Proyecto 3: Análisis de Artículos Científicos
Introducción a la Ciencia de Datos

Integrantes:	Canché, Elías; Sánchez, Hazel
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas
Profesor:	Dr. Marco Antonio Aquino López

Resumen

En este trabajo se realiza un análisis de lo que son la regresión lineal y la regresión logística. Analizamos algunos artículos donde aplican estos modelos y evaluamos su implementación, tanto de manera teórica como práctica.

1. Introducción

Los modelos de regresión son herramientas de la inferencia estadística aplicada y permiten establecer relaciones funcionales entre una variable de respuesta y un conjunto de variables explicativas. Entre ellos, la regresión lineal y la regresión logística ocupan un lugar central por su interpretación analítica, su fundamentación probabilística y su amplio rango de aplicaciones empíricas.

La regresión lineal tiene supuestos que se deben verificar, el cual uno de ellos, quizá el más importante, es la normalidad de los errores y utilizando estimación por mínimos cuadrados. Por su parte, la regresión logística puede verse como una extensión para variables dependientes binarias, mediante la modelación de la probabilidad de éxito a través de la función logística y la estimación por máxima verosimilitud.

En la práctica científica, ambas metodologías se han convertido en herramientas fundamentales del análisis en diversas disciplinas, no solo por su capacidad de explicar fenómenos, sino también por su capacidad de incorporar inferencia estadística, pruebas de hipótesis y medidas de ajuste. Sin embargo, la correcta especificación de los modelos, la verificación de supuestos y la interpretación de los coeficientes presentan desafíos recurrentes en la literatura aplicada. En este trabajo se realizará un análisis del uso de la regresión lineal y logística en artículos científicos publicados en revistas indexadas, con el objetivo de evaluar el rigor estadístico empleado y discutir la modelación en función de los objetivos de investigación.

2. Regresión Lineal y Regresión Logística

2.1. Regresión Lineal

La regresión lineal es un modelo estadístico que explica la relación entre una variable de respuesta Y y una variable predictora X, a través de una recta. El modelo se expresa como

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

donde β_0 es el valor esperado de Y cuando $X = 0$ (intercepto), β_1 es el cambio esperado en Y por cada unidad de cambio en X (pendiente) y ϵ es el término de error, que representa la variabilidad no explicada por el modelo. En el caso de la regresión lineal múltiple (con p variables independientes), el modelo generaliza a

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon.$$

Los parámetros $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ se estiman mediante el método de mínimos cuadrados ordinarios (MCO), que minimiza la suma de los cuadrados de los residuos. De forma matricial se resuelve como $\hat{\beta} = (X^T X)^{-1} X^T Y$, asumiendo que la matriz de diseño X (de dimensión $n \times (p+1)$) es de rango completo.

Para que la aplicación del modelo sea válida se requiere verificar los siguientes supuestos:

1. **Linealidad**: La relación entre las variables independientes y la dependiente es lineal en los parámetros. Es decir, $E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$.
2. **Independencia de los errores**: Los errores ϵ_i son independientes entre sí.

3. **Varianza constante de los errores**: $\text{Var}(\varepsilon_i) = \sigma^2$ para todo i.

4. **Normalidad de los errores**: Los errores ε_i siguen una distribución normal, $\varepsilon_i \sim N(0, \sigma^2)$.

5. **No multicolinealidad**: Las variables independientes no son linealmente dependientes, asegurando que $(X^T X)$ sea invertible.

Adicionalmente, en contextos de inferencia, se asume que las observaciones son una muestra aleatoria de la población.

2.2. Regresión Logística

La regresión logística es un modelo de regresión generalizado, utilizado para predecir variables dependientes categóricas, usualmente binarias (por ejemplo, éxito/fracaso, sí/no). A diferencia de la regresión lineal, la regresión logística modela la probabilidad de que la variable dependiente tome un valor específico mediante una función logística (o sigmoide).

Sea $Y \in \{0, 1\}$ una variable binaria dependiente de otra variable X. El modelo supone que Y sigue una distribución de Bernoulli con parámetro $p = P(Y = 1|X)$, y se expresa como

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p,$$

resolviendo para p,

$$p = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p))}.$$

Los parámetros β se estiman mediante máxima verosimilitud (MV), maximizando la función de log-verosimilitud

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)],$$

lo que requiere métodos numéricos ya que no hay solución cerrada. Para casos con más de dos categorías, se generaliza a la regresión logística multinomial.

