivo: Representa el promedio de tres medidas (felicidad, risa y disfrute) de las olas 3-7 de la Encuesta Mundial Gallup.
- Afecto Negativo: Es el promedio de las respuestas a tres medidas (preocupación, tristeza e ira) de las olas 3-7 de la Encuesta Mundial Gallup.
- Confianza Institucional: Se deriva del primer componente principal de medidas como la confianza en el gobierno, el sistema judicial, la honestidad electoral, la fuerza policial y la percepción de la corrupción. Se convierte en una medida binaria de alta confianza institucional utilizando el percentil 75 en la distribución global como punto de corte. No está disponible para todos los países debido a variaciones en las encuestas.

2.2. Limpieza de los datos:

- El conjunto de datos incluye, además de las variables mencionadas anteriormente, dos variables adicionales: país y año, que indican el país donde se realizó el informe y el año correspondiente. En nuestro caso, nos enfocamos en el año 2014 después de realizar pruebas con todos los años y determinar que era el año que proporcionaba el mejor modelo de regresión lineal múltiple (rlm).

  Adicionalmente, eliminamos estas dos variables del conjunto de datos, ya que no eran necesarias para la construcción del modelo de regresión lineal múltiple. También llevamos a cabo la traducción de los nombres de las variables de la siguiente manera:

```
colnames(datos)<-c("SatisfaccionVida", "LogPBI",
    "ApoyoSocial", "EsperanzaVidaSaludableNacer",
    "LibertadTomarDecisionesVida", "Generosidad", "PercepcionCorrupcion",
    "AfectoPositivo", "AfectoNegativo",
    "ConfianzaGobiernoNacional")
```

- También realizamos una limpieza de los valores nulos, ademas de hacer el casteo de los datos, esto se hizo de la siguiente manera:

```
# BÚSQUEDA DE DATOS NULOS
# Visualizamos un resumen de datos nulos en el conjunto de datos.
# View(summarise_all(datos, funs(sum(is.na(.)))))
datos <- na.omit(datos) # Eliminamos las filas con valores nulos en caso de haber

# CASTEO DE DATOS
# Convertimos todas las columnas a tipo numérico para asegurarnos de que sean interpretables.
attach(datos)
datos <- datos %>% mutate_all(as.numeric)
```

2.3. Estadísticas Descriptivas

El análisis de las estadísticas descriptivas proporciona una visión detallada de las variables presentes en el conjunto de datos. A continuación, se presentan los valores clave para cada variable:

Tabla 1: Estadísticas Descriptivas del Dataset

Variable	**Mínimo**	**1er Cuartil**	**Mediana**	**Media**	**3er Cuartil**	**Máximo**
SatisfaccionVida	2.839	4.483	5.313	5.408	6.449	7.508
LogPBI	6.640	8.475	9.476	9.368	10.217	11.638
ApoyoSocial	0.4443	0.7546	0.8323	0.8100	0.9000	0.9677
EsperanzaVidaSaludableNacer	47.66	57.90	65.10	63.12	68.18	73.48
LibertadTomarDecisionesVida	0.3692	0.6492	0.7524	0.7400	0.8527	0.9563
Generosidad	-0.28911	-0.09590	-0.00238	0.02314	0.11077	0.70638
PercepcionCorrupcion	0.1326	0.6696	0.8008	0.7449	0.8693	0.9763
AfectoPositivo	0.4095	0.5729	0.6613	0.6596	0.7598	0.8760
AfectoNegativo	0.1120	0.2207	0.2668	0.2747	0.3241	0.5636
ConfianzaGobiernoNacional	0.0951	0.3373	0.4686	0.4748	0.6093	0.9585

Estas estadísticas proporcionan una comprensión inicial de la distribución y variabilidad de las variables en el dataset, preparando el terreno para un análisis más detallado.

3. Análisis Exploratorio de los Datos

La exploración del gráfico de correlación entre las variables clave revela patrones significativos que pueden arrojar luz sobre las relaciones subyacentes en nuestro conjunto de datos. En particular, se observa una correlación destacada entre las variables LogPBI, SatisfacciónVida, ApoyoSocial, EsperanzaVidaSaludableNacer y AfectoPositivo.
La fuerte correlación positiva entre LogPBI (Producto Interno Bruto en escala logarítmica) y SatisfacciónVida sugiere una tendencia positiva en la relación entre la riqueza económica y la satisfacción general de vida. Además, la correlación positiva entre ApoyoSocial y SatisfacciónVida indica que un mayor apoyo social está asociado con niveles más altos de satisfacción.
La relación positiva entre EsperanzaVidaSaludableNacer y SatisfacciónVida resalta la importancia de la salud percibida en la percepción general de bienestar.

