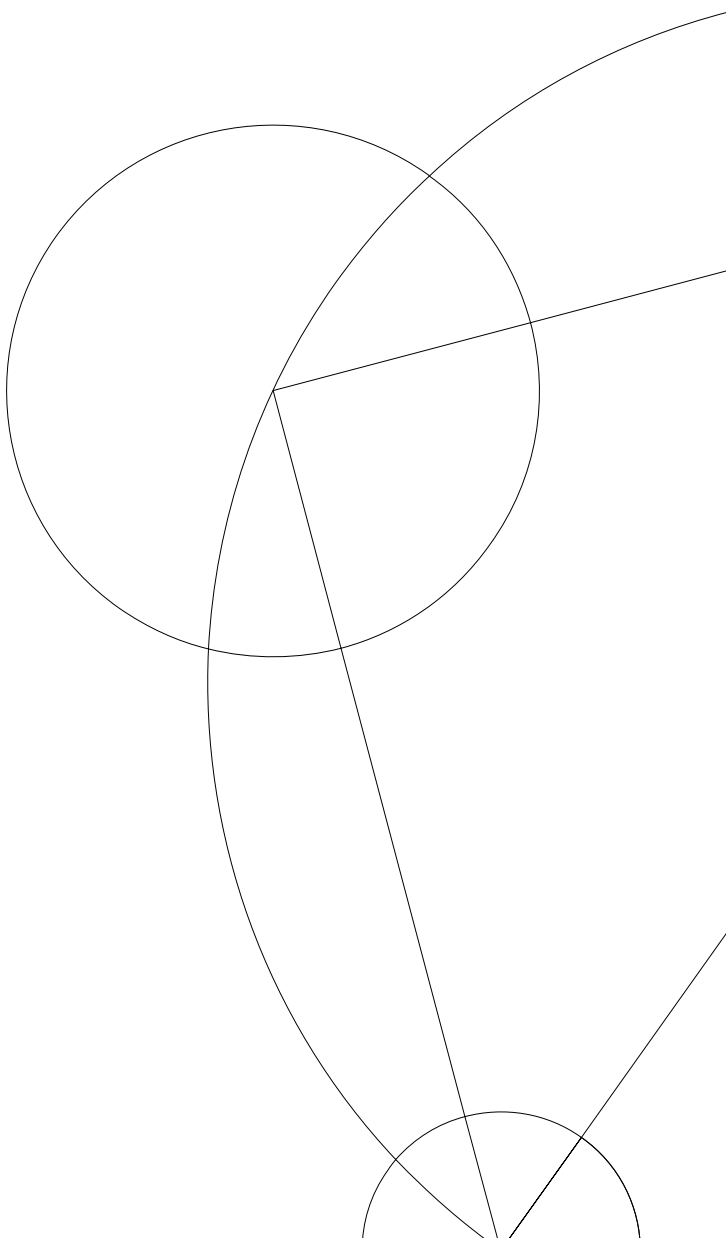
Análisis de Modelos en Artículos Académicos

Alfredo Bistrain y Jesús García

Introducción a la Ciencia de Datos

Tarea 3



October 14, 2025

Introducción

Introducción

El uso de modelos de regresión constituye una de las herramientas fundamentales en la investigación científica aplicada, pues permite describir, cuantificar y predecir relaciones entre variables bajo un marco estadístico formal. No obstante, la validez de los resultados inferenciales y predictivos depende críticamente de que los modelos ajustados cumplan los supuestos teóricos que los sustentan. En el caso de la regresión lineal, el teorema de Gauss–Markov establece las condiciones bajo las cuales los estimadores obtenidos mediante mínimos cuadrados ordinarios (OLS) son insesgados y eficientes dentro de la clase de estimadores lineales. De forma análoga, la regresión logística exige la correcta especificación funcional del modelo, la independencia de las observaciones y la ausencia de multicolinealidad para garantizar inferencias válidas sobre las probabilidades estimadas.

El presente trabajo tiene como objetivo analizar el cumplimiento de dichos supuestos en dos estudios científicos recientes que emplean modelos de regresión con propósitos explicativos y predictivos. El primero, titulado *Modelling the effect of prebiotics, probiotics and other functional additives on the growth, feed intake and feed conversion of European sea bass*, aplica un modelo de regresión lineal múltiple para evaluar los efectos de distintos aditivos funcionales en el crecimiento de la lubina europea. El segundo, *Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches*, utiliza regresión logística para identificar los principales factores de riesgo asociados con la diabetes mellitus tipo 2 en la población indígena Pima.

A partir de la replicación y el análisis de ambos modelos, se evalúan empíricamente los supuestos fundamentales de cada enfoque, incluyendo la linealidad, homocedasticidad, independencia y normalidad de los errores en el caso lineal, así como la especificación funcional, la multicolinealidad, la influencia de observaciones atípicas y la calibración en el modelo logístico. Además, se aplican pruebas de bondad de ajuste y criterios de información (AIC, BIC y R^2 ajustado) para valorar la calidad global de los modelos y la significancia de sus predictores.

1 Modelling the effect of prebiotics, probiotics and other functional additives on the growth, feed intake and feed conversion of European sea bass (*Dicentrarchus labrax*)

Se inicia con el artículo *Modelling the effect of prebiotics, probiotics and other functional additives on the growth, feed intake and feed conversion of European sea bass (Dicentrarchus labrax)*, en el cual se describe que la industria acuícola ha enfrentado en los últimos años nuevos desafíos para lograr una producción más económica y ambientalmente sostenible. En este contexto, se han desarrollado diversas estrategias dietéticas basadas en fuentes alternativas de proteínas que buscan reducir la dependencia de las materias primas marinas.

No obstante, la formulación de alimentos utilizando materias primas terrestres u otras fuentes alternativas de proteínas, como subproductos de origen terrestre o insectos, puede generar desequilibrios nutricionales que afecten negativamente el crecimiento y la salud de los peces. En conjunto, estos efectos pueden derivar en un aumento de los costos de producción, menores tasas de crecimiento y mayores requerimientos de alimento a lo largo del ciclo productivo.

Los efectos adversos asociados a la reducción de harina de pescado (FM) y aceite de pescado (FO) en las dietas resultan particularmente relevantes en especies carnívoras como la lubina europea (*Dicentrarchus labrax*). En algunos estudios se ha reportado que es posible reducir el nivel de FM en las dietas sin comprometer significativamente el crecimiento, en comparación con una dieta control con un 31.5% de FM. Sin embargo, la salud y el bienestar de los peces pueden verse comprometidos por este reemplazo, disminuyendo su resistencia a patógenos y su tolerancia al estrés.

En este sentido, la suplementación dietética con aditivos funcionales se presenta como una estrategia eficaz para contrarrestar dichos efectos negativos. Los aditivos funcionales son compuestos con la capacidad de mejorar el rendimiento del crecimiento, la salud y el bienestar de los peces mediante el aumento, por ejemplo, de la digestibilidad de los nutrientes. Sin embargo, la amplia variedad de ingredientes funcionales disponibles, junto con la diversidad de efectos asociados a los niveles de inclusión y

mecanismos aún no completamente comprendidos, dificulta la selección de estrategias adecuadas para la suplementación dietética.

