Proyecto 3: Regresiones lineal y logística
Introducción a la Ciencia de Datos

Integrantes:	Bueno Rivera, Oswaldo; Rodríguez Villagrán, Juan Pablo
Programa Educativo:	Maestría en Probabilidad y Estadística
Institución:	Centro de Investigación en Matemáticas
Profesor:	Dr. Marco Antonio Aquino López

Resumen

En este trabajo se presenta una revisión crítica a dos artículos: uno de biomarcadores sanguíneos en ratas y su reacción ante ciertos estímulos, y otro sobre equidad en labores domésticas y su relación con la formación de familias en Canadá. La crítica se hace en los modelos de regresión lineal y logística utilizados para argumentar. En el caso del artículo de biología se concluye que aún tras la propuesta alternativa, no se cumplen los supuestos de la regresión lineal y debería atenderse de otra manera. Adicionalmente, se propone un procedimiento de selección de variables con LASSO. En el caso del artículo de sociología, se exploran análisis alternativos al presentado ya que a partir del conjunto de datos no se pueden replicar directamente los resultados. Esto demuestra problemas de comunicación al no dar claridad en la codificación de variables.

1. Introducción

Un problema central en estadística consiste en modelar la relación entre una variable de interés Y y un conjunto de variables explicativas $X = (X_1, \ldots, X_d)$, usualmente más fáciles de observar o controlar experimentalmente. En este contexto, se cuenta con una muestra $(X_1, Y_1), \ldots, (X_n, Y_n)$ y se busca describir a Y como una función de X. Este problema, conocido como el *problema de regresión*, se expresa formalmente como

$$Y = f(X) + \epsilon,$$

donde ϵ representa un error no explicable por el modelo o errores de medición propios del experimento.

Entre las múltiples maneras de acercarse a este problema los modelos de *regresión lineal* y *regresión logística* destacan por su simplicidad interpretativa, su capacidad explicativa y su papel en análisis de datos, en general. Estas técnicas son ampliamente utilizadas en diferentes campos tanto por su carácter predictivo como por la facilidad con que permiten evaluar la significancia estadística de los predictores.

El objetivo principal de este trabajo es examinar críticamente el uso de los modelos de regresión lineal y la regresión logística en dos estudios provenientes de áreas distintas, analizando la metodología empleada y proponiendo posibles extensiones o mejoras. En particular, se siguen dos etapas:

1. *Replicación del método propuesto en la publicación.* Se revisará que el modelo especificado en cada publicación cumpla las hipótesis del método utilizado y que los resultados reportados sean estadísticamente consistentes con los datos.

2. *Propuesta de alternativas.* Se discutirán posibles modificaciones al enfoque original o métodos alternativos que puedan mejorar la interpretación, ajuste o capacidad predictiva de los modelos considerados.

2. Regresión lineal

La regresión lineal es uno de los modelos fundamentales del análisis estadístico, pues permite describir la relación promedio entre una variable de respuesta Y y un conjunto de predictores $X_1, \ldots, X_d$ mediante una combinación lineal de estos últimos. Según Christensen [Chr20], su poder radica tanto en la interpretación geométrica como una proyección ortogonal de Y sobre el espacio generado por los predictores, como en su simplicidad algebraica, que hace posible una inferencia transparente sobre la contribución de cada variable. El modelo clásico supone que los errores ϵ son independientes, con esperanza cero y varianza constante, y que la esperanza condicional de Y se expresa como $\mathbb{E}[Y|X] = X\beta$.

2.1. Descripción general

Para esto, utilizaremos el artículo de Prussia, A. y Demchuk, E. [PD25b] que estudia la relación entre la estructura de alquilos de flúor (PFAS) y cambios en biomarcadores sanguíneos en ratas. A continuación se presenta un resumen del contexto bajo el cual se desarrolla ese trabajo.

Las sustancias perfluoroalquiladas y polifluoroalquiladas (PFAS) son contaminantes orgánicos persistentes que, en muchos casos, presentan bajas tasas de eliminación en organismos vivos. Los PFAS de cadena larga (con seis o más carbonos perfluorados) tienden a tener una depuración particularmente lenta, por lo que suelen considerarse más perjudiciales para la salud humana que los PFAS de cadena corta.

Mediante la aplicación de la prueba de Dunnett, se analizaron 15 marcadores clínicos reportados por el Programa Nacional de Toxicología (NTP) tras una exposición de 28 días a siete tipos de PFAS, que incluían compuestos con estructuras perfluoroalquiladas tanto de cadena corta como de cadena larga.

El artículo [PD25b] se propuso identificar las dosis de PFAS que inducen cambios en marcadores de los sistemas hepático, renal, cardiovascular y metabólico, para posteriormente, emplear estos valores en modelos de regresión lineal múltiple con el fin de evaluar su relación estadística con variables estructurales de estos compuestos. Los resultados mostraron una dependencia log-lineal entre las alteraciones en los niveles de los marcadores y la estructura fluorada de los PFAS, cuantificada específicamente por el número de enlaces carbono-flúor (C-F). El análisis reveló que cada enlace C-F adicional se asociaba con una modificación de 0.45 ± 0.01 mmol/kg-día en el nivel de efecto del marcador.

El estudio del NTP utilizó ratas Sprague Dawley de ambos sexos, de entre 10 y 11 semanas de edad al inicio del experimento, con 10 machos y 10 hembras asignados a cada grupo de dosis y al grupo control. Prussia y Demchunk muestran la comparación entre el grupo de control y los grupos de dosis en [PD25b].

2.2. Presentación de los datos

Los datos utilizados por el artículo, recabados por los autores, se encuentra en [PD25a]. El conjunto de datos tiene distintas tablas, pero para puntualizar en este trabajo, sólo se considera `Table S4`. En la tabla 1 se presenta de manera sintética el tipo de información con la que se cuenta.

Atributo	**Descripción**	**Tipo**	**Posibles valores**
PFAS	Nombre del compuesto perfluoroalquilado	Categórica	PFHxA, PFHpA, PFOA, PFNA, PFDA, PFBS, PFHxS-K, PFOS
n(F)	Número de átomos de flúor o enlaces C–F en la molécula	Numérica	Enteros positivos
PFAS type	Clasificación química principal del compuesto	Categórica	Carboxylic, Sulfonic
Sex	Sexo del animal experimental	Categórica	Male, Female
POD type	Tipo de punto de partida toxicológico	Categórica	LOEL (Lowest Observed Effect Level), NOEL (No Observed Effect Level)
Marker	Biomarcador clínico medido	Categórica	ALT, AST, CHOL, CREA, GLU, TP, ALB, ALP, TBILI, TG, etc.
POD	Logaritmo (base 10) de la dosis correspondiente al punto de partida (mmol/kg·día)	Numérica	Valores reales negativos; más bajo implica mayor toxicidad

Cuadro 1: Detalle de variables de la *Table S4*: datos utilizados para la regresión lineal entre estructura de PFAS y nivel de dosis (LOEL/NOEL).

Para estimar la dosis de un PFAA en ratas que provoca un cambio en el nivel de un marcador clínico respecto al grupo control (es decir, el umbral dosis-efecto para cada par PFAA/marcador), Prussia y Demchuk calcularon los límites superior e inferior de la dosis umbral utilizando un nivel de significancia $\alpha = 0.05$.

Para cada marcador, se determinaron el nivel más bajo con efecto observado (LOEL) y el nivel sin efecto observado (NOEL) mediante la prueba de Dunnett, considerando todas las combinaciones entre los 15 marcadores clínicos y los 7 PFAAs, y estratificando por sexo.

Una observación importante a la metodología seguida por los autores, es que un requisito para aplicar la prueba de Dunnett es que los datos sigan una distribución lognormal. En el artículo mencionan que algunas variables fallaron la prueba de lognormalidad, pero aún así las consideraron como tal ya que de acuerdo a razones químicas, dichas variables deben seguir una distribución lognormal. Esto se detalla en [DA76], el cual muestra que la lognormalidad de las concentraciones está relacionada con el potencial químico y se deriva de la energética de las reacciones químicas que siguen la estadística de Maxwell-Boltzmann.

El objetivo final fue plantear un modelo de regresión para el parámetro POD, de modo que

$$\text{POD}_{ijk} = \beta_0 + \beta_1 n(F)_i + \beta_2 \text{PFAS type}_j + \beta_3 \text{Sex}_k + \beta_4 \text{POD type} + \epsilon_{ijk}.$$

2.2.1. Replicación de método

Prussia y Demchuk analizan con regresión lineal, tratando indistintamente NOEL y LOEL, cómo las propiedades del PFAS administrado afectan a cada marcador de los sistemas hepático, renal, cardiovascular y metabólico. En el artículo se presentan las gráficas de valores ajustados contra los reales; sin embargo, *no se presenta un análisis de los residuales para validar los supuestos de las regresiones.* Un detalle a resaltar es que los autores mencionan que omiten los datos faltantes, puesto que para cada tipo de PFAS siempre tienen ya sea NOEL o LOEL, y que no llevan a cabo ningún método de imputación al no ser necesario. Por nuestra parte, para tener consistencia con el trabajo original, decidimos no imputar datos y simplemente omitir los faltantes.