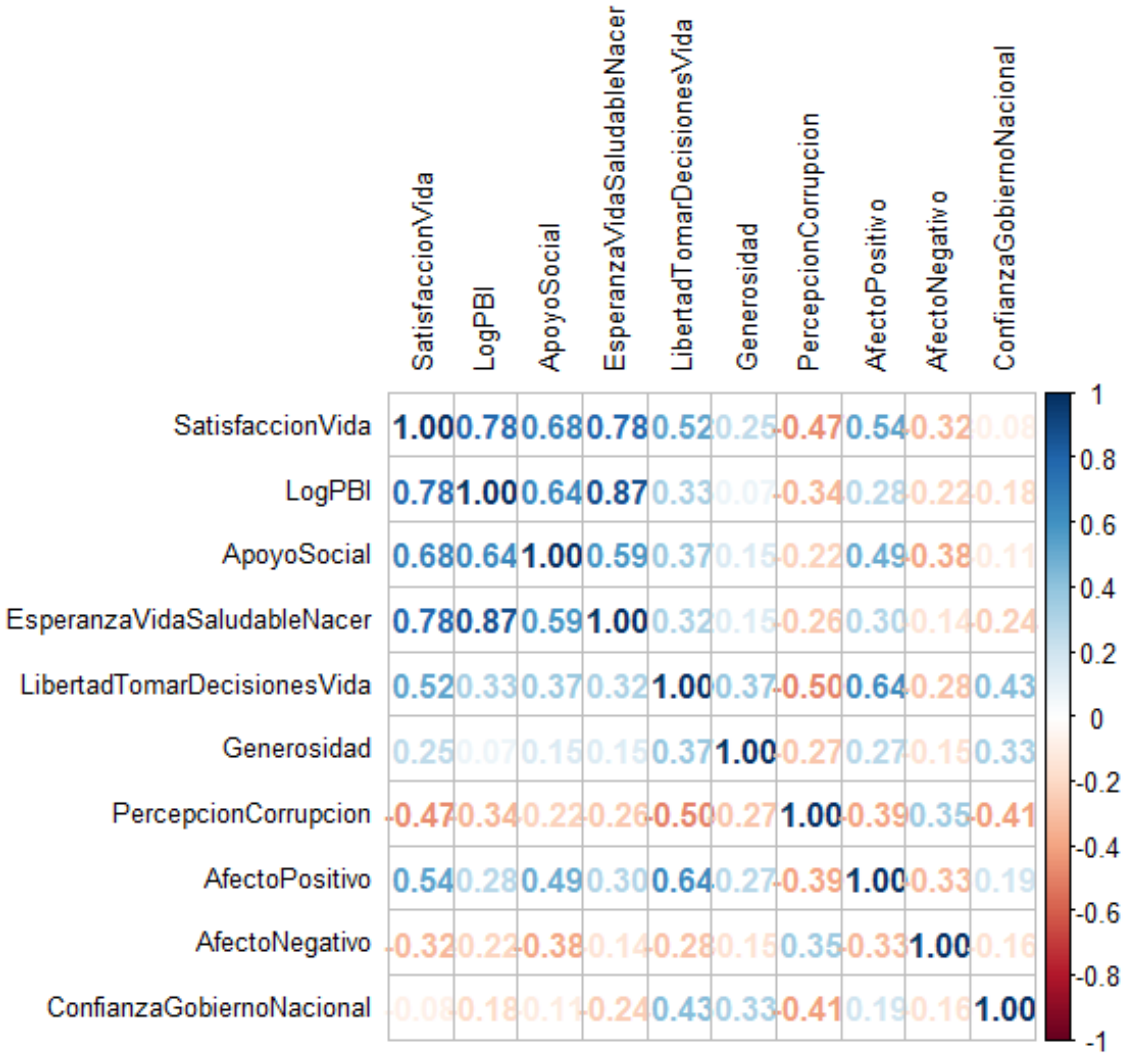

Figura 2: Matriz de correlación

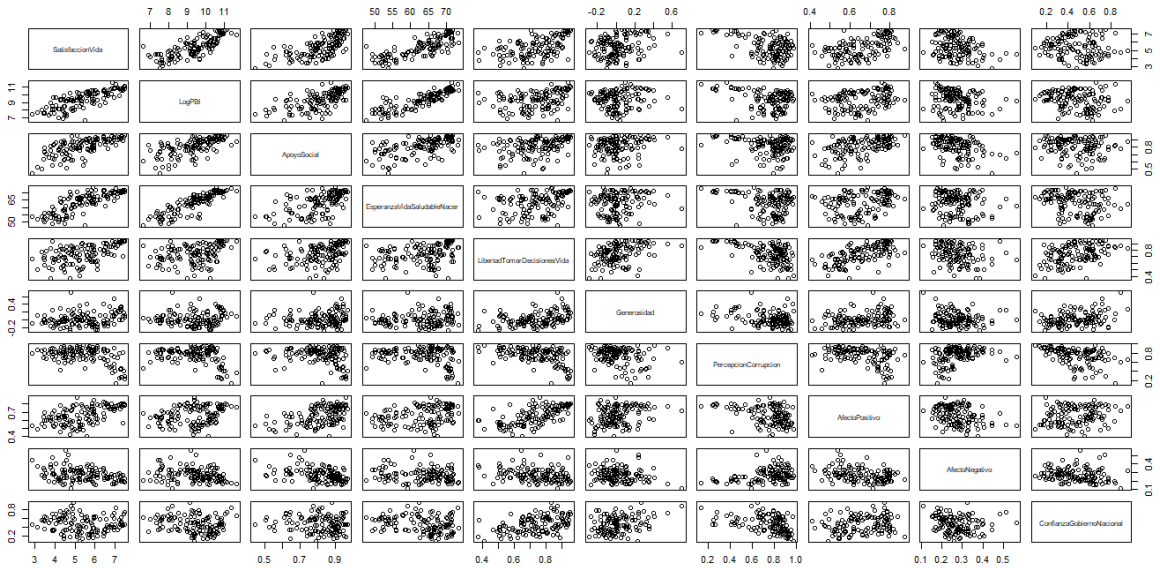

Figura 3: Diagrama de dispersión entre cada par de variables

4. Generación de Modelos

En el proceso de generación del modelo de regresión lineal múltiple, exploramos distintos métodos automáticos de selección de variables: 'forward', 'backward' y 'both'. A continuación, se presentan los resultados de cada método:

Método Forward: El método de selección hacia adelante identificó un subconjunto óptimo de variables predictoras para el modelo de regresión. El modelo resultante dió los siguientes resultados:
El R-cuadrado ajustado del modelo es 0.7783, y el F-statistic es significativamente alto (p-value $<$ 2.2e-16), indicando una buena capacidad predictiva.

Método Backward: El método de selección hacia atrás eliminó variables de manera iterativa para optimizar el modelo con los siguientes resultados:
El R-cuadrado ajustado del modelo es 0.7779, y el F-statistic es significativo (p-value $<$ 2.2e-16), indicando una buena capacidad predictiva.

Método Both (Bidireccional): La selección bidireccional combina aspectos de los métodos hacia adelante y hacia atrás. El modelo resultante es consistente con los modelos anteriores e incluye las mismas variables significativas.
El R-cuadrado ajustado y el F-statistic del modelo son similares a los obtenidos con los otros métodos, con un R-cuadrado ajustado de 0.7779 y un F-statistic significativo (p-value $<$ 2.2e-16).

5. Elección de un Modelo

La elección del modelo generado mediante el método backward se sustenta en la búsqueda de la parsimonia al seleccionar un conjunto más reducido de variables predictoras sin comprometer significativamente la capacidad predictiva del modelo y tener un R cuadrado ajustado ligeramente superior.

R cuadrado = 0.7904
R cuadrado ajustado = 0.7779
Estadístico F = 63.03.
P valor < 2.2e-16.