Para que la aplicación del modelo sea válida se requiere verificar los siguientes supuestos:

1. **Linealidad**: La relación entre las variables independientes y el logaritmo de la probabilidad es lineal.

2. **Independencia de las observaciones**: Las observaciones son independientes.

3. **No multicolinealidad**: Las variables independientes no son linealmente dependientes.

4. **Tamaño de muestra adecuado**: Se requiere un número suficiente de eventos por predictor (regla heurística: al menos 10 eventos por variable).

5. **Distribución binomial**: La variable dependiente sigue una distribución binomial (o Bernoulli para binaria).

Observación: Los supuestos son similares pero adaptados al contexto binomial y a diferencia de la regresión lineal, no se asume normalidad ni varianza constante de los errores, ya que en el modelo, la varianza de Y es $p(1 - p)$, lo cual es inherentemente dependiente de p.

En las matemáticas se enuncian supuestos, considerados ciertos (hipótesis) para luego dar consecuencia sobre estos. Si partiendo de supuestos se demuestra algún resultado, entonces cada que este supuesto se cumpla, el resultado también se cumplirá. Sin embargo, afirmarlo de manera contraria (el resultado implica los supuestos) es un razonamiento erróneo y frecuente entre las personas, más aún a esto se le conoce como la **falacia del consecuente**.

Especialmente en estos modelos, validar los supuestos es crucial ya que, de no cumplirse, puede llevar a estimaciones sesgadas, varianzas infladas o inferencias inválidas, lo que compromete la fiabilidad del modelo. Por ejemplo, en la regresión lineal, la varianza no constante de los errores, puede subestimar los errores estándar, llevando a pruebas de significancia erróneas (falsos positivos).

3. Revisión de Artículos

En esta sección veremos artículos científicos publicados en revistas indexadas; es decir, artículos en revistas científicas de alta calidad que ha sido incluida en una o más bases de datos o directorios de referencia bibliográfica a nivel mundial, con el fin de verificar la correcta implementación de los modelos.

3.1. Artículos que utilizaron Regresión Lineal

3.2. Water quality predictions through linear regression - A brute force algorithm approach

Este artículo fue elegido por una razón principal, además de haber sido publicado en una revista indexada, esta es por la utilidad de prevención sobre enfermedades provocadas por una mala calidad del agua. Si se puede conocer algún tipo de información sobre la calidad del agua, se podrían implementar acciones preventivas para minimizar enfermedades y más aún, si el modelo permite la información, se podrían tener acciones correctivas para mejorar la calidad del agua.

La cuenca del Río Ave (Ave River Basin - ARB), en el noroeste de Portugal, es una zona conocida por su mala calidad del agua desde hace décadas. El objetivo de este artículo es desarrollar un método para predecir parámetros de calidad del agua (SWP) en todas las subcuencas de la ARB, incluso en lugares donde no hay estaciones de medición. Se sabe que es más fácil y económico medir las características del territorio (usos del suelo, fuentes de contaminación) que monitorear directamente la calidad del agua en todos los puntos de la red hídrica. Por lo tanto, si se encuentra una relación sólida entre estas características del territorio (variables independientes) y los parámetros de agua (variables dependientes), se pudieran hacer predicciones sobre la calidad del agua.

El problema no es solo predecir la calidad del agua, sino cómo seleccionar el mejor modelo de regresión lineal de manera objetiva, rigurosa y automatizada cuando se tiene un número muy grande de variables predictoras. El artículo ofrece un método para ajustar un mejor modelo de regresión lineal.

3.2.1. Análisis Crítico

La base de datos consta de 29 muestras, lo cual hace que la regresión lineal sea un modelo candidato a utilizar. El modelo considera como variables dependientes los siguientes parámetros de calidad del agua: Conductividad eléctrica (EC), pH, Nitrato total, Alcalinidad total. Mientras que existen 73 variables potenciales; adicionalmente se realizaron transformaciones a alguna variables, por lo que se terminó con 365 variables.

Todos los supuestos se cumplen y fueron verificados de la siguiente manera.