Por otro lado, las condiciones experimentales de los diferentes estudios que analizan los efectos de los ingredientes funcionales sobre la salud y el crecimiento de los peces son altamente variables, lo que impide la comparación directa de los resultados obtenidos. En este contexto, el modelado matemático surge como una herramienta útil para analizar el crecimiento de los peces y la eficiencia en el uso del alimento bajo diferentes condiciones experimentales y composiciones de dieta.

En particular, se recurre a la regresión lineal múltiple como un método capaz de identificar los efectos de factores que varían entre estudios (factores de confusión), permitiendo obtener estimaciones normalizadas de la respuesta de los peces a distintos niveles de inclusión de ingredientes funcionales.

De esta manera, el objetivo principal del artículo fue desarrollar un modelo observacional de regresión lineal múltiple que permitiera aislar de forma robusta los efectos de los ingredientes funcionales dietéticos sobre los parámetros de crecimiento y utilización del alimento en juveniles de lubina europea, considerando simultáneamente los resultados de múltiples ensayos de crecimiento realizados en distintos contextos experimentales.

1.1 Criterios de selección de muestras

El artículo señala que se realizó una búsqueda bibliográfica en las bases de datos *Web of Science* y *Scopus*, utilizando las siguientes combinaciones de términos en título, resumen y palabras clave: {"Dicentrarchus labrax", "lubina europea", "European seabass"}, {"juveniles", "alevines"}, y {"alimentos funcionales", "ingredientes funcionales", "dietas funcionales", "suplementación dietética", "probióticos", "prebióticos", "fitogénicos", "aceites esenciales", "compuestos de origen vegetal", "aditivos fitogénicos para alimentación", "simbióticos"}. Los estudios seleccionados debían cumplir simultáneamente los siguientes criterios:

1. Experimentos realizados con juveniles de lubina europea.
2. Suplementación dietética con probióticos, prebióticos y/o compuestos de origen vegetal.
3. Reporte de al menos un parámetro de crecimiento o eficiencia alimenticia —tasa de crecimiento específico (SGR), tasa de conversión alimenticia (FCR) o ingesta de alimento (FI),
4. Información sobre peso corporal inicial (IBW) y final (FBW) con una medida de dispersión (por ejemplo, desviación estándar).
5. Duración del ensayo de alimentación.
6. Composición de los tratamientos dietéticos o, en su defecto, composición proximal con concentraciones de proteína (diet_CP) y energía (diet_GE).
7. Condiciones de cultivo, incluyendo al menos temperatura media del agua y oxígeno disuelto.

1.2 Acondicionamiento y análisis de datos

Toda la información recopilada se estandarizó a unidades comunes, expresando las variables en las siguientes magnitudes: peso y longitud corporal (g, cm), peso corporal promedio (ABW, g), ingesta individual de alimento (g por pez día^{-1}), temperatura del agua (°C), oxígeno disuelto (ppm), volumen del tanque (L), duración (días), composición de ingredientes y análisis proximal de la dieta (%, g/kg, MJ/kg).

Los aditivos funcionales se agruparon en tres categorías generales: *probióticos*, *prebióticos* y *otros* (compuestos de origen vegetal o simbióticos).

Explican que, dado que los peces son animales poiquilotermos, variables como la temperatura del agua influyen significativamente en la ingesta de alimento y el crecimiento. Por ello, las respuestas fueron normalizadas para eliminar el efecto de variables de confusión relevantes (tamaño y temperatura del pez) mediante la expresión:

$$\text{Rasgo normalizado} = \frac{\text{valor crudo medido del rasgo}}{\text{valor máximo del rasgo}}.$$

El valor crudo del rasgo se obtuvo a partir de los siguientes cálculos:

$$\text{SGR (día}^{-1}) = \frac{\ln(\text{FBW}) - \ln(\text{IBW})}{\text{días}}, \quad \text{FI (\% peso corporal/día)} = \frac{\text{ingesta individual}}{(\text{IBW} + \text{FBW})/2 \times \text{días}} \times 100,$$

$$\text{FCR (g alimento/g ganancia de peso)} = \frac{\text{ingesta individual}}{\text{FBW} - \text{IBW}}.$$

Table 1: Variables cuantitativas explicativas empleadas para el ajuste de los modelos completos de regresión lineal.

Abreviatura	Variable cuantitativa	Unidades
Temp	Temperatura	°C
Oxygen	Oxígeno disuelto	ppm
ABW	Peso corporal promedio	g
FI_norm	Ingesta de alimento individual normalizada	% peso corporal/día
diet_CP	Proteína cruda dietética	g/kg peso seco
diet_GE	Energía bruta dietética	MJ/kg
diet_CP/GE	Relación Proteína/Energía	g/MJ
diet_CL	Lípido crudo dietético	g/kg peso seco
diet_moisture	Humedad de la dieta	%
diet_ash	Cenizas dietéticas	g/kg peso seco
diet_Prebiotics	Prebióticos dietéticos	g/kg peso seco
diet_Probiotics	Probióticos dietéticos	g/kg peso seco
diet_Others	Aditivos dietéticos "otros"	g/kg peso seco

1.3 Resultados del modelo presentados en el artículo.

Quince estudios cumplieron los criterios de inclusión, conformando una base de datos con 61 tratamientos dietéticos: 12 con prebióticos, 13 con probióticos, 21 con otros aditivos funcionales y 15 dietas control sin suplementación.

El análisis mediante regresión lineal múltiple permitió describir las relaciones entre los factores ambientales, las características de la dieta y tres parámetros de desempeño en juveniles de lubina europea (*Dicentrarchus labrax*): la tasa de crecimiento específico (SGR), la ingesta individual de alimento (FI) y la tasa de conversión alimenticia (FCR). Todos los modelos presentaron valores de p significativos y coeficientes de determinación elevados (R^2 entre 0.90 y 0.97; R^2_{ajustado} entre 0.80 y 0.95), lo que evidencia un ajuste adecuado y una buena capacidad explicativa.

Modelo de crecimiento específico (SGR). El modelo ajustado para el SGR normalizado obtuvo $R^2 = 0.96$, $R^2_{\text{ajustado}} = 0.92$ y $p = 7.2 \times 10^{-8}$, siendo el de mejor desempeño global. Las variables más influyentes fueron la temperatura, el oxígeno disuelto y la relación proteína–energía de la dieta, todas con efectos positivos sobre el crecimiento. La inclusión de prebióticos como variable dietética mejoró significativamente los criterios de selección frente al modelo simple ($R^2 = 0.67$; $R^2_{\text{ajustado}} = 0.65$), lo que refuerza su papel como factor determinante en el crecimiento. En contraste, los probióticos y otros aditivos no mostraron efectos significativos sobre este parámetro.