Nuestros resultados presentados en las gráficas de valores ajustados contra los reales concuerdan con aquellas obtenidas por los autores. Las gráficas correspondientes se encuentran en el apéndice A. La selección de variables llevada a cabo por nosotros concordó en gran medida con aquella obtenida por los autores. Al igual que ellos, es importante notar que el número de enlaces C-F se mantuvo como una variable significativa para todos los marcadores clínicos, mostrando que es muy influyente en el efecto de los PFAS. A continuación, en la tabla 2, presentamos los valores de R^2 ajustada, el $RMSE$, las variables seleccionadas y el p-valor de la prueba de Shapiro-Wilk aplicada a los residuales de la regresión lineal de cada marcador clínico.

Marcador	$\mathbf{R^2}$	**RMSE**	**Variables Seleccionadas**	p**-valor S-W**
AA	0.8465	0.5020	n(F), PFAS type_sulfonic, Sex_male	0.4328
Alb	0.8506	0.4121	n(F)	0.6326
AP	0.9264	0.3421	n(F), Sex_male, POD type_NOEL	0.6582
AspA	0.8809	0.4561	n(F), PFAS type_sulfonic, Sex_male	0.9262
BilesSalts	0.8531	0.4776	n(F), PFAS type_sulfonic	0.2403
Cholesterol	0.6440	0.6924	n(F), PFAS type_sulfonic, Sex_male	0.5498
CK	0.8311	0.4709	n(F)	0.6981
Creatinine	0.7847	0.5753	n(F), Sex_male	0.2955
DirectBilirubin	0.8613	0.4476	n(F)	0.1014
Glucose	0.8525	0.4406	n(F), Sex_male	0.7921
SorbitolDehydrogenase	0.8410	0.4700	n(F)	0.8631
TotalBilirubin	0.8268	0.4797	n(F)	0.2696
Protein	0.8272	0.5094	n(F), Sex_male	0.9368
Triglycerides	0.7269	0.6469	n(F), PFAS type_sulfonic, Sex_male	0.2306
UreaNitrogen	0.7906	0.5876	n(F)	0.6794

Cuadro 2: Resumen de modelos de regresión por marcador clínico.

Dado que los coeficientes para los enlaces C-F resultaron similares, los autores realizaron una regresión categórica conjunta utilizando una única variable categórica de marcador. Dicho esto, efectuamos una regresión lineal considerando las mismas variables que antes (que describen el tipo de PFAS), pero además consideramos todos los distintos marcadores clínicos como una variable categórica. En la figura 1 se presentan las gráficas correspondientes a este ajuste.

Para este ajuste de regresión lineal obtuvimos la raíz del error cuadrático medio igual a $RMSE = 0.38$, mientras que la R^2 ajustada tuvo un valor de $R^2 = 0.90$. Además, consideramos la prueba de normalidad de Shapiro-Wilk obteniendo un p-valor igual a $p = 0.05$, bajo el cual se rechaza la hipótesis de normalidad de los residuales. Lo anterior, junto con el mal comportamiento de las gráficas para evaluar residuales pudiera indicar que considerar los marcadores clínicos como variable categórica no sea adecuado.

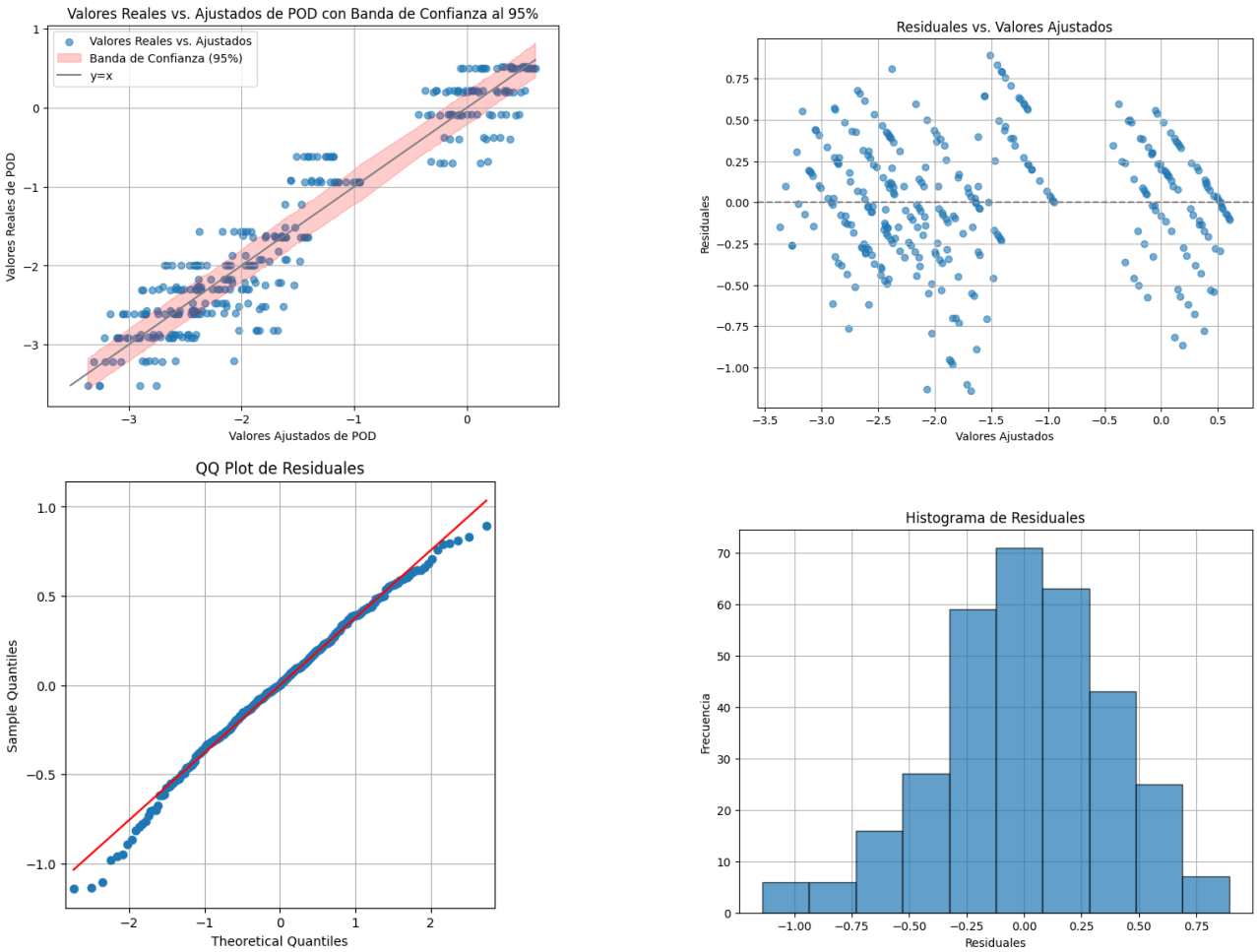


Figura 1: Análisis de regresión considerando los marcadores clínicos

2.2.2. Método propuesto

En el trabajo original, los autores no mencionan haber hecho un proceso de selección de variables, sino que sólo consideraron los 15 marcadores clínicos como valores de una variable categórica. Al tener un número tan alto de variables es razonable pensar en una reducción de éstas. Para la selección, proponemos llevar a cabo un ajuste de regresión LASSO, el cual elimina aquellas variables que no muestren ser significativas para la regresión [Tib96]. En la figura 2 y en la tabla 3 se presentan los resultados de la regresión LASSO.