- **Linealidad:** Asumida en el modelo de regresión lineal ordinaria (OLS).
- **Varianza cosntante de los errores:** Verificada con 4 pruebas: Breusch-Pagan, Harvey-Collier, Glejser y Goldfeld-Quandt.
- **Normalidad de los errores:** Verificada con 5 pruebas: Jarque-Bera, Anderson-Darling, Shapiro-Wilk, Kolmogorov-Smirnov y Omnibus.
- **Independencia de errores:** Se verificó la autocorrelación espacial con la prueba de Moran's I.
- **No multicolinealidad:** No se probó que las variables descriptivas sean independientes de las demás. Por la naturaleza de las variables, quizá sea suficiente con una descripción de ellas para poder argumentar no multicolinealidad, pero no realizó.

Los coeficientes de regresión se interpretaron según el signo esperado (positivo o negativo) basado en el conocimiento del dominio. Por ejemplo, un coeficiente positivo para emisiones difusas indica un aumento en la contaminación, mientras que un coeficiente negativo para áreas forestales sugiere un efecto positivo en la calidad del agua.

En este contexto, podemos afirmar que la regresión lineal fue un método adecuado y su implementación fue correcta, por que la muestra era pequeña ($n = 29$), es un método interpretable y aceptado en el contexto del problema, se validaron supuestos mediante un script automatizado. Como observación adicional, se complementó con Geographically Weighted Regression (GWR) para capturar variaciones espaciales, aunque los resultados fueron muy similares a OLS.

Los autores de este artículo de *Open Access*, pusieron a disposición del público en general los códigos (uno para cada variable de respuesta) de Python con los cuales se realizó el análisis, por lo que se procedió a revisarlos.

En general, los autores de este artículo justificaron e implementaron de manera correcta la regresión lineal, más aún fueron transparentes en la implementación del modelo citando y verificando los supuestos, además de proporcionar el script en Python para revisar los resultados. Sin emabrgo, podemos comentar algunas limitantes como el tiempo de cómputo el cual es alto (entre 10 y 60 horas por parámetro), el enfoque de fuerza bruta prueba millones de combinaciones y al replicar los resultados pudimos comprobar que el método tarda considerablemente, por lo que el método no es escalable; con cientos de variables, el método se vuelve inviable. Quizá la crítica más fuerte es la transformación de las variables, al transformarlas e incluir todas las variables, resultan correlaciones que no se contemplaron, lo cual podría perjudicar al ajuste final del método.

Algunas opciones que se podrían utilizar para este problema son métodos como Ridge o Lasso por el manejo de multicolinealidad y selección automática de variables.

El códgigo del análisis se encuentra en el Appendice B: Material Suplementario del url original del artículo y su cita APA es la siguiente:

Fernandes, A. C. P., Fonseca, A. R., Pacheco, F. A. L., & Sanches Fernandes, L. F. (2023). Water quality predictions through linear regression - A brute force algorithm approach. MethodsX, 10, 102153. https://doi.org/10.1016/j.mex.2023.102153

mientras que los códigos, la revisión de estos y el artículo se pueden encontrar en la carpeta `LR Art 1` del repositorio de Github de este reporte.

3.2.2. Influential Factors on California Regional Housing Price Analysed by Multiple Linear Regression.

El análisis de factores que influyen en el precio de la vivienda representa un problema fundamental en econometría y planificación urbana. Este reporte compara dos aproximaciones al mismo problema: una implementación práctica en Python y un estudio académico formal publicado en proceedings de conferencia. Ambas investigaciones utilizan datos del mercado inmobiliario de California pero difieren en aspectos metodológicos clave.

Ambos estudios fueron realizados con el mismo set de datos del mercado de viviendas de California, pero con algunas diferencias en el tratamiento de datos. A continuación comparamos estas diferencias:

Cuadro 1: Comparación de Estrategias de Preprocesamiento

Implementación Propia	**Articulo de referencia**
Relleno de valores faltantes en *total_bedrooms* con la mediana	Eliminación de todas las observaciones con valores faltantes
Justificación: Valores MAR (Missing At Random)	No se especifica justificación para eliminación
Preservación de outliers identificados en boxplots	Muestra aleatoria de 500 observaciones sin mención de outliers

En ambos caso se tomo una muestra aleatoria de 500 observaciones y una división 400-100 para entrenamiento y prueba respectivamente.