Modelos de FI y FCR. El modelo de FI alcanzó un $R^2 = 0.97$ y $R^2_{\text{ajustado}} = 0.95$, mientras que el de FCR obtuvo $R^2 = 0.90$ y $R^2_{\text{ajustado}} = 0.80$, ambos con $p < 10^{-5}$. En estos modelos, la relación proteína–energía y la energía bruta dietética fueron las principales covariables asociadas a la eficiencia alimenticia. Los tratamientos con prebióticos tendieron a mostrar una menor FI y un menor FCR, coherentes con una mejor conversión alimenticia.

Síntesis e interpretación. Los resultados globales confirman la relevancia del equilibrio nutricional en el rendimiento productivo de la lubina europea (Azevedo et al., 2002; Oliva-Teles, 2012; Méndez-Martínez et al., 2021). En particular, el modelo de SGR —que será el foco del análisis de replicación— evidenció que los prebióticos generan un efecto positivo sobre el crecimiento, correlacionándose de forma directa con el SGR ($\beta = 0.32$) y de manera inversa con la FI y la FCR ($\beta = -0.44$ en ambos casos). Este patrón coincide con lo reportado por Torrecillas et al. (2011) y respalda la hipótesis de que la suplementación con prebióticos mejora la eficiencia metabólica y el aprovechamiento del alimento en los peces, promoviendo así una acuicultura más sostenible.

1.4 Réplica del modelo

1.4.1 Evaluación del preprocesamiento de datos

En este modelo se encontraron algunos problemas no mencionados en el artículo. Estos son manejo de datos faltantes, outliers, puntos de influencia y el criterio con el que se hizo la selección de variables hacia atrás. Iniciando con los datos faltantes, algunos de los consumos diarios y porcentajes de la dieta eran faltantes. Estos se imputaron con medias asegurando que la suma por día fueran del 100%, en el caso de los porcentajes. Es importante mencionar que algunos de los datos originales sumaban más o menos del 100%.

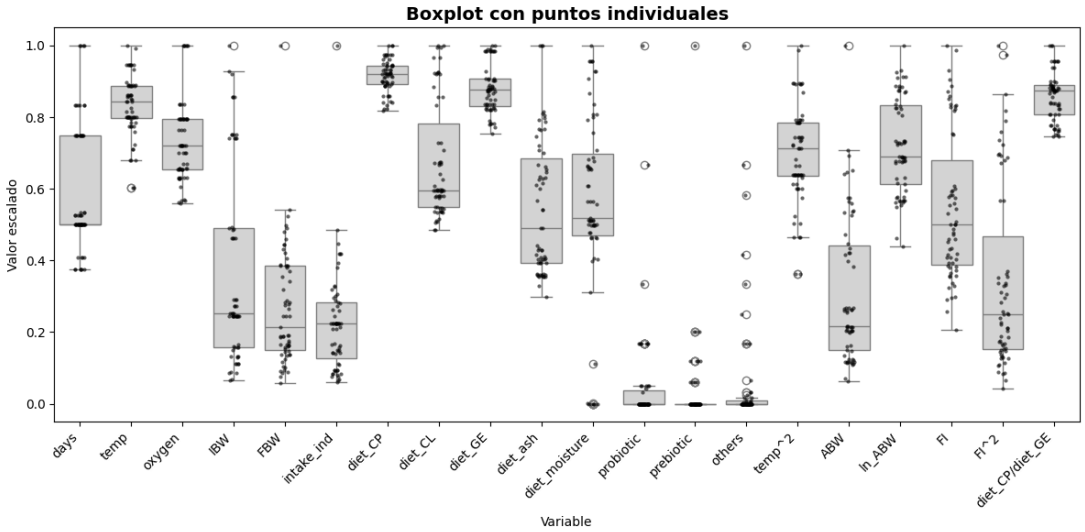


Figure 1: Distribución de las variables explicativas tras el escalamiento (división entre el valor máximo). Se incluyen los valores individuales para visualizar la dispersión y posibles valores atípicos.

La Figura 1 muestra los diagramas de caja de las variables explicativas después del escalamiento, procedimiento que permitió comparar las magnitudes relativas de los predictores dentro de un rango común $[0, 1]$ y evitar que diferencias de escala afectaran los coeficientes estimados. La mayoría de las variables presentan una dispersión moderada, mientras que *probiotic* y *prebiotic* exhiben distribuciones más concentradas y algunos valores atípicos, posiblemente asociados con efectos experimentales específicos o limitada variabilidad entre estudios.

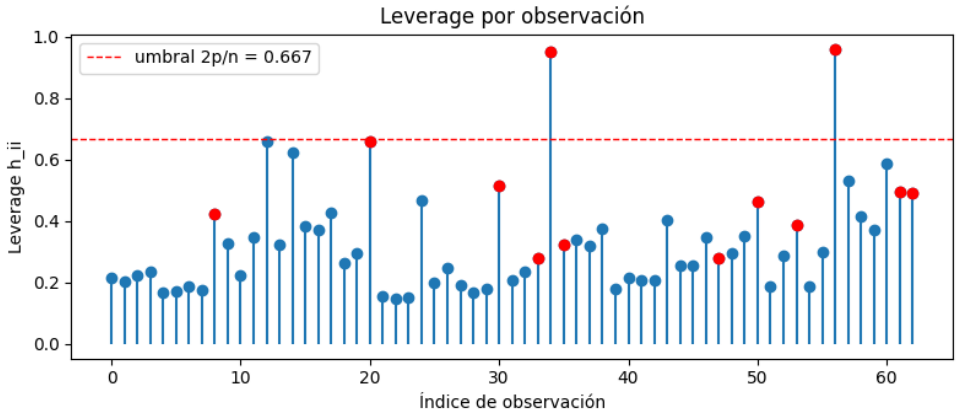


Figure 2: Valores de leverage (h_{ii}) por observación. La línea discontinua roja indica el umbral $2p/n$, utilizado como criterio práctico para identificar observaciones potencialmente influyentes.

En la Figura 2 se presentan los valores de *leverage* (h_{ii}) derivados de la matriz sombrero del modelo, que cuantifican la influencia potencial de cada observación sobre el ajuste de los coeficientes. El umbral $2p/n$ (con p el número de columnas y n el número de filas), indicado por la línea discontinua roja, sirve

como referencia práctica para detectar puntos influyentes. Aunque algunas observaciones superan este límite, la mayoría se sitúa por debajo, lo que indica que el modelo no está dominado por valores extremos en el espacio de predictores.