El modelo resultante, incluye las siguientes variables significativas: LogPBI, ApoyoSocial, EsperanzaVidaSaludableNacer, LibertadTomarDecisionesVida, PercepcionCorrupcion, AfectoPositivo, ConfianzaGobiernoNacional

El gráfico de dispersión sobre nuestro modelo revela una dispersión de puntos alrededor de la línea de regresión, indicando una representación adecuada de la variabilidad en la relación entre nuestras variables independientes y la variable dependiente. Esta dispersión sugiere que la línea de regresión ha capturado de manera efectiva la tendencia central de los datos, permitiendo que la variación natural en la respuesta se manifieste de manera coherente.
La adecuada dispersión alrededor de la línea de regresión es un indicador positivo de la capacidad del modelo para explicar la variabilidad observada en la variable dependiente a partir de las variables independientes consideradas. La presencia de puntos que se distribuyen uniformemente alrededor de la línea sugiere que la relación modelada es consistente a lo largo del rango de las variables predictoras, lo cual es fundamental para la validez y la fiabilidad de nuestras inferencias.

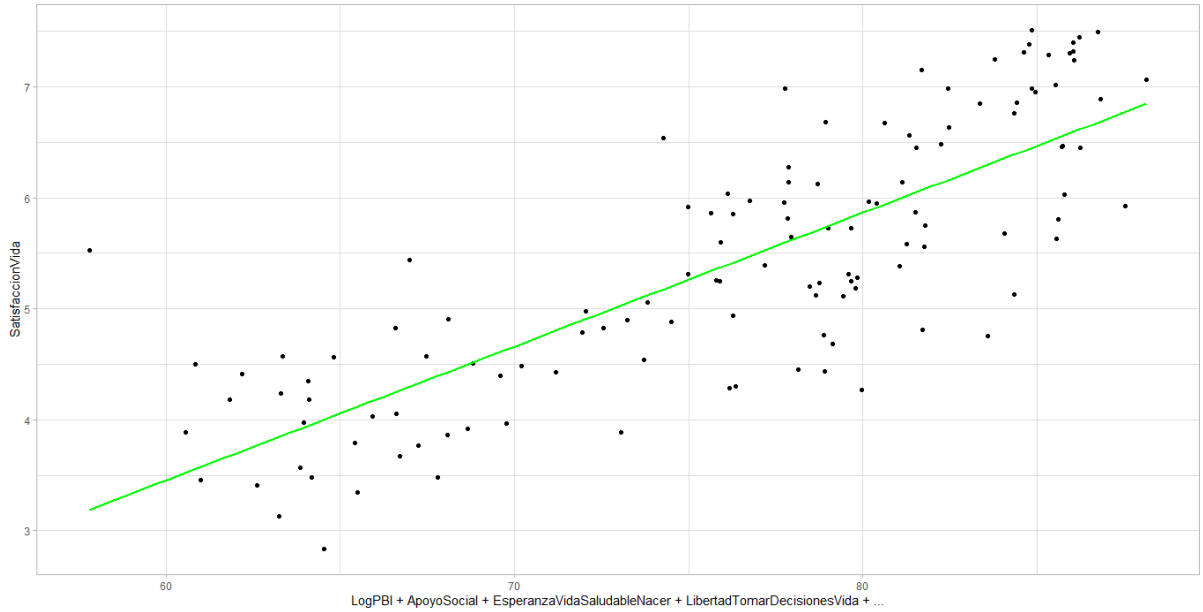

Figura 4: Diagrama de dispersión entre variable dependiente e independientes

6. Verificación de Supuestos de RLM

6.1. Gráficos

Al examinar la distribución de los residuos en nuestro modelo de regresión lineal múltiple, observamos que el histograma de los residuos presenta una notable similitud con la forma de una campana de Gauss. Este patrón es indicativo de una distribución normal de los residuos alrededor de cero, lo cual es una suposición fundamental para los modelos lineales.