Cuadro 2: Comparación de Métricas del Modelo

Métrica	**Impl. Propia**	**Artículo (Train)**	**Artículo (Test)**	**Diferencia**
R^2 / R^2 ajustado	0.5633	0.592	0.637	+5.1 % / +13.1 %
MAE	58,159.22	No reportado	No reportado	-
MSE	5.9458e9	No reportado	No reportado	-
RMSE	77,109.04	No reportado	No reportado	-
AIC	No calculado	8,051.3	3,094.1	-

El diagnóstico del Modelo propio es el siguiente:
Ambos estudios convergen en hallazgos clave:

Cuadro 3: Comparación de Coeficientes del Modelo Final

Variable	**Impl. Propia**	**Artículo (Modelo 2)**	**Diferencia**	**Dirección**
Intercept	-61,205	-57,233	-6.9 %	Consistente
Median Income	46,623.69	45,309	+2.9 %	Consistente (+)
Housing Median Age	1,661.97	2,460.26	-32.5 %	Consistente (+)
Households	291.18	190.80	+52.6 %	Consistente (+)
Population	-65.19	-53.77	+21.2 %	Consistente (-)
Total Rooms	-15.05	Excluída	-	Divergente

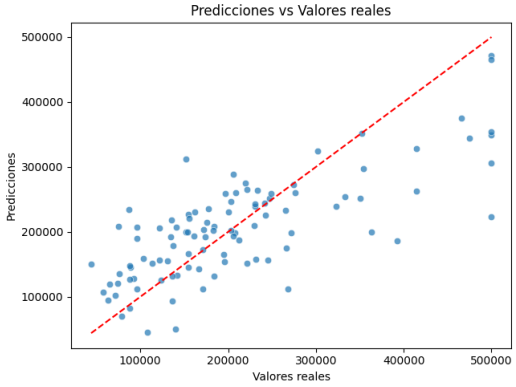

Figura 1: Predicciones vs Valores Reales - Implementación Propia

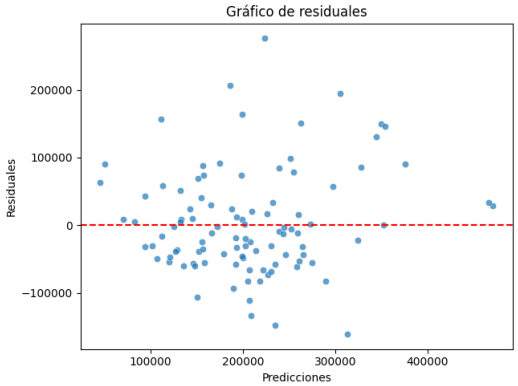

Figura 2: Gráfico de Residuales - Implementación Propia

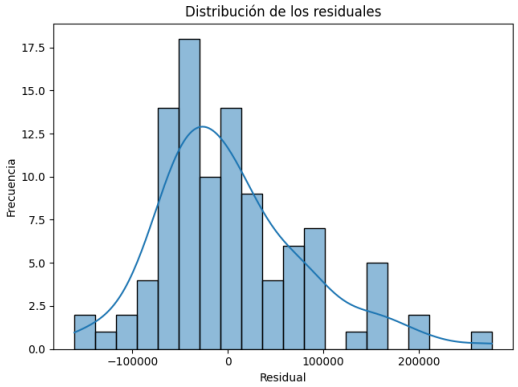

Figura 3: Distribución de Residuales - Implementación Propia

- **Ingreso medio como factor dominante**: Ambos modelos identifican el ingreso medio como la variable más influyente, con coeficientes muy similares (46,623 vs 45,309).
- **Consistencia en direcciones**: Todas las variables comunes mantienen el mismo signo en ambos estudios,

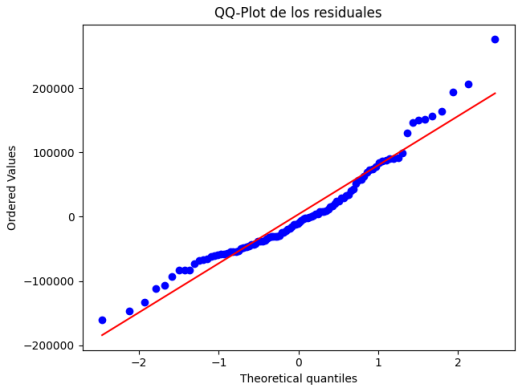

Figura 4: QQ-Plot de Residuales - Implementación Propia

lo que sugiere relaciones direccionales robustas.