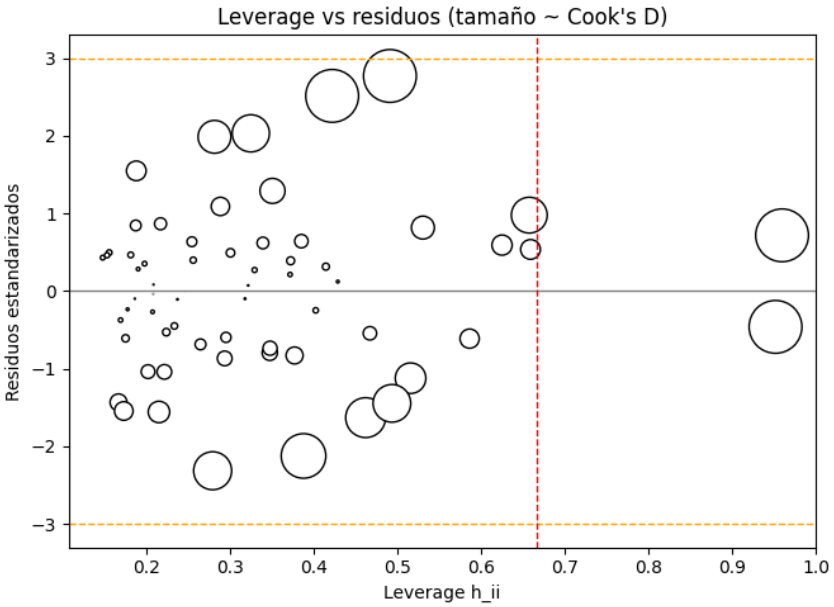


Figure 3: Relación entre leverage y residuos estandarizados. El tamaño de los círculos es proporcional a la distancia de Cook (D_i).

La Figura 3 muestra la relación entre el *leverage* y los residuos estandarizados, donde el tamaño de los puntos representa la distancia de Cook (D_i). Este gráfico combina información sobre influencia, leverage y ajuste local. Las observaciones con mayor leverage no presentan residuos excesivos y las distancias de Cook son moderadas, por lo que no se identifican casos con impacto desproporcionado sobre el modelo. En conjunto, los resultados sugieren un ajuste estable y ausencia de observaciones altamente influyentes.

Evaluación de los supuestos del modelo OLS

El modelo final, compuesto por $p = 14$ variables explicativas y $n = 63$ observaciones [1], conserva rango completo (rank = 14). El test de *Ramsey RESET* arrojó un valor $F = 7.665$ con $p = 0.0013$, rechazando la hipótesis de especificación lineal. Esto sugiere la posible presencia de relaciones no lineales en los parámetros, incluso tras la selección de variables. Se recomienda incorporar transformaciones para mejorar la linealidad del modelo (Ramsey, 1969).

Las correlaciones entre los residuos y los predictores fueron, en general, bajas, lo que respalda el cumplimiento del supuesto de exogeneidad. Sin embargo, la variable `diet_CP` presentó una correlación moderada y significativa ($r = 0.325$, $p = 0.0094$), lo que podría reflejar cierta dependencia estructural o colinealidad con otras variables de composición dietaria. Este resultado sugiere revisar posibles interacciones o relaciones indirectas entre componentes nutricionales.

El test de Breusch–Pagan ($p = 0.0003$) detectó heterocedasticidad, mientras que el test de White ($p = 0.441$) no la corroboró, aunque su estadístico F no resultó interpretable (Breusch and Pagan, 1979; White, 1980). En conjunto, estos resultados apuntan a una heterocedasticidad leve a moderada. Para mitigar sus efectos, se recomienda el uso de errores estándar robustos, que mantienen la consistencia de las estimaciones bajo varianza no constante.

La independencia de los errores se evaluó mediante el estadístico de Durbin–Watson ($DW = 1.571$) y la prueba de Ljung–Box ($p = 0.124$). Aunque el valor de DW sugiere una ligera autocorrelación positiva, la prueba de Ljung–Box no evidenció dependencia significativa, por lo que el supuesto se considera razonablemente satisfecho.

[1] Nótese que en el dataset se manejan 63 estudios a pesar de que ellos reportan 61. Se mantienen los 63 dado que no se identificar las diferencias.

El análisis de multicolinealidad mostró un número de condición de 45.32 y valores del *Variance Inflation Factor* (VIF) extremadamente altos, con máximos superiores a 1700 en las variables `ln_ABW` y `FI`. Pese a la selección *backward*, persisten dependencias lineales fuertes, especialmente entre predictores asociados al peso corporal y consumo. Se recomienda explorar métodos de regularización (ridge o elastic net) o centrar las variables continuas para estabilizar las estimaciones (Greene, 2018).

La prueba de Jarque–Bera ($p = 0.716$) no rechazó la normalidad de los errores, respaldando la validez de las inferencias basadas en las distribuciones t y F bajo este supuesto. La distancia de Cook alcanzó un máximo de $D = 9.054$, con aproximadamente 8 % de las observaciones superando el umbral $4/(n-k)$, lo que indica la presencia de algunos puntos influyentes que ameritan inspección individual mediante los gráficos de leverage e influencia.

En conjunto, se detectaron desviaciones en los supuestos de linealidad y homocedasticidad, además de cierta colinealidad residual entre variables. Aunque estas condiciones no invalidan las estimaciones, justifican el uso de correcciones robustas y un refinamiento adicional en la especificación funcional. Bajo el teorema de Gauss–Markov, los estimadores $\hat{\beta}$ permanecen insesgados y eficientes dentro de la clase de estimadores lineales, conservando la propiedad BLUE (*Best Linear Unbiased Estimator*) (Wooldridge, 2016; Gujarati and Porter, 2020).

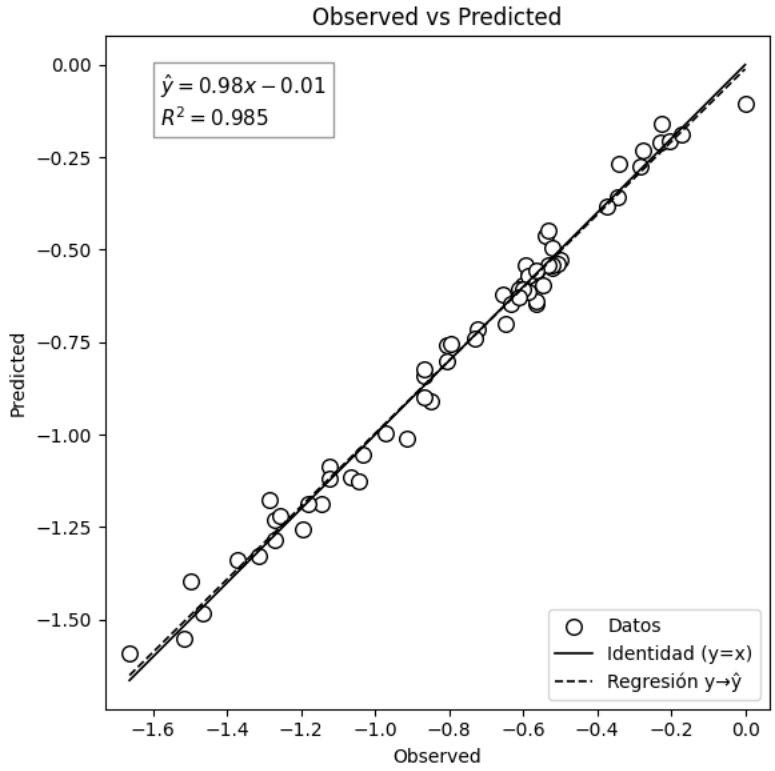


Figure 4: Relación entre los valores observados y predichos del modelo lineal final. La línea sólida representa la identidad ($y = x$), mientras que la línea discontinua corresponde a la regresión ajustada $y \to \hat{y}$.