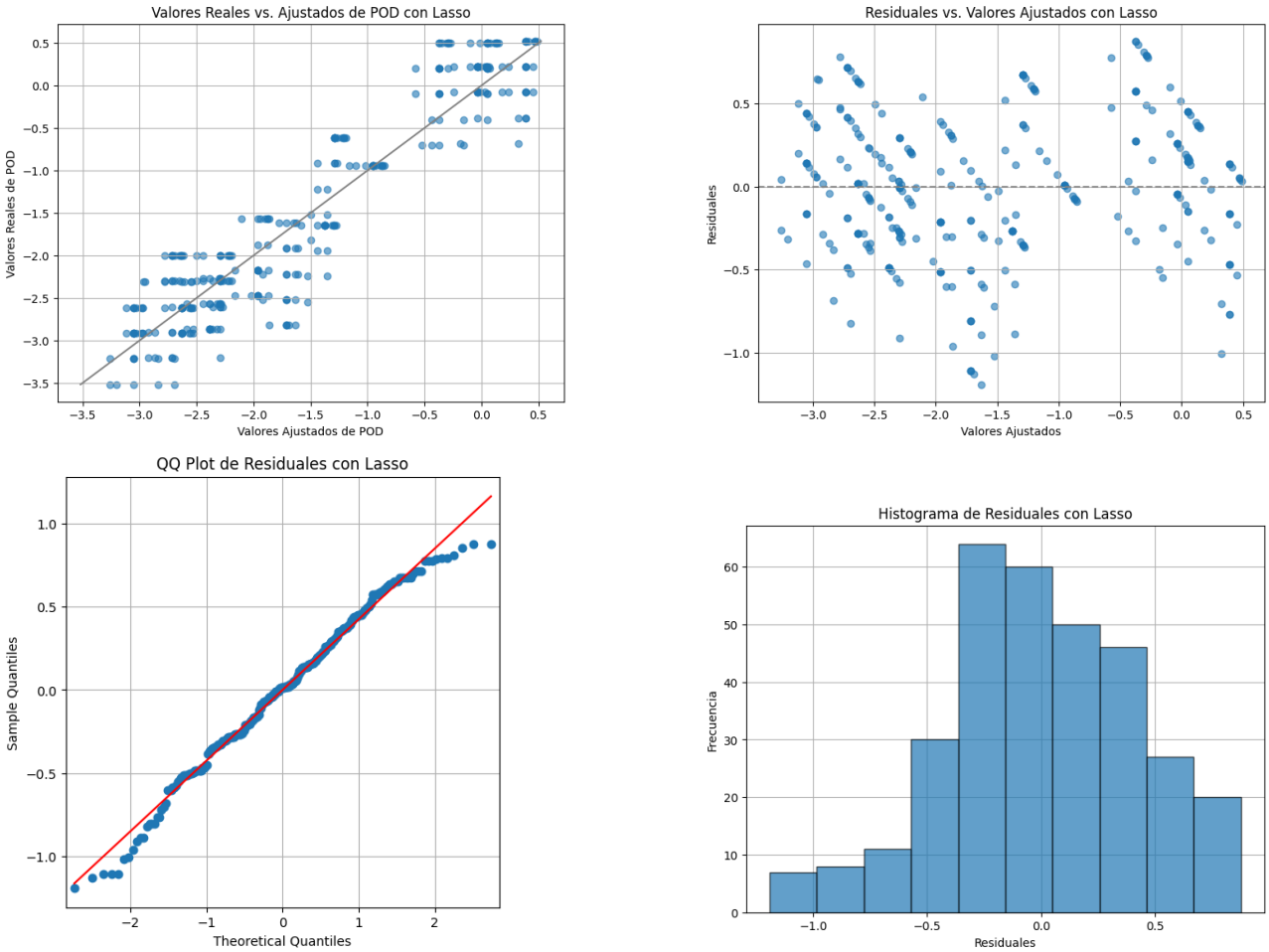


Figura 2: Análisis de regresión LASSO considerando los marcadores clínicos

Variable	**Coeficiente**
n(F)	-1.1483
PFAS type_sulfonic	-0.1633
Sex_male	-0.2115
Marker_AP	-0.0553
Marker_Alb	-0.0159
Marker_CK	0.0230
Marker_Cholesterol	-0.0356
Marker_Creatinine	0.0197
Marker_DirectBilirubin	0.0158
Marker_Glucose	0.0201
Marker_SorbitolDehydrogenase	0.0203
Marker_Triglycerides	0.0051
α óptimo	0.0192
RMSE	0.4254
R^2	0.8786

Cuadro 3: Resultados de la regresión LASSO.

Podemos resaltar que el histograma de residuales está cargado hacia la derecha, lo cual es indicio de que los residuales no siguen una distribución normal. Esta afirmación es sustentada por la QQ plot y la gráfica de residuales contra valores ajustados. La prueba de Shapiro-Wilk da un p-valor de $p = 0.01$ de manera que se rechaza la normalidad de los residuales.

Los análisis de residuales del ajuste tanto de los autores como el propuesto, indican que las regresiones no satisfacen los supuestos necesarios. Dado lo anterior, el estudiar la variable POD considerando los marcadores como una variable categórica podría ser inadecuado. También podría ser adecuado estudiar a los datos con valores NOEL y LOEL por separado, así se da un énfasis único a cada tipo de umbral; sin embargo, esto no aseguraría que un ajuste de regresión lineal sea adecuado y posiblemente sea necesario otro tipo de análisis.

3. Regresión logística

La regresión logística es un modelo estadístico diseñado para analizar la relación entre un conjunto de variables explicativas $X_1, \ldots, X_d$ y una variable de **respuesta categórica binaria** $Y \in \{0, 1\}$. A diferencia de la regresión lineal, en la que la esperanza condicional $\mathbb{E}[Y|X]$ se modela directamente, en la regresión logística se modela la probabilidad de éxito mediante la función de enlace logit,

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = X\beta,$$

lo que garantiza que las probabilidades estimadas se mantengan dentro del intervalo $(0, 1)$. El que la regresión se haga sobre *una función de la media* hace que la regresión logística forme parte de la familia de *modelos lineales generalizados*.
Este modelo permite interpretar los coeficientes β_j en términos de razones de momios (odds ratio), proporcionando una medida clara del cambio relativo en la probabilidad de ocurrencia de $Y = 1$ ante variaciones en X_j.
Más detalles de este modelo se pueden revisar en el texto de Hosmer [HLS13].

3.1. Descripción general

Para ilustrar la aplicación de este modelo, utilizamos el artículo de Erfani, A. y Pilon, L. [EP25] el cual estudia la relación entre la equidad en las labores domésticas y las intenciones de formar familias en $N = 1589$ mujeres de entre 18 y 39 años que se encontraban en relaciones heterosexuales. El conjunto de datos utilizado proviene del *General Social Survey – Cycle 31: Family (2017)* [Sta17], una encuesta telefónica transversal representativa de la población canadiense, llevada a cabo por *Statistics Canada*.

El objetivo principal del estudio fue, mediante una regresión logística, contrastar hipótesis sociológicas sobre la relación entre la distribución equitativa de las tareas domésticas y las intenciones de tener hijos. En particular, las autoras analizaron si la igualdad en el reparto de responsabilidades dentro del hogar (como cocinar, limpiar, lavar o hacer las compras) influye positivamente en la disposición de las mujeres a ampliar su familia, y si dicha relación depende del nivel educativo o de la situación laboral de las entrevistadas.

El artículo formula tres hipótesis principales

H1. Las mujeres que comparten tareas del hogar de manera equitativa tienden más a querer tener hijos

H2. Las mujeres que tienen un empleo y comparten las tareas del hogar de manera equitativa tienden más a querer tener hijos

H3. Las mujeres con estudios universitarios que comparten las tareas del hogar de manera equitativa tienden más a querer tener hijos.

Los resultados obtenidos mostraron que la probabilidad de expresar intención de tener hijos fue significativamente mayor entre las mujeres que compartían las tareas domésticas de forma equitativa, especialmente entre aquellas con empleo a tiempo completo y con estudios universitarios. En contraste, entre mujeres sin empleo o con menor nivel educativo, la asociación fue débil o incluso nula, lo que sugiere la presencia de un efecto de moderación por nivel educativo y condición laboral.

3.2. Presentación de los datos

Los datos utilizados por el estudio, recabados por *Statistics Canada*, se encuentran disponibles públicamente en [Sta17], que se pueden revisar directamente en este enlace. El conjunto contiene variables demográficas, familiares y laborales que permiten construir un modelo de probabilidad de intención de fertilidad. En la tabla 4 se presentan las variables empleadas en el análisis y sus posibles valores.

