Análisis de Artículos Científicos

INTRODUCCIÓN A CIENCIA DE DATOS

14 de octubre de 2025

Debany Jazmín Hernández Camacho debany.hernandez@cimat.mx
Eric Ernesto Moreles Abonce eric.moreles@cimat.mx
Luis Erick Palomino Galván luis.palomino@cimat.mx

1. Selección de artículos

Con el propósito de realizar un análisis crítico y replicable sobre el uso de modelos de regresión en investigaciones científicas recientes, se seleccionaron dos artículos publicados en revistas de la familia *Nature*. Cada uno emplea un tipo distinto de modelo: el primero, de regresión lineal, se orienta al estudio de fenómenos políticos en el tiempo; el segundo, de regresión logística, aborda un problema biológico predictivo con datos binarios. Ambos fueron elegidos por su claridad metodológica, disponibilidad de resultados y relevancia interdisciplinaria, lo que los hace adecuados para el trabajo de este reporte.

1.1. Regresión lineal: atención presidencial al cambio climático en México (1994–2018)

El primer artículo seleccionado, Balderas Torres et al. (2020), examina cómo la atención presidencial al cambio climático en México ha evolucionado entre 1994 y 2018. A partir del análisis de 968 documentos oficiales (discursos, informes y comunicados), los autores cuantifican la "agenda presidencial" en torno al cambio climático y distinguen entre dos dimensiones: la *agenda sistémica*, determinada por factores internacionales y legales (como el Protocolo de Kioto o el Acuerdo de París), y la *agenda gubernamental*, vinculada a los Planes Nacionales de Desarrollo.

El estudio utiliza herramientas de regresión lineal segmentada para identificar cambios estructurales en la relación entre el tiempo y el nivel de atención presidencial. En particular, se emplean pruebas de cambio de pendiente (*Chow tests*) y regresiones por tramos para estimar las diferencias en tendencia antes y después de eventos clave de política ambiental. La variable dependiente corresponde a un índice cuantitativo de atención presidencial por periodo, mientras que las variables explicativas reflejan la presencia o ausencia de instrumentos de política o coyunturas internacionales.

Este artículo nos resulta idóneo para el presente reporte por las siguientes razones,

1. Nos proporciona un ejemplo claro de regresión sobre una variable continua, con un diseño longitudinal que permite discutir los supuestos de linealidad, independencia y homocedasticidad en datos temporales.

2. La descripción de resultados, tablas y figuras (notablemente las Figs. 1 y 2 del artículo) nos ofrece suficiente información para reconstruir una versión simplificada del conjunto de datos y replicar el modelo original o estimar uno equivalente con regresión segmentada.

3. El análisis podría, posteriormente, extenderse hacia el inciso 8, explorando algún modelo alternativo o enfoque complementario que nos permita comparar el ajuste y la validez del modelo original. La elección de dicho enfoque dependerá de la naturaleza de los datos que se repliquen y de los objetivos específicos que se planteen en la siguiente etapa del trabajo.

Además, el tema combina un enfoque cuantitativo riguroso con un problema sustantivo de relevancia pública, lo que nos permite reflexionar sobre el papel de los modelos lineales en el análisis de políticas y en la representación de procesos sociales complejos.

1.2. Regresión logística: predicción de eventos de *host-shift* del virus de la rabia

El segundo artículo seleccionado, Boutelle et al. (2025), presenta un modelo de regresión logística multivariable para predecir el próximo evento de cambio de hospedero (*host-shift event*, HSE) del virus de la rabia. A partir de un conjunto de datos compuesto por $19,170$ pares reservorio–especie susceptible, el estudio busca estimar la probabilidad de transmisión interespecífica considerando características fisiológicas, ecológicas y evolutivas de las especies implicadas.

Las variables predictoras incluyen la diferencia de temperatura corporal entre especies, el grado de parentesco filogenético, la masa corporal, el tamaño de camada y el linaje viral (canino o de murciélago). La variable dependiente es dicotómica, indicando si un par de especies ha registrado o no un evento de host-shift. El modelo se ajusta mediante máxima verosimilitud y se valida con un esquema *leave-one-out cross-validation* (LOOCV), obteniendo una precisión de aproximadamente $90\,\%$, sensibilidad de $90\,\%$ y especificidad de $82\,\%$.

Este artículo es especialmente útil para el análisis comparativo por las siguientes razones,

1. Constituye un ejemplo aplicado de regresión logística en biología predictiva, en el que la variable respuesta es binaria y las covariables representan información ecológica y evolutiva cuantificable.

2. La descripción metodológica y las tablas suplementarias permiten replicar o aproximar el ajuste del modelo mediante datos simulados o subconjuntos reducidos, lo cual es esencial para los siguientes puntos del reporte.

3. Facilita la exploración de modelos alternativos, tales como regresión logística penalizada (Ridge o Lasso) o enfoques bayesianos con priors débiles, evaluando su desempeño con métricas como AUC, precisión y curvas de calibración.

Además de su relevancia biológica, el artículo destaca por su diseño estadístico transparente y su capacidad predictiva, lo que nos permite discutir de manera crítica la interpretación de los coeficientes, la influencia de la multicolinealidad y el equilibrio entre precisión y generalización del modelo.