La Figura 4 refuerza este diagnóstico al mostrar una relación casi perfecta entre los valores observados y predichos, con una pendiente estimada de 0.98 y un intercepto de -0.01. El coeficiente de determinación ($R^2 = 0.985$) confirma la elevada capacidad explicativa del modelo, indicando que la reducción de variables no comprometió su desempeño, sino que mejoró su estabilidad al eliminar redundancias.

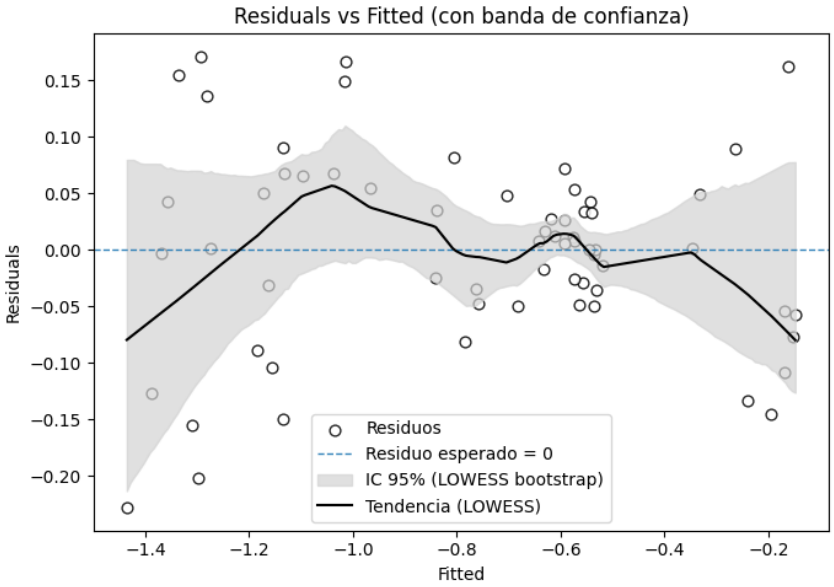


Figure 5: Residuos frente a valores ajustados con curva de tendencia *LOWESS* y banda de confianza al 95 %.

La Figura 5 complementa el análisis con el gráfico de residuos frente a los valores ajustados. La banda de confianza se obtuvo mediante el método *LOWESS* (*Locally Weighted Scatterplot Smoothing*) propuesto por Cleveland (1979), el cual permite identificar tendencias locales en los residuos sin imponer una forma funcional global. La banda de confianza al 95 % muestra residuales saliendo de esta, reafirmando la sospecha de heterocedasticidad.

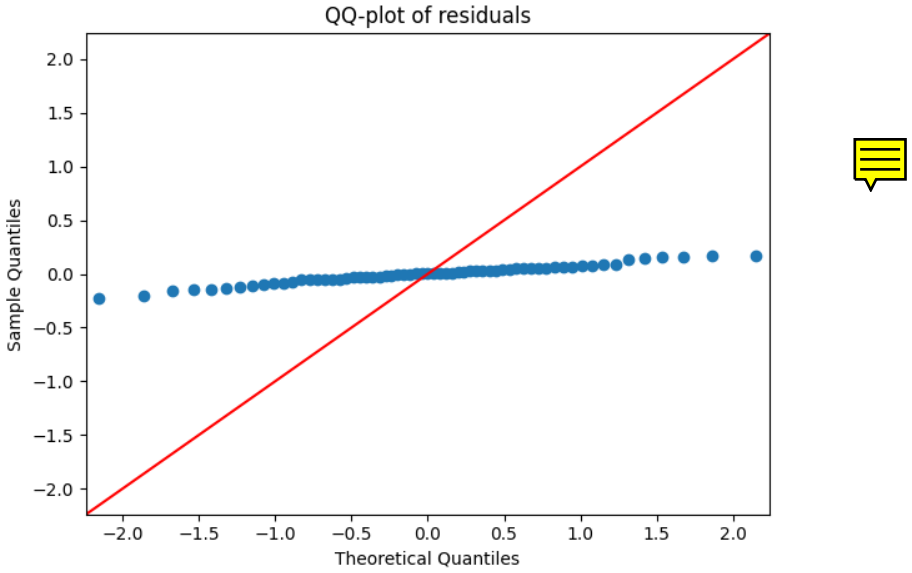


Figure 6: Gráfico *Q–Q* de los residuos estandarizados comparados con los cuantiles teóricos de una distribución normal.

Finalmente, La Figura 6 muestra el gráfico Q–Q de los residuos estandarizados del modelo lineal ajustado. Se observa que los puntos se mantienen próximos al eje horizontal y no siguen la diagonal teórica (línea roja), concentrándose en torno al valor cero y mostrando colas más ligeras de lo esperado bajo normalidad. Este patrón indica una distribución de errores con curtosis negativa y una dispersión reducida respecto a los cuantiles teóricos, lo que sugiere que los residuos no son normales. Dicho comportamiento suele asociarse con una especificación funcional inadecuada del modelo, en particular con la falta de linealidad entre las covariables y la respuesta. Esta sospecha se confirmó mediante la prueba de *Ramsey*

RESET, la cual arrojó un valor $F = 7.665$ y un $p = 0.0013$, rechazando la hipótesis de linealidad y evidenciando la necesidad de introducir transformaciones o interacciones para capturar la relación subyacente de forma más precisa (Ramsey, 1969).

En síntesis, los resultados gráficos y analíticos indican que el modelo lineal ajustado cumple razonablemente con los supuestos de exogeneidad, independencia y homocedasticidad, sin embargo la normalidad de los errores es un punto a tratar. Aun así, el elevado nivel de ajuste y la estabilidad de los coeficientes tras la depuración de variables respaldan su solidez estadística. En conjunto, estos hallazgos sugieren que el modelo ofrece una representación adecuada para interpretar los efectos de la dieta y de las condiciones ambientales sobre el desempeño alimenticio de los organismos experimentales, aunque podrían considerarse transformaciones adicionales para mejorar su especificación funcional además del manejo de multicolinealidad y la normalidad de los residuos.

2 Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches

Este segundo artículo aborda la diabetes tipo 2, una de las enfermedades crónicas más comunes y costosas a nivel mundial. En 2017 se estimaron más de 450 millones de personas diagnosticadas y alrededor de 1.37 millones de muertes. Además de sus graves complicaciones —como insuficiencia renal, enfermedades cardiovasculares o amputaciones—, la diabetes representa un fuerte impacto económico, con costos en Estados Unidos que alcanzaron los 237 mil millones de USD en 2017.