Atributo	Descripción	Tipo	Posibles valores
FERINT	Intención de tener (otro) hijo en los próximos 3 años	Categórica binaria	1 = Sí (Definitivamente/Probablemente), 0 = No (Definitivamente/Probablemente)
HHDW001–HHDW004	Quién realiza principalmente las tareas domésticas rutinarias (preparar comidas, lavar platos, lavar ropa, limpiar)	Categórica ordinal	1 = Principalmente la entrevistada; 2 = Su pareja; 3 = Compartido; 4 = Ninguno
HHDW005–HHDW007	Quién realiza principalmente las tareas domésticas intermitentes (compras, organización social, finanzas y pagos)	Categórica ordinal	1 = Principalmente la entrevistada; 2 = Su pareja; 3 = Compartido; 4 = Ninguno
EDUDR04	Máximo nivel educativo alcanzado	Categórica	0 = Secundaria o técnica; 1 = Licenciatura o superior
LFSSTAT / EMPLST	Situación laboral en la semana anterior / Empleo actual	Categórica binaria	1 = Empleada; 0 = No empleada
AGEC	Edad de la entrevistada	Numérica continua	18–39 años (muestra restringida)
CHILDNUM	Número de hijos vivos (incluye embarazo actual)	Numérica discreta	0, 1, 2, 3+
INCGRP	Ingreso familiar antes de impuestos	Categórica ordinal	$< 50,000$, 50–75k, 75–100k, 100–125k, 125k+
REGION	Región de residencia	Categórica	Atlántico, Quebec, Ontario, Praderas, Columbia Británica
BRTHC	Lugar de nacimiento de la entrevistada	Categórica binaria	1 = Canadá, 0 = Fuera de Canadá
MARSTAT / SEXPR	Estado civil y sexo de la pareja conviviente	Categórica	Filtrado a mujeres con pareja masculina (casadas o en unión libre)

Cuadro 4: Detalle de variables del *General Social Survey – Cycle 31: Family (2017)* utilizadas en el estudio de Erfani y Pilon (2025).

El modelo principal especificado por las autoras se centra en la probabilidad de que una mujer declare intención de tener hijos dentro de los próximos tres años, en función del grado de equidad en la división de las tareas domésticas, nivel educativo y situación laboral. Formalmente, el modelo puede escribirse como:

$$\begin{aligned}\text{logit}\left(\mathbb{P}[\text{FERINT} = 1]\right) = \beta_0 + \beta_1(\text{DomesticEquality}) + \beta_2(\text{Employment}) + \beta_3(\text{Education}) \\ + \beta_4(\text{Controles}) + \beta_5(\text{Interacciones}),\end{aligned}$$

donde los coeficientes se interpretan en términos de las razones de momios (odds ratios), de modo que $\exp(\beta_j)$ indica el cambio multiplicativo en las probabilidades de tener intención de fertilidad ante un incremento unitario en la variable X_j.

La manera en la que validan las hipótesis H1, H2 y H3 es proponiendo modelos con distintas características

M1 El primer modelo incluye sólo los índices de equidad de género en división de trabajo doméstico

M2 El segudo modelo ajusta los efectos de estos índices por variables de control

M3 El tercer modelo añade interacciones para examinar las hipótesis 2 y 3, además de que agrega variables independientes y de control.

3.2.1. Replicación de método

La variable de respuesta representa la intención de formar una familia `FERINT` (*fertility intention*). Los datos de esta pregunta se obtuvieron con la pregunta "¿piensa tener un hijo en los siguientes tres años?", con posibles respuestas: "Definitivamente sí", "Probablemente sí", "Probablemente no", "Definitivamente no" y "No estoy segura". Una decisión metodológica fue formar una única categoría llamada "Sí" combinando "Definitivamente sí" y "Probablemente sí", y análogamente formar una única categoría llamada "No" combinando "Definitivamente no" y "Probablemente no".

La variable independiente más importante *Equidad de género en la división de labores domésticas* se midió preguntando directamente de labores domésticas como: preparar la comida, limpieza, lavado de trastes, lavado de ropa, hacer la despensa, jardinería, reparaciones, organización de vida social, finanzas domésticas y pago de impuestos. Otra decisión metodológica fue descartar las columnas de jardinería y de reparaciones dado que no aplican para parejas que vivían en condominios o lugares donde fuera imposible.

La replicación del método no es inmediata ya que a partir de los datos, no es claro cómo hacer la agrupación para definir la variable de respuesta ni las variables relacionadas con la equidad de división de labores. Posiblemente sea un caso de censura de variables ya que la *fertility intention* podría considerarse un dato sensible.

Dado que no se cuenta directamente con el dato de *fertility intention*, se hace inferencia sobre la variable *children*. Siguiendo el método propuesto en el artículo, con la agrupación de variables indicada, la regresión logística da una matriz de confusión como la presentada a continuación en la figura 3. Dos de los principales supuestos que debe cumplir la muestra para que se pueda utilizar la regresión logística son: no multicolinealidad y linealidad. En la tabla 5 se presentan los factores de inflación de varianza para el tercer modelo, de la cual

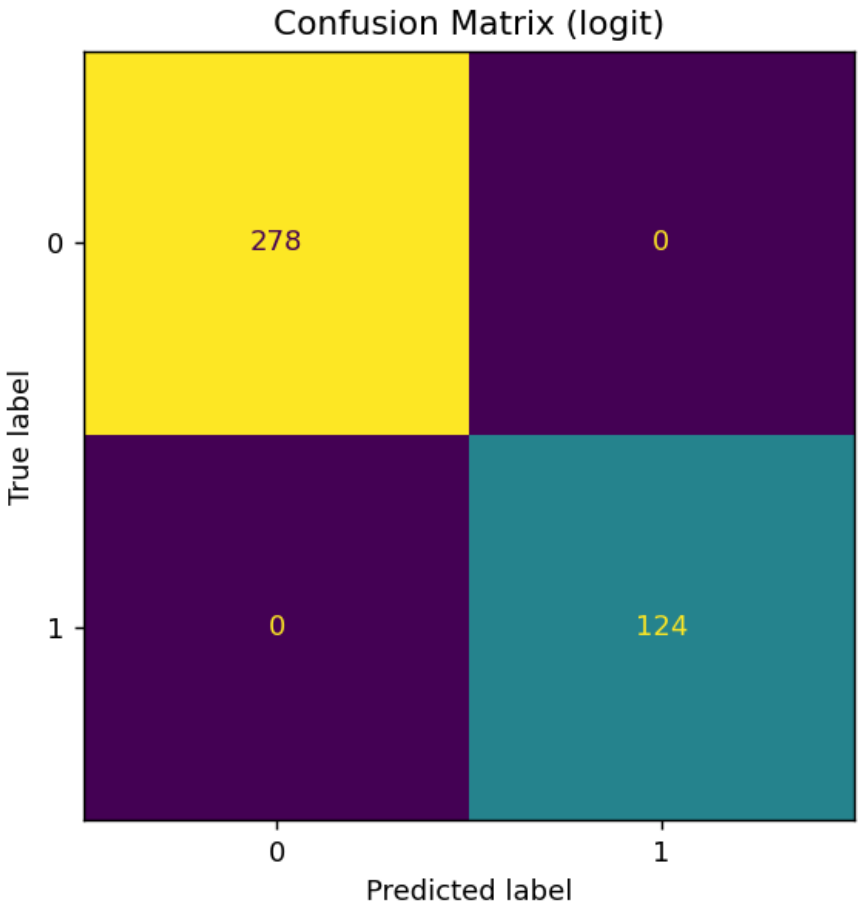


Figura 3: Matriz de confusión para la pregunta si una mujer tendrá hijos o no respondida con regresión logística. Ésta es una pregunta muy diferente a si piensa tener hijos en un transcurso de tres años.

podemos ver que no hay problemas de multicolinealidad dada la regla empírica de no superar el valor de 5. Del mismo modo, en la tabla 6 se presenta el valor de la prueba de Box-Tidwell, la cual permite afirmar, en este caso, que sí no hay suficiente evidencia para descartar la posibilidad de que se describa con una regresión logística.

Predictor	**VIF**
region_cat_3	1.736
region_cat_2	1.691
region_cat_4	1.612
age_cat_35_39	1.489
income_cat_2.0	1.481
income_cat_3.0	1.46
age_cat_30_34	1.393
region_cat_5	1.385
kids_grp	1.293
income_cat_4.0	1.256
university_bin	1.195
income_cat_5.0	1.096
born_out_bin	1.063
income_cat_6.0	1.05
employed_bin	0
routine_idx	
intermittent_idx	

Cuadro 5: VIF por predictor (Modelo 3).

Variable	p-value
kids_grp	0.9992

Cuadro 6: Prueba de Box–Tidwell (linealidad en el logit).

En el apéndice B se presentan los odds ratios del modelo logístico (tabla 8), el desempeño predictivo (tabla 9), la medida de desbalance (tabla 10) y el orden de importancia de variables respecto a distancia de Cook (tabla 11). Todas estas tablas sustentan el utilizar la regresión logística para la descripción de estos datos. Un resumen de los modelos de regresión logística se presenta a continuación en la tabla 7. Es importante notar que el único de estos modelos que tiene problemas es el modelo 1. El valor tan alto en su desvianza sugiere que el modelo subestima la información, sugiriendo que no hay evidencia en contra de agregar las interacciones que se consideran en los modelos 2 y 3.