- **Significancia estadística**: Las variables incluidas en ambos modelos finales muestran significancia estadística.

Las diferencias observadas pueden atribuirse a varios factores metodológicos:

- *Tratamiento de Valores Faltantes:* La implementación propia utilizó imputación por mediana para *total_bedrooms*, mientras el artículo original eliminó observaciones incompletas. Esto podría introducir sesgos diferentes en cada modelo.
- *Validación de Supuestos:* El artículo original realizó verificaciones exhaustivas de multicolinealidad (VIF) y supuestos de regresión, mientras la implementación propia se limitó a diagnósticos básicos de residuales por simplicidad.
- *Poder Predictivo:* La diferencia en R^2 (56.3 % vs 63.7 %) sugiere que la metodología más rigurosa del articulo original produce modelos con mejor capacidad explicativa.

Ambos estudios presentan limitaciones importantes que deben ser consideradas en la interpretación de los resultados. En primer lugar, la generalizabilidad de estos estudios está límitada por el uso exclusivo de datos de California, lo que impide generalizar las conclusiones a otros contextos geográficos con características demográficas, económicas y regulatorias diferentes. Esto ya que ambos modelos tienen el problema de omitir variables del dataset original, ya que factores críticos como la ubicación geográfica específica, la calidad de construcción, la proximidad a servicios esenciales, amenities locales y variables macroeconómicas regionales no fueron incorporados en los análisis.

Finalmente, la reducción del tamaño muestral a 500 observaciones, aunque metodológicamente justificada, puede afectar la estabilidad de los coeficientes estimados y reducir la potencia estadística para detectar relaciones significativas, particularmente en subgrupos poblacionales o para variables con efectos moderados.

La comparación revela que, si bien ambos enfoques identifican patrones fundamentales similares en los determinantes del precio de la vivienda, la metodología más rigurosa del estudio académico produce resultados más confiables y con mayor poder explicativo. Los códigos y el análisis realizado está en la carpeta `Segunda Parte` del repositorio de GitHub de este reporte y la cita APA de este artículo es la siguiente:

Cao, Y. (2024). Influential factors on California regional housing price analysed by multiple linear regression. Advances in Economics, Management and Political Sciences, 77, 48–53. EWA Publishing.

3.3. Artículos que utilizaron Regresión Logística

3.3.1. Logistic regression technique for prediction of cardiovascular disease.

Este artículo fue elegido debido a la importancia de la detección oportuna de enfermedades. El artículo aplica **regresión logística** para predecir la precencia de alguna enfermedad cardiovascular usando el dataset `Cleveland` del UCI (303 registros, 13 características), cual es un dataset público.

3.3.2. Análisis Crítico

La variable dependiente es una binaria, la cual es 1 en presencia de enfermedad y 0 en ausencia de enfermedad. Se seleccionaron las variables predictoras, con mayor correlación positiva con la variable objetivo, según la Tabla 1 del artículo. Exang, Cp, Oldpeak, Thalach, Ca, Slope. Esto es crítico, pues puede llegar a contradecir el modelo, pues no se cumpliría el supuesto de linealidad. Más aún, el artículo no verifica ningún supuesto más que la independencia de las observaciones; en efecto, cada observación es un paciente y la información de un paciente no influye en otros.

Dado que se requiere resolver un problema de clasificación binaria y no modelar relaciones complejas, elegir la regresión logística es buena opción. Sin embargo, la implementación rigurosa deja mucho que desear, pues no se verificaron los supuestos del modelo, no se evaluó la calibración del modelo, no se exploraron interacciones entre variables, solo se utilizó una base de datos, no se reportan intervalos de confianza ni significancia estadística, no se menciona validación cruzada, solo división entrenamiento/prueba y existe la posibilidad de desbalance de clases, pues no se menciona si las clases están balanceadas. Todo el análisis está sobre la precisión y dadas las deficiencias del análisis, esto puede llevar a interpretaciones como que el análisis fue exclusivo para mejorar la precisión de en esa base de datos, lo cual no aporta nada al campo de la investigación, pues esta base de datos ya tiene identificado qué pacientes tuvieron enfermedad; el método podría ser inservible en datos cuya clasificación real no se conozca.