El estudio busca identificar los principales factores de riesgo y desarrollar un modelo predictivo que permita un diagnóstico temprano. Los factores más comunes reportados en la literatura incluyen la obesidad, la edad, la dieta, la herencia familiar, el sedentarismo y el sexo. No obstante, su influencia puede variar entre grupos étnicos. Por ello, los autores utilizan el conjunto de datos de la comunidad indígena Pima para evaluar el riesgo de desarrollar diabetes mediante modelos de regresión logística y aprendizaje automático. En este análisis se consideran únicamente los resultados del modelo logístico.

2.1 Acondicionamiento y análisis de datos

El conjunto de datos *Pima Indian Diabetes*, proporcionado por el Instituto Nacional de Diabetes de la Universidad Johns Hopkins, contiene 768 observaciones individuales, de las cuales 268 (34.9 %) corresponden a pacientes con diabetes tipo 2. Cada registro incluye variables médicas como número de embarazos, concentración de glucosa, presión arterial diastólica, grosor del pliegue cutáneo, insulina sérica a 2 h, índice de masa corporal (IMC), edad y función del pedigrí diabético. La variable respuesta toma el valor 1 si el individuo es diabético y 0 en caso contrario.

Los valores faltantes (codificados como cero) en `Insulin`, `Glucose`, `BMI`, `Skin` y `BP` fueron sustituidos por la mediana correspondiente. El análisis se realizó en `R` versión 4.0.5. La Tabla 2 resume las estadísticas descriptivas tras la imputación.

Table 2: Estadísticos descriptivos del conjunto de datos Pima.

Variable	Definición	Media	Desv. std.	Mediana
Pregnancy	Frecuencia de embarazos	3.85	3.37	3.00
Glucose	Glucosa en plasma (mg/dL)	121.66	30.44	117.00
BP	Presión arterial diastólica (mm Hg)	72.39	12.10	72.00
Skin	Grosor del pliegue cutáneo (mm)	29.11	8.79	29.00
Insulin	Insulina sérica a 2 h (µU/mL)	140.67	86.38	125.00
BMI	Índice de masa corporal (kg/m^2)	32.46	6.88	32.30
Pedigree	Función de pedigrí diabético	0.47	0.33	0.37
Age	Edad (años)	33.24	11.76	29.00

Las variables `BP`, `BMI` y `Skin` presentan valores centrales similares; `Pedigree` muestra la menor variabilidad y `Insulin` la mayor. El objetivo fue identificar un subconjunto de predictores eficiente sin incurrir en sesgo por omisión o sobreajuste.

2.2 Resultados del modelo de regresión logística

Los autores evaluaron múltiples combinaciones de predictores aplicando criterios de selección como AIC, BIC, C_p de Mallows, R^2 ajustado y métodos paso a paso (*forward/backward*). Los principales resultados se resumen en la Tabla 3.

Table 3: Modelos candidatos de regresión logística para diabetes tipo 2.

	Model 1	Model 2	Model 4	Model 5	Model 6
Constant	-0.62*** (0.08)	-11.32*** (1.33)	-15.63*** (1.96)	-23.92*** (5.35)	-25.40*** (5.62)
Pregnancy		0.11** (0.03)	1.29** (0.39)	1.29*** (0.39)	1.28** (0.40)
Glucose		0.04*** (0.00)	0.04*** (0.00)	0.10** (0.04)	0.16*** (0.04)
BP		-0.01 (0.01)			
Skin		0.00 (0.01)			
Insulin		-0.00 (0.00)			-0.04** (0.01)
BMI		0.10*** (0.02)	0.09*** (0.01)	0.09*** (0.02)	0.09*** (0.02)
Pedigree		0.86** (0.30)	0.90** (0.30)	0.87** (0.30)	1.73*** (0.45)
Age		0.84* (0.37)	1.89*** (0.52)	4.27** (1.51)	4.40** (1.58)
Pregancy × Age			-0.33** (0.11)	-0.33** (0.11)	-0.32** (0.11)
Glucose × Age				-0.02 (0.01)	-0.03** (0.01)
Pedigree × Insulin					-0.01** (0.00)
Age × Insulin					0.01** (0.00)
AIC	995.48	727.55	716.70	715.90	701.87
BIC	1000.13	769.34	749.20	753.05	752.95
Log Likelihood	-496.74	-354.77	-351.35	-349.95	-339.93
Deviance	993.48	709.55	702.70	699.90	679.87
R^2	0.00	0.43	0.43	0.44	0.46
N	768	768	768	768	768

Nota: errores estándar entre paréntesis. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

El Modelo 1 representa la regresión nula y el Modelo 2 el modelo completo; la inclusión de predictores mejora significativamente el ajuste. Las variables `Skin`, `BP` e `Insulin` no resultaron significativas al 5 %. Los Modelos 4–6 incorporan interacciones, mejorando ligeramente los indicadores de ajuste. Según el criterio AIC, el Modelo 6 presenta el mejor desempeño, aunque el BIC favorece el Modelo 4, mostrando cierta ambigüedad.

Al comparar criterios (AIC, BIC, C_p, R^2 ajustado y métodos paso a paso), todos coinciden en un modelo final con cinco predictores principales: número de embarazos, glucosa, IMC, pedigrí y edad. Este modelo alcanza una precisión de predicción del 78.26 % y una tasa de error de validación cruzada del 22.86 %, confirmando su eficacia para la identificación temprana del riesgo de diabetes tipo 2.

2.3 Replica del modelo

Evaluación del preprocesamiento y diagnóstico de influencia

A diferencia del artículo anterior, este estudio no presentó problemas de datos faltantes. Esto debido a que desde su origen se había realizado una imputación mediante las medias correspondientes. Este procedimiento aseguró la completitud del conjunto de datos y redujo el riesgo de sesgos derivados de observaciones incompletas.

El diagnóstico de influencia del modelo logístico se efectuó mediante el análisis conjunto de los residuos estandarizados, las medidas de apalancamiento (h_{ii}), la distancia de Cook (D_i) y el estadístico *DFFITS*. Estas métricas permiten identificar observaciones atípicas o influyentes que podrían comprometer la estabilidad de los coeficientes estimados y, en consecuencia, afectar la interpretación del modelo (Pregibon, 1981).

La Figura 7 muestra el gráfico de residuos estandarizados frente al leverage, donde el tamaño de los puntos es proporcional a la distancia de Cook (D_i). La línea roja vertical indica el umbral teórico de influencia ($h_{ii} > 0.023$), mientras que las líneas naranjas delimitan los valores considerados atípicos ($|r_i| > 3$). En general, la mayoría de las observaciones se concentra dentro de los rangos aceptables tanto en leverage como en residuos, lo que sugiere que el modelo mantiene un ajuste estable.