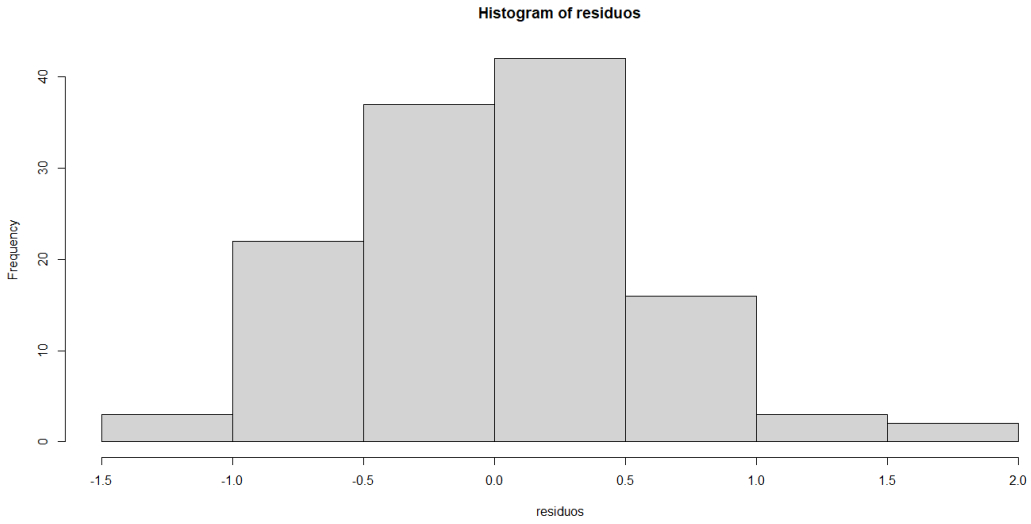

Figura 5: Histograma de Residuos

La inspección del gráfico de residuos revela una dispersión sin patrón aparente alrededor de la línea cero. Este fenómeno indica que la variación no explicada por el modelo se distribuye de manera aleatoria y no sigue ninguna tendencia discernible. La falta de patrón en los residuos respalda la asunción de homocedasticidad, que implica que la varianza de los residuos es constante a lo largo de todo el rango de las variables independientes.

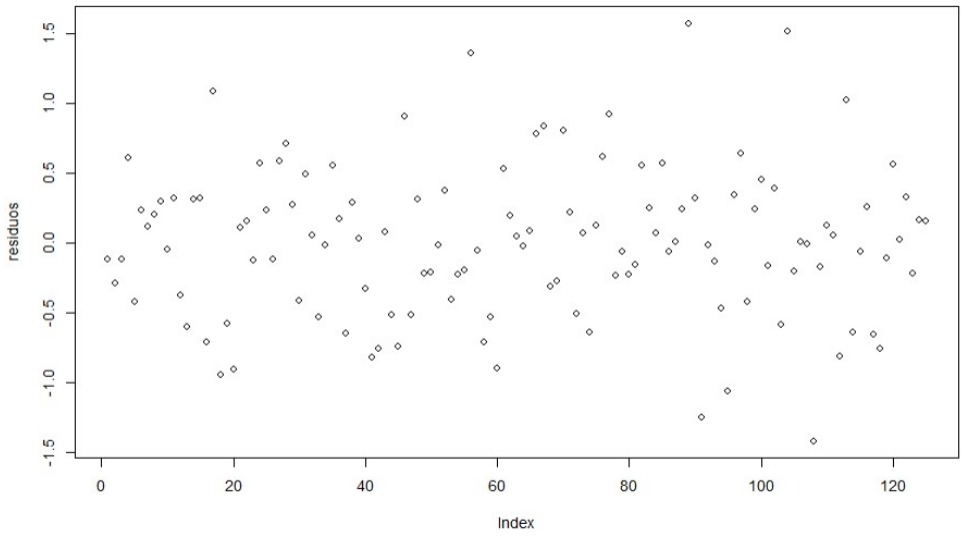

Figura 6: Dispersión de Residuos

El análisis de los residuos es crucial para evaluar la validez de un modelo de regresión lineal múltiple. En nuestro estudio, hemos explorado cuatro gráficos fundamentales:

- **Residuals vs Fitted:** En este gráfico, no se observa un patrón sistemático en la dispersión de los residuos en función de los valores ajustados. La ausencia de tendencias evidentes sugiere que la relación lineal entre las variables independientes y dependientes se mantiene.
- **QQ Plot de Residuos:** Revela una distribución de residuos que se asemeja a la normalidad. La mayoría de los puntos siguen la línea diagonal, indicando que los residuos se distribuyen de manera aproximada según una distribución normal, consistente con uno de los supuestos básicos de la regresión lineal.
- **Scale-Location:** Este análisis muestra una dispersión constante de los residuos en función de los valores ajustados, indicando principalmente homocedasticidad en el modelo. Aunque se observa una ligera disminución en la línea de tendencia a medida que los valores ajustados aumentan, la magnitud de la variación es moderada, aproximadamente 0.4 unidades. Esta disminución gradual podría indicar variabilidad ligeramente no constante de los errores, pero la magnitud sugiere que el impacto puede no ser significativo.
- **Residuals vs Leverage:** No se identifican puntos con un impacto desproporcionado en la regresión en este gráfico. No hay puntos que destaquen como valores atípicos influenciales, indicando que no hay observaciones con una influencia desmesurada en los resultados del modelo.