Para las limitaciones se recomienda explorar lo descrito en la tabla 4

Cuadro 4: Limitaciones del análisis y alternativas propuestas

Limitación	**Alternativa**
No verificación de supuestos	Regresión logística penalizada (Ridge/Lasso) para manejar multicolinealidad
Posible no linealidad	Modelos aditivos generalizados (GAM) o transformaciones de variables
Solo un dataset	Validación cruzada repetida o validación externa con otro dataset
Poco tamaño de muestra	Regresión logística bayesiana para incorporar información previa y mejorar estimaciones con poca data
Desbalance de clases	Balanceo de datos (SMOTE) o uso de F1-score como métrica principal

Por otro lado, se descargó la base de datos utilizada y se realizó una réplica del experimento de manera independiente. Cabe mencionar que en la base de datos existen pocos valores faltantes, que al parecer son MAR, por lo tanto en nuestro análisis se procede a imputarlos con la mediana (robusto para outliers). Por su parte, los autores no mencionan el tratamiento de los valores faltantes, ellos afirman que el dataset çontiene información completa"después del preprocesamiento, dado que conocemos la base de datos (es pública) sabemos que hay valores faltantes, por lo que podremos suponer que eliminaron los registros con valores faltantes.

Al replicar los resultados se obtuvo lo siguiente. En la tabla 5 se pudo obtener que las mismas variables que en la tabla 1 del artículo, pero con otro coeficiente de correlación. En la tabla 6 se pueden ver los resultado para diferentes proporciones entre conjunto de entrenamiento y conjunto de prueba, hay diferencia para las proporciones 60/40 y 70/30, esto se puede dar por la imputación de datos. Por último, en la figura 5 podemos observar cierta similitud con la figura 3 del artículo original, nuevamente, esta diferencia puede deberse a la imputación de datos.

Cuadro 5: Selección de variables por correlación con la variable objetivo

Variable	Thal	Ca	Exang	Oldpeak	Cp	Slope
Correlación	0.522057	0.460033	0.431894	0.424510	0.414446	0.339213

Como exploración alternativa se implementó una regresión logística con regularización $L1/L2$ (con la misma imputación de datos), cuya gráfica de residuos es la figura 6, lo cual indica que quizá no sea un buen modelo.

Cuadro 6: Resumen de resultados del modelo de regresión logística

Entrenamiento (%)	**Prueba (%)**	**Exactitud**	**AUC**
50	50	0.815789	0.890885
60	40	0.795082	0.888874
70	30	0.780220	0.867006
80	20	0.852459	0.904634
90	10	0.870968	0.940171

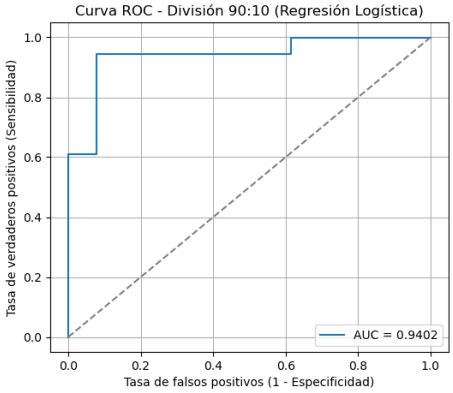

Figura 5: Curva ROC para la proporción 90/10

En conclusión, el artículo aplica regresión logística de manera estándar para predecir enfermedad cardiovascular, con una precisión competitiva (87.10 %), pero carece de rigor matemático en la verificación de supuestos y profundidad en el análisis de los coeficientes. Sería recomendable complementar con técnicas de validación más robustas y explorar modelos que capturen relaciones no lineales o incluyan regularización.

El código implementado para replicar los resultados se encuentra en la carpeta `Log R Art 2` en el repositorio de GitHub de este reporte, mientras que la cita APA de este artículo es el siguiente.

G., A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. *Global Transitions Proceedings*, 3, 127–130. https://doi.org/10.1016/j.gltp.2022.04.008

4. Conclusiones

Las herramientas que nos brinda la estadística son muy útiles y sus aplicaciones pueden llegar a ser extraordinarias, sin embargo, como no dejan de ser objetos matemáticos, es fundamental verificar que el modelo sea adecuado y más aún, se deben verificar que se cumplan los supuestos; de esta manera podremos utilizar todo el potencial de estas herramientas. Estos artículos sirven como evidencia de que un modelo bien implementado da buenos resultados, pero también muestran que sin la interpretación correcta, el modelo no es aplicable.

Figura 6: Grafica de residuos: Heart Disease