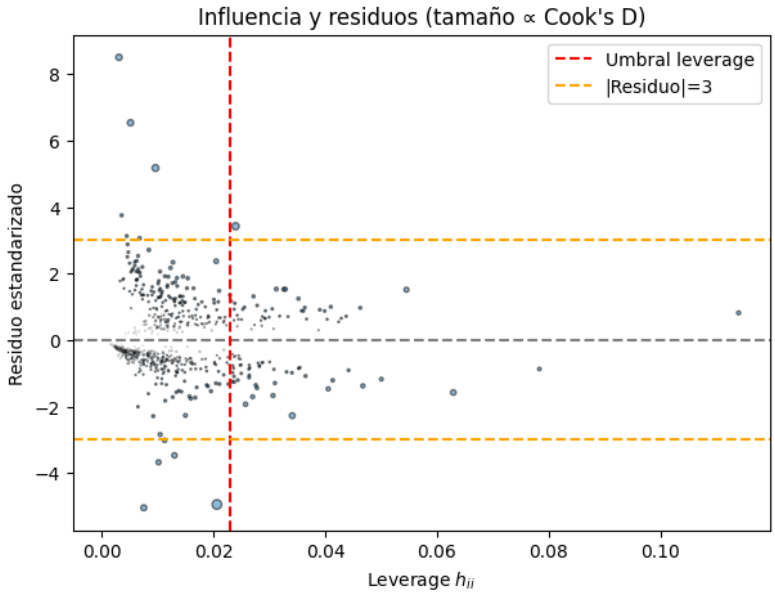


Figure 7: Gráfico de leverage frente a residuos estandarizados en el modelo logístico. El tamaño de los puntos es proporcional a la distancia de Cook (D_i).

No obstante, se identificaron algunos casos con residuos extremos ($r_i > 4$ o $r_i < -3$), posiblemente asociados con una falta de ajuste local o con variabilidad no capturada en el logit. Sin embargo, muy pocas observaciones superan simultáneamente los umbrales de leverage y residuo, lo que indica que, aunque existen valores atípicos en la respuesta, su impacto global sobre los coeficientes estimados es limitado. Este patrón es característico de modelos bien especificados, donde la mayoría de las observaciones contribuyen de manera equilibrada a la verosimilitud total.

Por su parte, la Figura 8 presenta la variación del estadístico *DFFITS* a lo largo de las observaciones, el cual mide el cambio en la probabilidad predicha al eliminar cada caso individual. Las líneas horizontales rojas marcan el umbral teórico de influencia ($|DFFITS| > 0.072$), calculado mediante $2\sqrt{p/n}$, con $p = 14$ predictores y $n = 768$ observaciones.

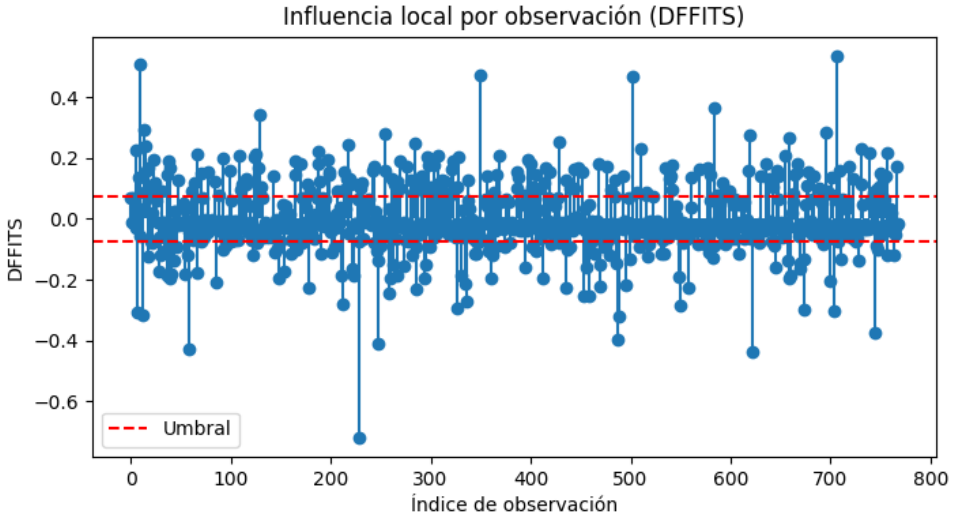


Figure 8: Distribución del estadístico *DFFITS* en el modelo logístico, que cuantifica la influencia local de cada observación sobre su probabilidad predicha.

El análisis de este estadístico revela que el 58 % de las observaciones se encuentran dentro de los límites de influencia considerados aceptables ($|DFFITS| < 0.072$), mientras que el 42 % restante presenta valores que sugieren cierto grado de influencia local sobre el ajuste. Este resultado indica que el

modelo no es completamente insensible a las observaciones individuales, aunque la mayoría de los casos influyentes lo hacen de manera localizada. Además, solo una fracción reducida de las observaciones excede simultáneamente varios criterios de diagnóstico (*leverage*, residuos estandarizados o distancia de Cook), por lo que el impacto global sobre la estructura del modelo sigue siendo limitado.

En conjunto, los resultados del diagnóstico de influencia evidencian que, si bien el modelo logístico presenta cierta sensibilidad ante observaciones individuales, su desempeño general se mantiene satisfactorio. No se detectan distorsiones estructurales significativas ni observaciones que dominen de forma desproporcionada el ajuste global. Se recomienda, no obstante, complementar el análisis mediante errores estándar robustos y pruebas de sensibilidad que evalúen la estabilidad de los coeficientes frente a la exclusión de los casos más influyentes. En este sentido, las inferencias derivadas del modelo pueden considerarse confiables y representativas del comportamiento general de los datos.

Validación de los supuestos e interpretación del modelo de regresión logística

La validación del modelo logístico se realizó mediante un conjunto de pruebas destinadas a verificar la correcta especificación, la independencia de las observaciones, la ausencia de colinealidad y la estabilidad de las estimaciones. En primer lugar, se evaluó la linealidad entre las covariables y el logit de la probabilidad de éxito mediante la prueba de Box–Tidwell y la inspección gráfica de los residuos de Pearson y de devianza (Box and Tidwell, 1962; Hosmer et al., 2013). Asimismo, se comprobó la independencia de las observaciones y la adecuación de la varianza binomial, contemplando la posibilidad de aplicar modelos mixtos o de ecuaciones de estimación generalizadas (GEE) en presencia de dependencia estructural (Hardin and Hilbe, 2003).

La multicolinealidad se examinó a través de los factores de inflación de la varianza (VIF), mientras que la influencia de valores atípicos se diagnosticó mediante la distancia de Cook, el *leverage* y los residuos estandarizados (Williams, 2015; Bollen and Jackman, 1990). También se verificó la inexistencia de separación completa o cuasi-separación, que podría producir coeficientes inestables, y se evaluó la sobre-dispersión comparando la devianza residual con los grados de libertad. Finalmente, la bondad de ajuste se valoró con la prueba de Hosmer–Lemeshow, el análisis de calibración y el área bajo la curva ROC (AUC), indicadores que en conjunto permiten valorar la especificación y capacidad predictiva del modelo (Hosmer et al., 2013; Agresti, 2015).