Modelo	N	Log-Likelihood	Deviance	Pseudo R^2 (CS)
Modelo 1	1606	-1.673e+06	3.346e+06	-5.8e-13
Modelo 2	1606	-2.394e-09	6.584e-09	1
Modelo 3	1606	-2.394e-09	6.584e-09	1

Cuadro 7: Resumen compacto de ajuste de los modelos logísticos.

3.2.2. Método propuesto

Como se trata de un modelo de clasificación binaria, se propone un clasificador Naïve Bayes para trabajar estos datos. Dado lo comentado ya en el propio modelo, el resultado del clasificador Naïve Bayes, en forma de matriz de confusión en la figura 4 sólo confirma que la clasificación original es apropiada.

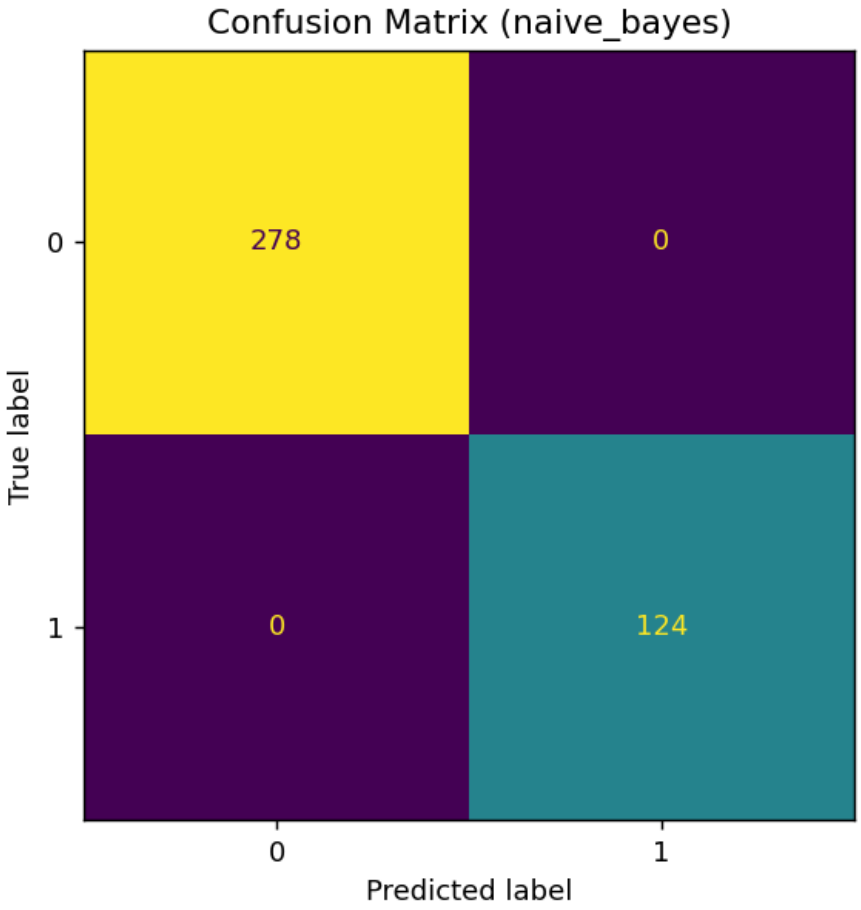


Figura 4: Matriz de confusión para la pregunta si una mujer tendrá hijos o no respondida con Naïve Bayes.

4. Conclusiones y discusión

Entre otras cosas, se puso en evidencia que el presentar un modelo de regresión no necesariamente significa que se verificó su validez. Además, hace falta claridad en la exposición de métodos en los trabajos ya que existe la posibilidad de que los datos hayan sido preprocesados y se haya trabajado con una versión distinta a la que está disponible.

Sobre el primer artículo revisado.

Una cuestión de formato importante es que no se presenta una descripción apropiada del conjunto de datos.

El análisis por separado de cada marcador clínico indicó buen comportamiento al analizar los residuales, mientras que al haber ajustado un modelo de regresión a la variable POD considerando los marcadores clínicos como una variable categórica solo obtuvimos un mal comportamiento de los residuales. Dicho esto, es mejor evaluar cada marcador clínico por separado, o bien, agruparlos de acuerdo al sistema al que están asociados (cardiovascular, metabólico, etc) y entonces revisar cómo las propiedades del PFAS afectan a dichos indicadores.

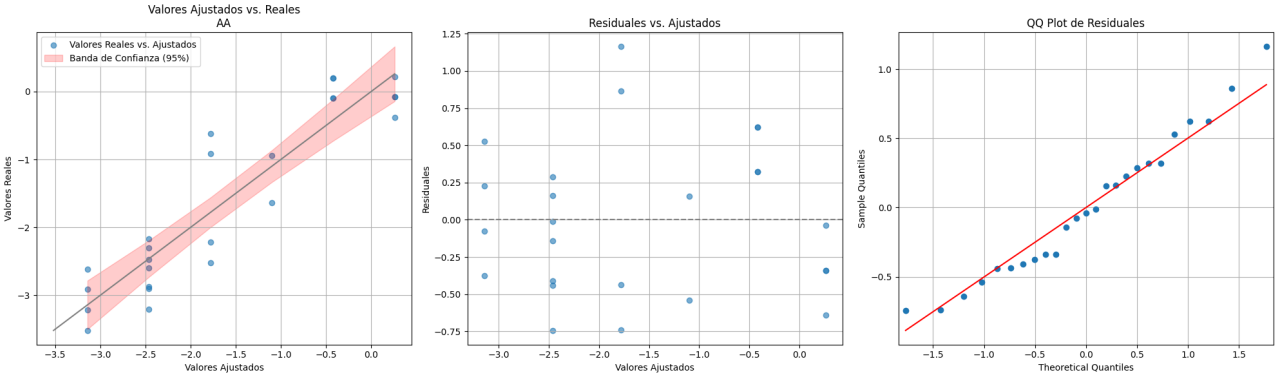
Sobre el segundo artículo revisado.

Al explorar directamente los datos es muy complicado tan siquiera validar los resultados presentados. El conjunto de datos original tiene miles de columnas codificadas de una manera no explicada en el artículo. Además, el hecho de que se agruparon artificialmente valores para construir variables binarias *descartando uno de estos valores* introduce sesgos en el análisis. Continuando la discusión iniciada al hablar de la regresión logística, al existir la posibilidad de que la variable de respuesta sea un dato sensible, sería importante reportar este preprocesamiento.

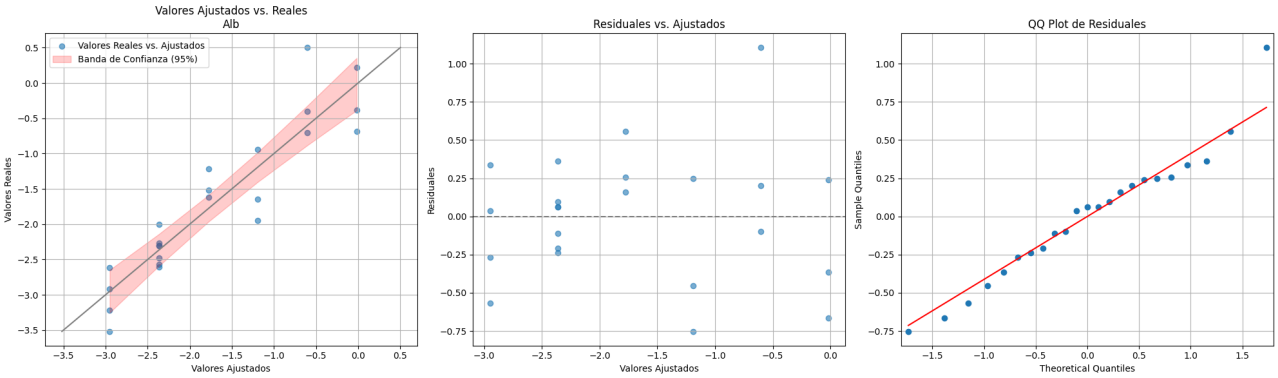
Tras revisar el problema propuesto desde la perspectiva original con regresión logística y la perspectiva propuesta con Naïve Bayes, notamos que los resultados son muy parecidos. Una posible razón para esto es el proceso tan brusco de selección de variables y de reducción de tamaño de la muestra, lo que podría fomentar que los modelos estén sobreespecificados.