En conjunto, los dos artículos seleccionados representan contextos científicos distintos (uno político-social y otro ecológico-biológico), pero ambos ilustran el uso riguroso y justificado de modelos de regresión. Su análisis nos permitirá contrastar la aplicación de supuestos, la interpretación de parámetros y la validación de resultados en diferentes disciplinas, cumpliendo plenamente con los objetivos planteados en esta segunda parte del reporte.

2. Análisis crítico

El propósito de esta sección es examinar de manera crítica el uso de modelos de regresión en los artículos seleccionados, atendiendo tanto a los supuestos estadísticos que sustentan cada enfoque como a la forma en que los autores interpretan y validan sus resultados. Este análisis no se limita a una descripción de los procedimientos empleados, sino que busca valorar la coherencia entre los objetivos de investigación, el tipo de variable respuesta, los predictores considerados y la idoneidad del modelo elegido.

En términos generales, la evaluación se organiza en tres niveles. En primer lugar, vamos a analizar las variables y la estructura del modelo, identificando cómo se relacionan las covariables con la variable dependiente y si estas relaciones cumplen los supuestos teóricos de la regresión utilizada. En segundo lugar, se revisan los supuestos estadísticos y metodológicos, como la linealidad, independencia, homocedasticidad y normalidad en el caso del modelo lineal, o la linealidad en el logit y la ausencia de multicolinealidad en el modelo logístico. Finalmente, se discute la interpretación y adecuación de los resultados, considerando la validez del modelo, la significancia de los parámetros estimados, las limitaciones reconocidas por los autores y las posibles alternativas de modelación que podrían fortalecer el análisis.

El objetivo no es únicamente replicar los resultados reportados, sino comprender el razonamiento estadístico detrás de las decisiones de modelado y su pertinencia en el contexto de cada estudio. A continuación, se presentan los análisis críticos correspondientes a los dos artículos seleccionados: el primero, que aplica un modelo de regresión lineal para estudiar la atención presidencial al cambio climático en México, y el segundo, que emplea un modelo de regresión logística para predecir eventos de cambio de hospedero del virus de la rabia.

2.1. Estudio de caso: aplicación de la regresión lineal

El estudio de Balderas Torres et al. (2020) aplica un enfoque de regresión lineal segmentada para analizar cómo la atención presidencial al cambio climático en México varió entre 1994 y 2018. A partir del análisis de 968 documentos oficiales (discursos, informes y comunicados), los autores cuantifican la frecuencia con la que los temas ambientales aparecen en la agenda presidencial y examinan cómo esta atención se relaciona con eventos internacionales y nacionales de política climática. Su objetivo es identificar si los compromisos multilaterales, como el Protocolo de Kioto o el Acuerdo de París, influyeron en el nivel de prioridad otorgado al tema dentro del discurso político mexicano.

Variables dependientes e independientes. La variable dependiente corresponde a un índice cuantitativo de atención presidencial al cambio climático, calculado a partir de la frecuencia de menciones en documentos oficiales por periodo (mensual o anual). Las variables independientes representan factores de agenda, tanto sistémicos (eventos internacionales, acuerdos globales, cumbres ambientales) como gubernamentales (planes nacionales y políticas domésticas). Además, el tiempo actúa como covariable principal, permitiendo evaluar cambios de tendencia a lo largo de la serie temporal.

Supuestos implícitos del modelo. El modelo asume las condiciones clásicas de la regresión lineal:

- linealidad en la relación entre la atención presidencial y el tiempo o los factores de agenda,
- independencia de los errores entre observaciones sucesivas,

- homocedasticidad de la varianza de los residuos, y
- normalidad de los errores.

Sin embargo, debido a la estructura temporal de los datos, es probable que el supuesto de independencia no se cumpla plenamente, ya que la atención presidencial de un año puede depender de la del año anterior. Tampoco se documenta en el artículo la verificación de homocedasticidad ni de normalidad de los residuos, lo que limita la evaluación formal del ajuste.

Interpretación de los coeficientes. Los coeficientes de las regresiones por segmentos reflejan la pendiente de cambio en el nivel de atención presidencial dentro de cada periodo. Un coeficiente positivo indica un aumento sostenido en la frecuencia de menciones ambientales, mientras que un valor cercano a cero sugiere estabilidad o desinterés relativo. Los autores encuentran un incremento en la pendiente posterior a eventos internacionales clave, lo cual interpretan como evidencia de que la política exterior y los compromisos internacionales influyen directamente en la agenda interna de la Presidencia mexicana.

Adecuación del modelo. La regresión lineal segmentada es una elección razonable para describir variaciones en series temporales donde se anticipan cambios estructurales. Permite estimar pendientes diferenciadas antes y después de eventos relevantes, y facilita la interpretación substantiva de los resultados. No obstante, para fines de inferencia estadística, el modelo clásico puede resultar limitado al no incorporar la autocorrelación ni la posible naturaleza discreta del índice de atención (que se deriva de conteos). En este sentido, el modelo cumple con el objetivo descriptivo y exploratorio del estudio, pero podría ser mejorado en términos de precisión estadística y validez de los supuestos.