En conjunto, estos resultados respaldan la validez de los supuestos subyacentes en el análisis de regresión lineal múltiple para nuestro modelo. Estos hallazgos fortalecen la confianza en la robustez y adecuación de nuestro modelo a los datos disponibles.

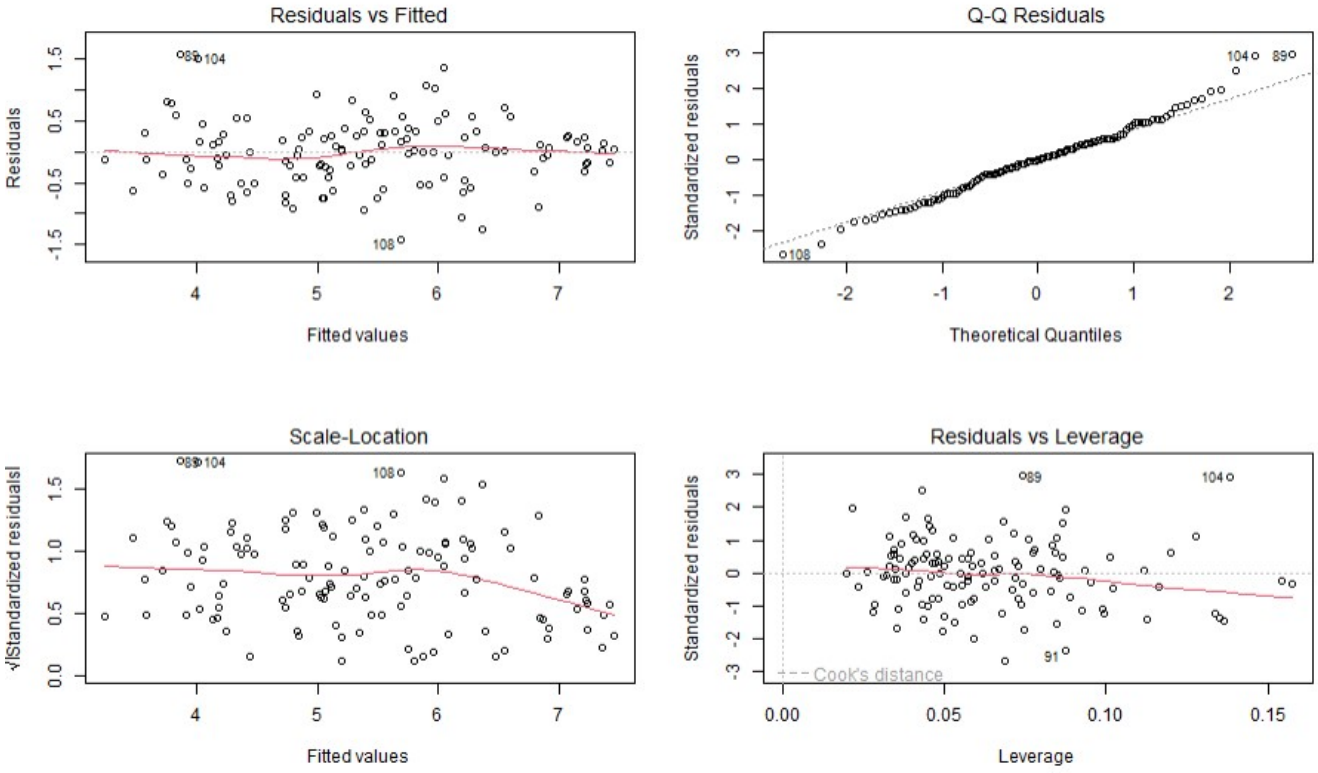

Figura 7: Gráficos del Modelo

El boxplot muestra cuatro valores atípicos dentro de un rango de 0.5 unidades de los bigotes, indicando desviaciones entre predicciones y observaciones en casos particulares. Aunque la mediana en cero sugiere un buen rendimiento general del modelo, la presencia de outliers resalta áreas específicas que podrían necesitar una atención. Este descubrimiento aporta información clave sobre la solidez y confiabilidad del modelo en diferentes situaciones y subconjuntos de datos.