Apéndices

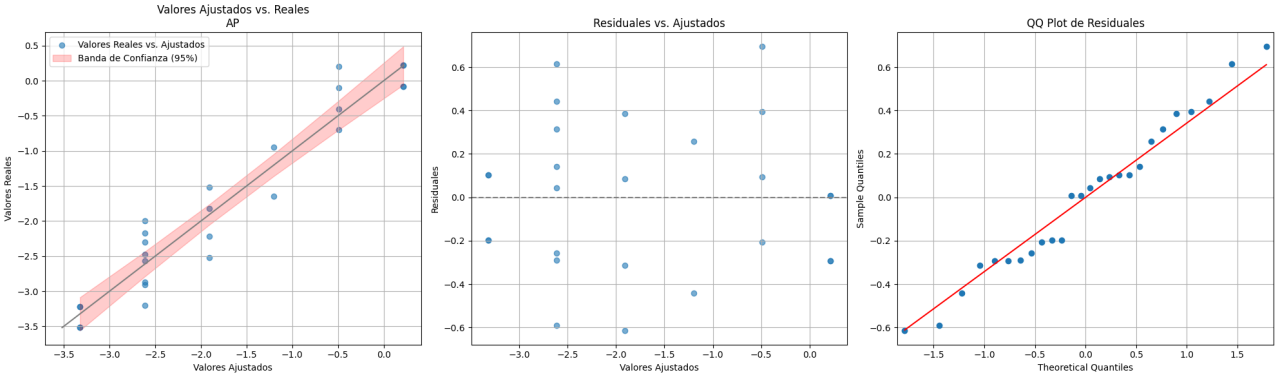
A. Complementos a regresión lineal



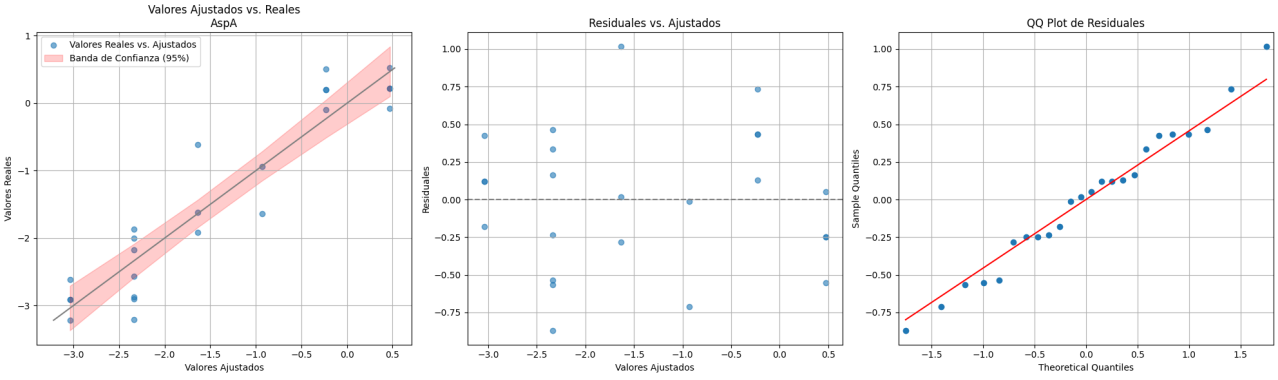
(a) Marcador clínico alanina aminotransferasa (AA)



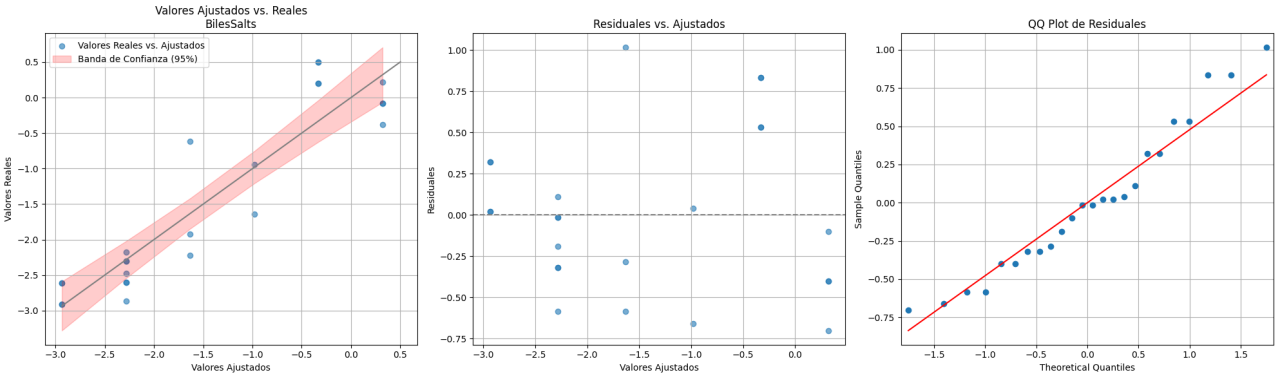
(b) Marcador clínico albúmina (Alb)



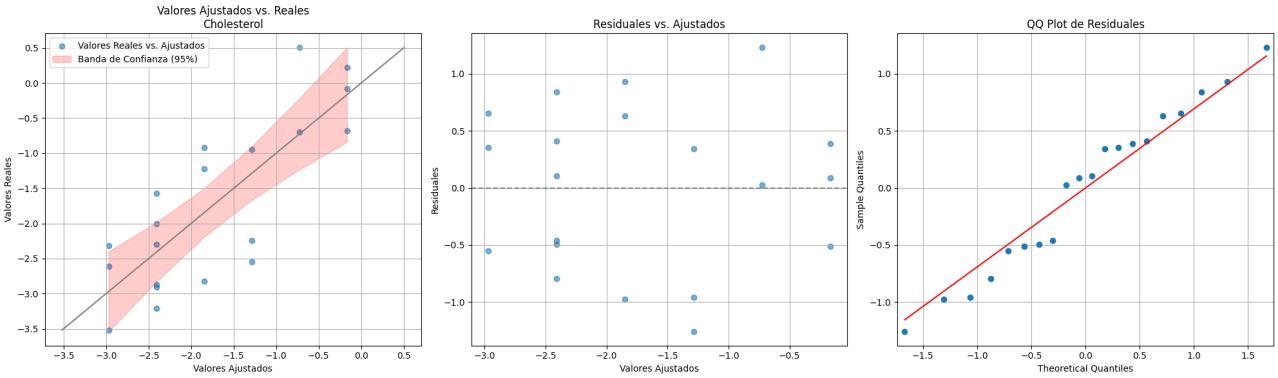
(c) Marcador clínico fosfatasa alcalina (AP)



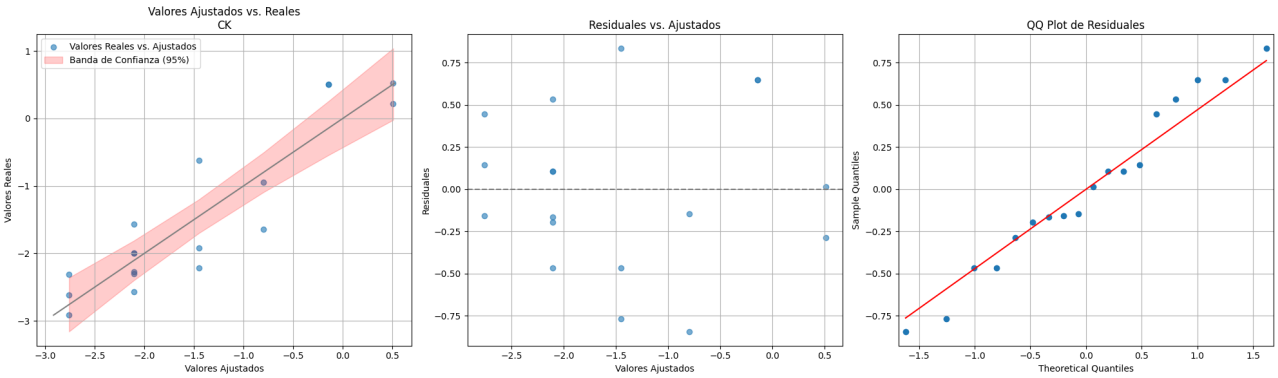
(d) Marcador clínico aspartato aminotransferasa (AspA)



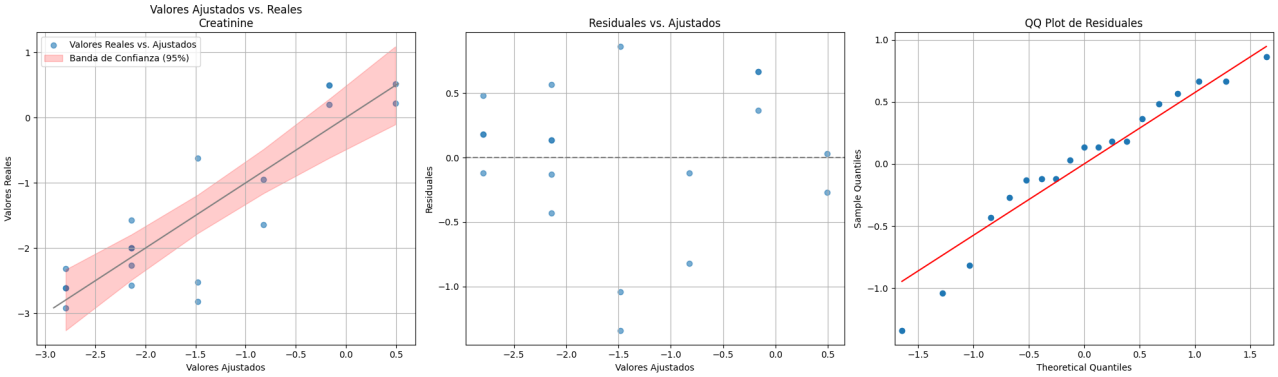
(e) Marcador clínico sales biliares (BilesSalts)