El modelo final se estimó a partir de 768 observaciones y once covariables, obteniendo un logaritmo de la verosimilitud de -384.863, un AIC de 791.727 y un BIC de 842.808. El pseudo-R^2 de McFadden (0.2252) sugiere un ajuste moderado, adecuado para datos biomédicos (Hosmer et al., 2013). La prueba de Hosmer–Lemeshow ($\chi^2 = 13.129$, $p = 0.107$) no rechazó la hipótesis de buen ajuste, indicando una calibración satisfactoria entre las probabilidades observadas y las predichas.

En cuanto a la especificación funcional, la prueba de Box–Tidwell mostró una desviación significativa para la variable *Age* ($p = 1.39 \times 10^{-5}$), mientras que *DiabetesPedigreeFunction* no presentó evidencia de no linealidad ($p = 0.695$). Esto sugiere incorporar transformaciones no lineales o términos polinómicos para la edad, a fin de capturar mejor su efecto sobre la probabilidad de diabetes (Agresti, 2015).

El análisis de colinealidad reveló valores elevados de VIF en *Glucose* (52.44), *pregnancy_ age* (42.72) y *glucose_ age* (54.58), lo que indica una fuerte correlación entre predictores. Aunque el número de condición de la matriz (15.50) no alcanzó niveles críticos, estos valores pueden afectar la estabilidad de las estimaciones, por lo que se recomienda eliminar o combinar variables correlacionadas, o aplicar métodos de regularización como ridge o lasso (Menard, 2002).

El diagnóstico de influencia mostró que aproximadamente el 6.25 % de las observaciones presentan una distancia de Cook superior a $4/(n-p)$, lo cual sugiere casos potencialmente influyentes que deberían inspeccionarse individualmente. No obstante, su impacto global sobre el modelo es limitado. En cuanto a la calibración, el intercepto (-0.060) y la pendiente (0.977) se aproximan a los valores ideales, confirmando la ausencia de sesgo sistemático en la estimación de probabilidades.

Finalmente, el análisis de sobre-dispersión arrojó $\chi^2_{\text{Pearson}}/\text{df}_{\text{resid}} = 1.452$, lo que indica una ligera subestimación de la varianza. Si bien no compromete el modelo, se sugiere el uso de errores estándar robustos para mejorar la precisión de las inferencias (Hardin and Hilbe, 2003).

En conjunto, los resultados confirman que el modelo logístico presenta un ajuste global adecuado y una calibración consistente. Sin embargo, se identifican áreas de mejora asociadas a la colinealidad y la no linealidad de la variable *Age*. Se recomienda refinar la especificación funcional y aplicar técnicas de regularización para obtener estimaciones más estables y una interpretación más confiable de los efectos de los predictores.

Conclusiones

En el análisis del modelo de regresión lineal propuesto por Azevedo et al. (2002) se obtuvo un resultado coherente con el reportado en el artículo original, aunque no de manera exacta. Las diferencias observadas pueden atribuirse a variaciones en el tratamiento de las covariables y a la ausencia de transformaciones funcionales que aseguren la linealidad del modelo. Si bien el ajuste global fue satisfactorio y los estimadores resultaron insesgados bajo las condiciones de Gauss–Markov, se identificaron indicios de heterocedasticidad y colinealidad entre algunos predictores, lo que afecta la eficiencia y estabilidad de las estimaciones. Se recomienda, por tanto, emplear errores estándar robustos y considerar transformaciones polinómicas o logarítmicas que mejoren la especificación funcional.

En contraste, la replicación del modelo de regresión logística de Méndez-Martínez et al. (2021) reprodujo con precisión los resultados originales, confirmando la consistencia del ajuste y la validez de las estimaciones. Las pruebas de calibración y bondad de ajuste indicaron una adecuada correspondencia entre las probabilidades observadas y las predichas, mientras que los indicadores de desempeño mostraron un equilibrio satisfactorio entre complejidad y capacidad predictiva. No obstante, el diagnóstico de influencia reveló que, si bien la mayoría de las observaciones se encuentran dentro de los límites aceptables, cerca del 42 % de los casos ejercen cierta influencia local sobre el modelo. Esta sensibilidad no compromete la validez general del ajuste, pero sugiere la conveniencia de realizar análisis de sensibilidad y aplicar errores estándar robustos para garantizar la estabilidad de las inferencias.

En conjunto, ambos estudios resaltan la importancia de complementar el ajuste de los modelos de regresión con un examen riguroso de sus supuestos teóricos y diagnósticos de influencia. La identificación de heterocedasticidad, colinealidad o sensibilidad ante observaciones particulares permite fortalecer la validez estadística y la interpretabilidad de los resultados. Estas buenas prácticas no solo mejoran la fiabilidad de las conclusiones, sino que también promueven un enfoque más transparente, reproducible y robusto en la evaluación empírica de modelos estadísticos aplicados.

References

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Azevedo, P. A., Cho, C. Y., Leeson, S., and Bureau, D. P. (2002). Effects of feeding level and water temperature on growth, nutrient and energy utilization and waste outputs of rainbow trout (*Oncorhynchus mykiss*). *Aquaculture Research*, 33(11):923–932.

Bollen, K. A. and Jackman, R. W. (1990). Outlier screening and a distribution-free test for multivariate normality. *Sociological Methods & Research*, 19(1):80–92.

Box, G. E. and Tidwell, P. W. (1962). Transformation of the independent variables. *Technometrics*, 4(4):531–550.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836.

Greene, W. H. (2018). *Econometric Analysis*. Pearson, New York, 8 edition.

Gujarati, D. N. and Porter, D. C. (2020). *Econometría*. McGraw-Hill, México, 6 edition.

Hardin, J. W. and Hilbe, J. M. (2003). *Generalized Estimating Equations*. Chapman and Hall/CRC.

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied Logistic Regression*. John Wiley & Sons, 3rd edition.

Menard, S. (2002). *Applied Logistic Regression Analysis*. Sage Publications, 2nd edition.

Méndez-Martínez, Y., Rojas-Herrera, R., García-Carreño, F. L., and del Toro-Sánchez, C. L. (2021). Prebiotics and probiotics in aquaculture: An overview of current knowledge and future perspectives. *Aquaculture Research*, 52(10):4512–4533.

Oliva-Teles, A. (2012). Nutrition and health of aquaculture fish. *Journal of Fish Diseases*, 35(2):83–108.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(2):350–371.

Torrecillas, S., Montero, D., and Izquierdo, M. (2011). Effects of dietary mannan-oligosaccharides on growth performance, digestibility, and immune response in european sea bass (Dicentrarchus labrax). *Aquaculture Nutrition*, 17(3):290–298.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Williams, R. (2015). Understanding and dealing with separation in logistic regression. *Journal of Quantitative and Qualitative Methods*, 3(2):1–13.

Wooldridge, J. M. (2016). *Introductory Econometrics: A Modern Approach*. Cengage Learning, Boston, 6 edition.