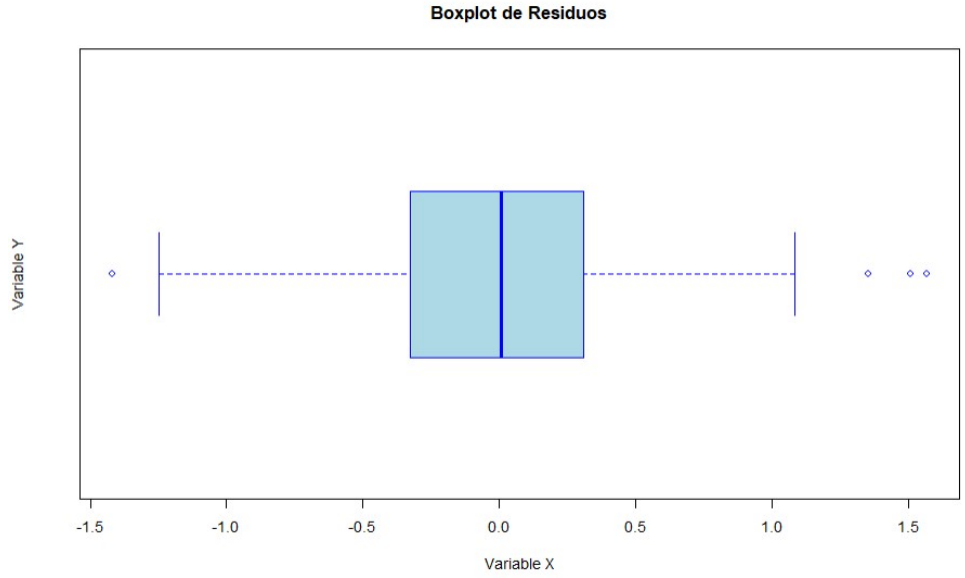

Figura 8: Boxplot de Residuos

La exploración del gráfico de distancia de Cook revela que tenemos varios puntos que superan el umbral de 0.05, indicando cierta influencia en la estimación de los coeficientes del modelo. Específicamente, observamos tres puntos con distancias de Cook entre 0.05 y 0.10, sugiriendo una moderada influencia en la estabilidad del modelo. Adicionalmente, identificamos otro punto con una distancia de Cook aproximada de 0.17, señalando una influencia más pronunciada.

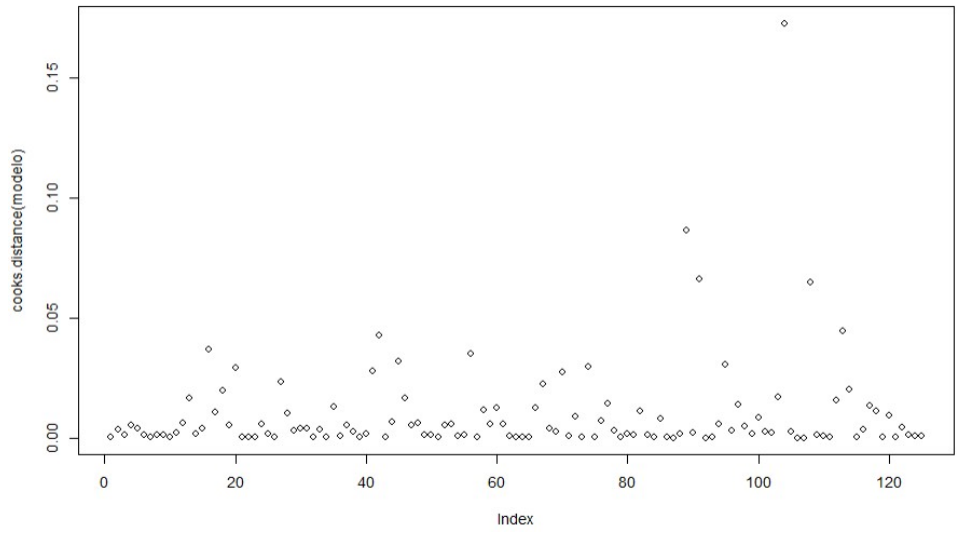

Figura 9: Cooks Distance

6.2. Tests

Test de Durbin-Watson para Independencia de Residuos

Para evaluar la independencia de los residuos, se aplicó el test de Durbin-Watson. El resultado obtenido fue un estadístico DW de 2.1275, con un p-valor de 0.7566. Dado que el p-valor es significativamente alto (cercano a 1), no hay evidencia suficiente para rechazar la hipótesis nula de no autocorrelación entre los residuos. Este hallazgo sugiere que los errores no están correlacionados, respaldando así la validez del supuesto de independencia de residuos.