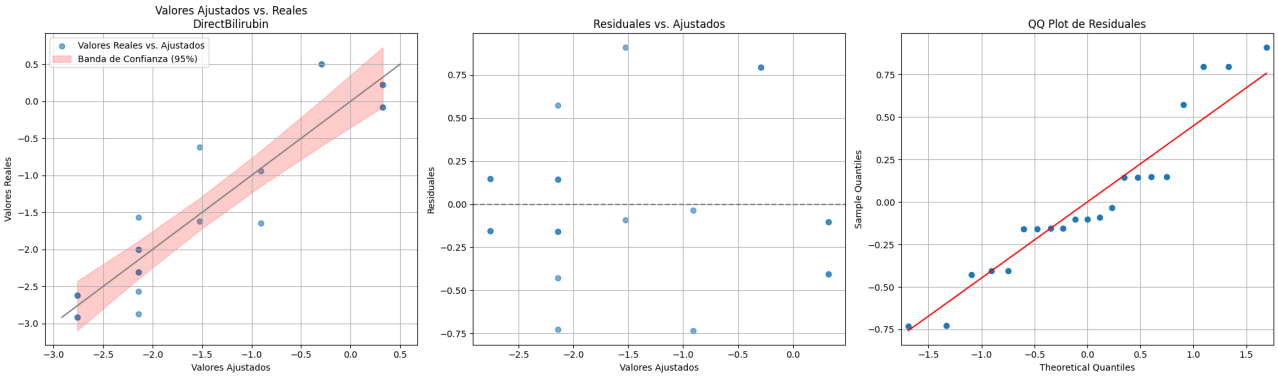
(f) Marcador clínico colesterol (Cholesterol)



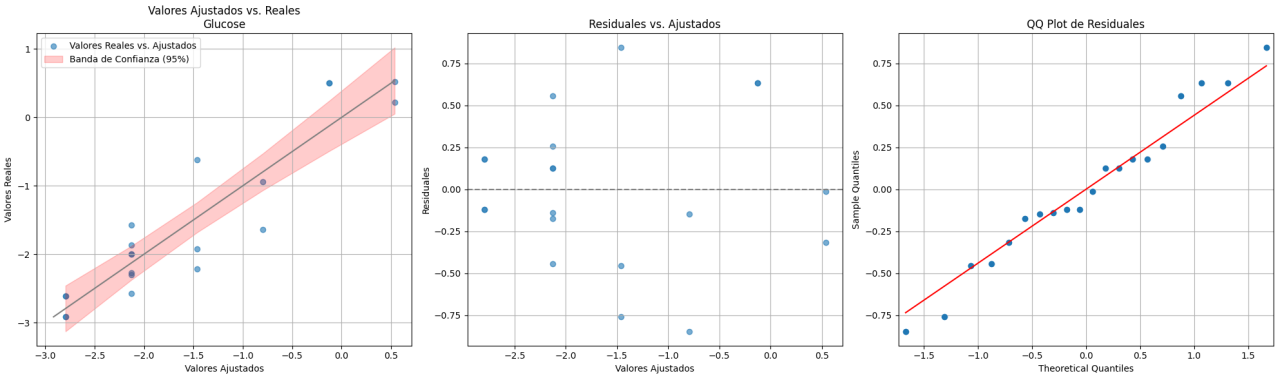
(g) Marcador clínico creatina kinasa (CK)



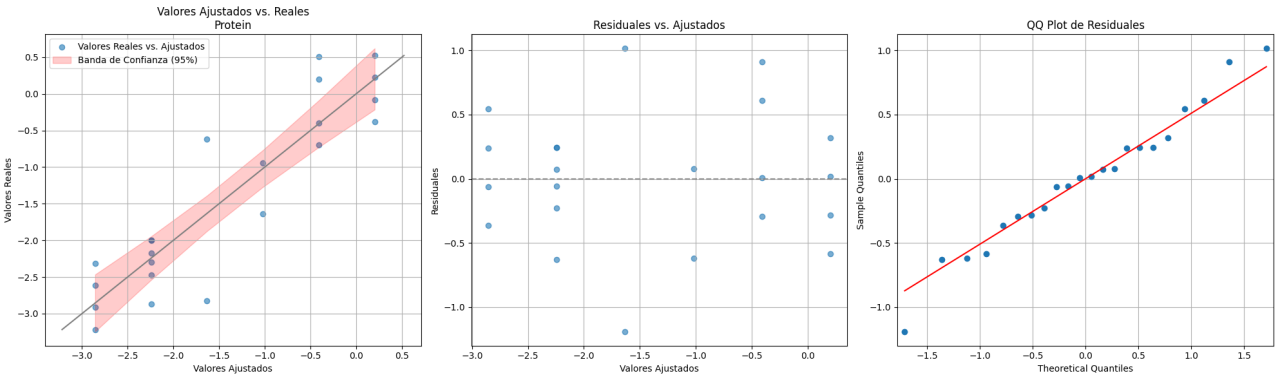
(h) Marcador clínico creatinina (Creatinine)



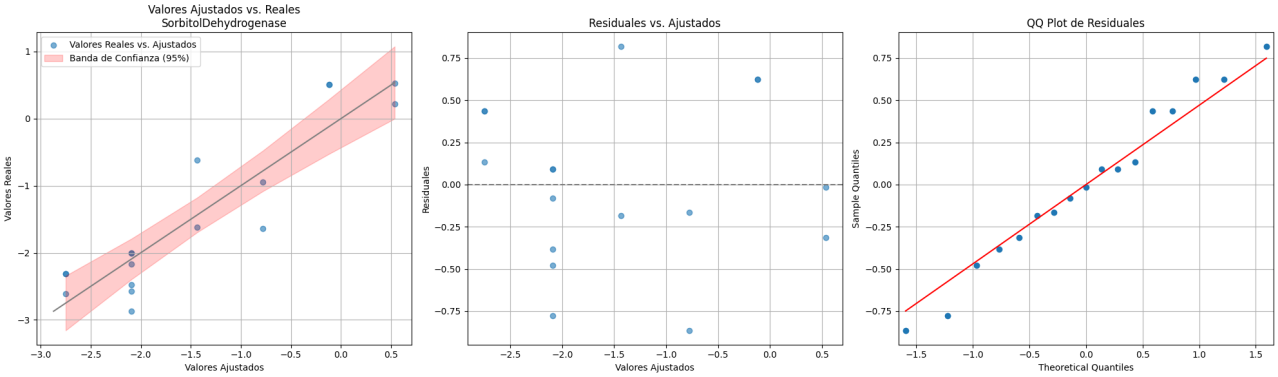
(i) Marcador clínico bilirrubina directa (DirectBilirubin)



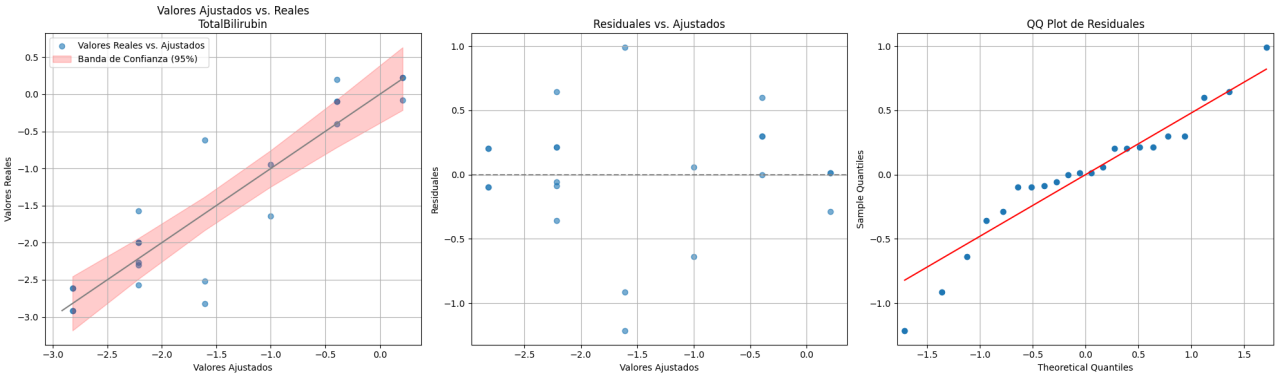
(j) Marcador clínico glucosa (Glucose)



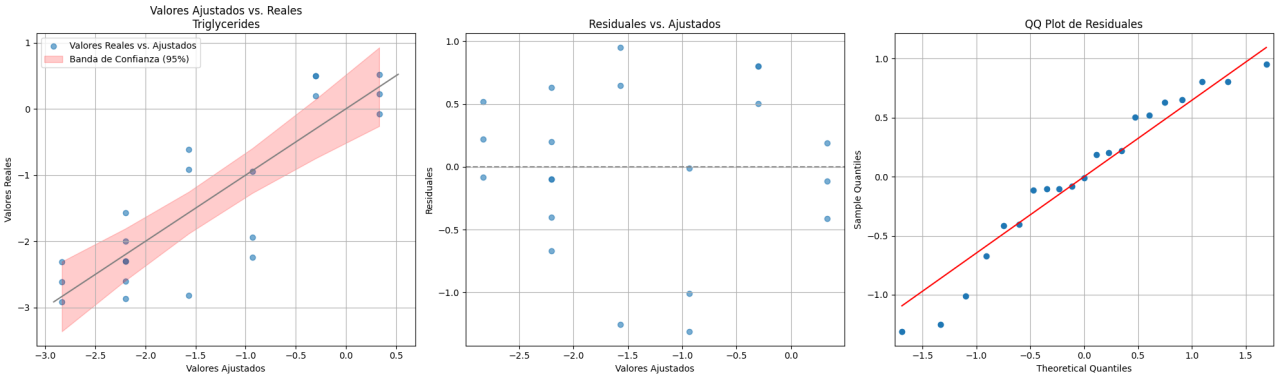
(k) Marcador clínico proteínas (Proteine)



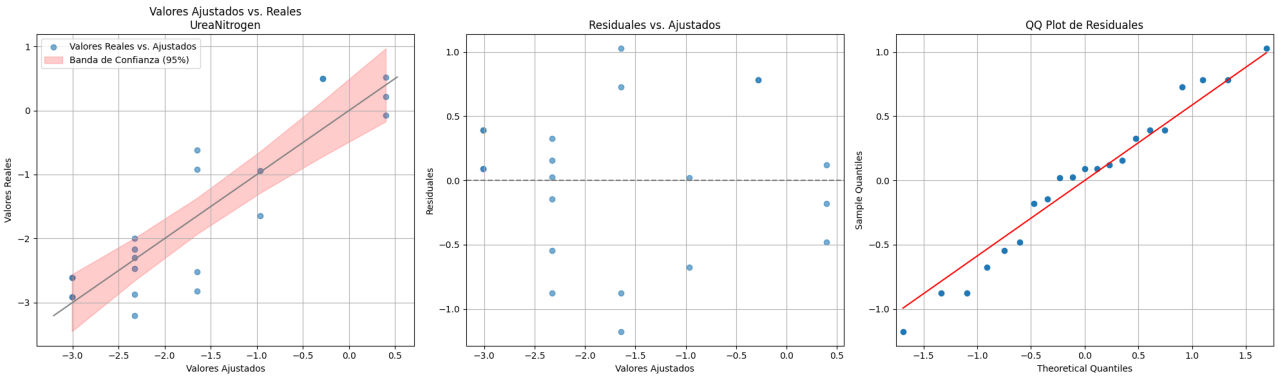
(l) Marcador clínico sorbitol deshidrogenasa (SorbitolDehydrogenase)



(m) Marcador clínico bilirrubina total (TotalBilirubin)



(n) Marcador clínico triglicéridos (Triglycerides)



(ñ) Marcador clínico nitrógeno uréico (UreaNitrogen)

B. Complementos a regresión logística

Term	OR	2.5 %	97.5 %	p-value
C(kids_grp)[T.1.0]	4.852e+08	2.061e-09	4.852e+08	0.9986
C(kids_grp)[T.2.0]	4.852e+08	2.061e-09	4.852e+08	0.9987
C(kids_grp)[T.3.0]	4.852e+08	2.061e-09	4.852e+08	0.999
employed_bin	3.119e-08	2.061e-09	4.852e+08	0.9995
Intercept	3.119e-08	2.061e-09	4.852e+08	0.9995
C(income_cat)[T.4.0]	1	2.061e-09	4.852e+08	1
university_bin	1	2.061e-09	4.852e+08	1
C(income_cat)[T.3.0]	1	2.061e-09	4.852e+08	1
born_out_bin	1	2.061e-09	4.852e+08	1
C(income_cat)[T.6.0]	1	2.061e-09	4.852e+08	1
C(region_cat)[T.3]	1	2.061e-09	4.852e+08	1
C(age_cat)[T.30_34]	1	2.061e-09	4.852e+08	1
C(region_cat)[T.4]	1	2.061e-09	4.852e+08	1
C(region_cat)[T.5]	1	2.061e-09	4.852e+08	1
C(age_cat)[T.35_39]	1	2.061e-09	4.852e+08	1
C(income_cat)[T.5.0]	1	2.061e-09	4.852e+08	1
C(region_cat)[T.2]	1	2.061e-09	4.852e+08	1
C(income_cat)[T.2.0]	1	2.061e-09	4.852e+08	1
rout_x_emp_12	1	1	1	
rout_x_emp_34	1	1	1	
int_x_emp_1	1	1	1	
int_x_emp_23	1	1	1	
rout_x_edu_12	1	1	1	
rout_x_edu_34	1	1	1	
int_x_edu_1	1	1	1	
int_x_edu_23	1	1	1	

Cuadro 8: Odds ratios (IC95 %) del modelo logístico.

Modelo	AUC
Logistic Regression (balanced)	1
GaussianNB	1

Cuadro 9: Métricas de desempeño en validación (AUC).

Métrica	Valor
Prevalencia clase positiva (Has children=1)	0.309

Cuadro 10: Desbalanceo de clases en la muestra.

Fila	Cook's D
1191	7.771e-18
1223	7.576e-18
605	6.494e-18
129	5.642e-18
199	5.045e-18
1063	4.602e-18
1114	3.923e-18
52	3.331e-18
385	2.998e-18
1592	2.455e-18

Cuadro 11: Top 10 observaciones por distancia de Cook (Modelo 3).

Referencias

[Chr20] Ronald Christensen. *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer Texts in Statistics. Springer Cham, 2020. DOI: 10.1007/978-3-030-32097-3.

[DA76] McQuarrie DA. *Statistical Mechanics*. New York, NY: Harper Row., 1976.

[EP25] Amir Erfani y Leandra Pilon. "Gender Equality in the Division of Housework and Fertility Intentions in Canada: The Moderating Effect of Employment and Education". En: *Journal of Family Issues* 46.5 (2025). © The Author(s) 2024, págs. 784-804. DOI: 10.1177/0192513X241299419.

[HLS13] David W. Hosmer, Stanley Lemeshow y Rodney X. Sturdivant. *Applied Logistic Regression*. Third. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons, 2013. DOI: 10.1002/9781118548387.

[PD25a] Andrew J. Prussia y Eugene Demchuk. *Data from "Investigating the quantitative toxicological relationship between PFAS alkyl fluorine structure and exposure levels leading to changes in blood-based clinical markers in rats"*. Dataset. 2025. URL: https://figshare.com/articles/dataset/Investigating_the_quantitative_toxicological_relationship_between_PFAS_alkyl_fluorine_structure_and_exposure_levels_leading_to_changes_in_blood-based_clinical_markers_in_rats/29488614.

[PD25b] Andrew J. Prussia y Eugene Demchuk. "Investigating the quantitative toxicological relationship between PFAS alkyl fluorine structure and exposure levels leading to changes in blood-based clinical markers in rats". En: *Journal of Toxicology and Environmental Health, Part A* 88.23 (2025), págs. 981-996. DOI: 10.1080/15287394.2025.2520427. URL: https://pubmed.ncbi.nlm.nih.gov/40620146/.

[Sta17] Statistics Canada. *General Social Survey (GSS), Cycle 31: Family, 2017 [Public Use Microdata File]*. Retrieved from the Statistics Canada Data Liberation Initiative (DLI). Contains anonymized microdata from a representative national sample of Canadians aged 15 years and over. Ottawa, Ontario, Canada: Statistics Canada, 2017. DOI: 10.25318/4501-eng. URL: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4501.

[Tib96] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". En: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), págs. 267-288. DOI: 10.1111/j.2517-6161.1996.tb02080.x. URL: https://www.jstor.org/stable/2346178.