Test de Normalidad de Shapiro-Wilk para Residuos

La normalidad de los residuos se evaluó mediante el test de Shapiro-Wilk y un gráfico QQplot. El p-valor del test de Shapiro-Wilk fue 0.5379, indicando que no hay suficiente evidencia para rechazar la hipótesis nula de normalidad en los residuos. El gráfico QQplot también respalda visualmente la aproximación a una distribución normal. Estos resultados sugieren que los residuos se ajustan razonablemente bien a una distribución normal.

Evaluación de Multicolinealidad

Para evaluar la multicolinealidad, se calculó el Factor de Inflación de la Varianza (VIF) para cada variable independiente. Los resultados muestran que todos los VIF están por debajo del umbral de 5, indicando una multicolinealidad aceptable. Este hallazgo sugiere que las variables independientes no están altamente correlacionadas entre sí, lo cual es esencial para la validez de la estimación de los coeficientes. En conjunto, estos resultados respaldan la validez de los supuestos del modelo de regresión lineal múltiple.

Variable Independiente	**VIF**
LogPBI	5.074599
ApoyoSocial	2.122405
EsperanzaVidaSaludableNacer	4.457307
LibertadTomarDecisionesVida	2.337508
PercepcionCorrupcion	1.671092
AfectoPositivo	2.048834
ConfianzaGobiernoNacional	1.763040

Tabla 2: Resultados del Factor de Inflación de la Varianza (VIF)

7. Predicción

Para probar nuestro módelo, hacemos uso de nuevos datos para realizar la predicción, estos son:

```
nuevo <- data.frame(
  LogPBI = 7.65,
  ApoyoSocial = 0.52,
  EsperanzaVidaSaludableNacer = 53.2,
  LibertadTomarDecisionesVida = 0.50,
  Generosidad = 0.10,
  PercepcionCorrupcion = 0.87,
  AfectoPositivo = 0.49,
  AfectoNegativo = 0.37,
  ConfianzaGobiernoNacional = 0.40
```

En base a la prueba del modelo utilizando nuevos datos, observamos que al introducir valores casi idénticos a los de la primera fila de los datos originales, la predicción resultante (3.29) fue muy cercana al valor original (3.13). Esta proximidad sugiere que el modelo es capaz de hacer predicciones de buena manera y generalizar bien a partir de datos no vistos previamente. Sin embargo, es crucial tener en cuenta que la ligera diferencia entre la predicción y el valor real podría deberse a diversos factores, puede ser una combinación de variabilidad aleatoria en los datos (ruido) y la capacidad del modelo para capturar todas las complejidades en la relación entre las variables predictoras y la variable de respuesta.

8. Conclusión

En conclusión, la regresión lineal múltiple realizada revela que la satisfacción de vida puede ser explicada en gran medida por factores como el Producto Interno Bruto per cápita, el apoyo social, la esperanza de vida saludable al nacer, la libertad para tomar decisiones de vida, la percepción de corrupción, el afecto positivo y la confianza en el gobierno nacional. El modelo resultante muestra un buen ajuste a los datos, con un R cuadrado ajustado de 0.7779, lo que indica una capacidad predictiva sólida.

El análisis exploratorio destaca relaciones positivas significativas entre la satisfacción de vida y la riqueza económica, el apoyo social, la esperanza de vida y el afecto positivo. Además, la validación del modelo muestra que se cumplen los supuestos fundamentales de la regresión lineal, como la normalidad de los residuos y la independencia entre ellos.

Aunque se identifican algunos outliers, la capacidad del modelo para predecir nuevos datos se demuestra mediante la prueba con valores de entrada adicionales. Este estudio respalda la utilidad de la regresión lineal múltiple como una herramienta efectiva para entender y predecir la satisfacción de vida en función de múltiples variables.

9. Bibliografía

- World Happiness Report - Data
  https://worldhappiness.report/data/