Limitaciones y posibles alternativas. Entre las principales limitaciones destacan:

- la posible dependencia temporal entre observaciones consecutivas,
- la falta de evidencia sobre homocedasticidad y normalidad de los residuos, y
- la naturaleza discreta de la variable dependiente.

Para abordar estos puntos, se podrían considerar alternativas como:

- Un modelo de regresión Poisson o binomial negativa, adecuado para datos de conteo con sobredispersión.
- Un modelo de series de tiempo con estructura autoregresiva (AR o ARIMA), que permita controlar la correlación serial de los errores.
- Un modelo bayesiano jerárquico, que incorpore la incertidumbre en los parámetros y permita comparar formalmente los efectos de las agendas sistémica y gubernamental.

En síntesis, el artículo constituye una aplicación sólida del modelo lineal con fines descriptivos, al mostrar cómo los cambios en la política internacional y los compromisos multilaterales se reflejan en la atención presidencial. Sin embargo, desde el punto de vista estadístico, aún existen oportunidades para fortalecer el análisis mediante modelos que consideren explícitamente la estructura temporal y la naturaleza discreta de los datos.

2.2. Estudio de caso: aplicación de la regresión logística

El artículo de Boutelle et al. (2025) presenta una aplicación reciente del modelo de regresión logística multivariable en el campo de la epidemiología y la biología evolutiva. El objetivo de los autores es desarrollar un modelo predictivo capaz de estimar la probabilidad de que ocurra el próximo evento de cambio de huésped (host-shift event, HSE) del virus de la rabia, a partir de información biológica, filogenética y ecológica de las especies involucradas. Este tipo de eventos representa una de las principales vías de emergencia de nuevas enfermedades infecciosas, por lo que comprender los factores asociados a su aparición tiene una gran relevancia en salud pública.

Variables dependientes e independientes. La variable dependiente es dicotómica y representa si un par de especies reservorio–susceptible ha registrado un evento documentado de cambio de huésped (1) o no (0). Las variables explicativas incorporan características biológicas de ambas especies, como la diferencia en temperatura corporal, la distancia filogenética, el tamaño de camada, el peso corporal promedio y el linaje viral (de origen canino o de murciélago). Estas covariables permiten explorar simultáneamente la influencia de rasgos fisiológicos y evolutivos en la probabilidad de transmisión interespecífica. El conjunto de datos comprende $19,170$ pares de especies, lo que confiere al estudio un tamaño muestral amplio y adecuado para este tipo de modelo.

Supuestos implícitos del modelo. El modelo de regresión logística supone que la variable respuesta sigue una distribución Bernoulli y que la relación entre las covariables y el logaritmo de las probabilidades (logit) es lineal. También se asume la independencia entre observaciones, ausencia de multicolinealidad entre los predictores y varianza binomial de los errores. Si bien el artículo menciona la validación cruzada tipo "leave-one-out (LOOCV)", no se detalla la verificación empírica de los supuestos de linealidad en el logit ni de colinealidad entre predictores. En este tipo de datos biológicos, algunas variables, como la temperatura corporal y el tamaño de cuerpo, tienden a estar correlacionadas, por lo que una exploración adicional de estas relaciones podría fortalecer el modelo.

Interpretación de los coeficientes. Los coeficientes estimados se interpretan en términos de .oddsratios", es decir, como el cambio multiplicativo en las probabilidades de un evento de host-shift asociado a una unidad de cambio en cada covariable, manteniendo las demás constantes. Por ejemplo, una disminución en la diferencia de temperatura corporal entre especies incrementa la probabilidad de transmisión del virus, lo cual coincide con la hipótesis biológica de que una mayor similitud fisiológica facilita la adaptación viral. Asimismo, los resultados muestran que la proximidad filogenética y ciertos rasgos reproductivos aumentan la susceptibilidad al cambio de huésped, mientras que linajes virales específicos (como los caninos) presentan una propensión mayor a la transmisión interespecífica. Estas interpretaciones son coherentes con la teoría evolutiva y refuerzan la validez biológica del modelo.

Adecuación del modelo. El uso de una regresión logística multivariable es apropiado dado que la variable respuesta es binaria y las covariables son de naturaleza continua y categórica. Este modelo permite estimar probabilidades ajustadas e identificar factores de riesgo significativos, lo cual cumple con los objetivos del estudio. Los autores aplican un procedimiento de validación cruzada exhaustivo (LOOCV) que reporta una precisión cercana al 90 %, con sensibilidad de 90 % y especificidad de 82 %, lo que indica un alto poder predictivo y una buena capacidad de generalización. Sin embargo, el artículo no detalla si se realizó una evaluación de calibración de probabilidades, ni si se analizaron posibles sesgos derivados de un desbalance entre clases (eventos positivos y negativos), lo que podría ser relevante para la robustez del modelo predictivo.

Limitaciones y posibles alternativas. Entre las principales limitaciones se encuentra la posible multicolinealidad entre las variables biológicas y la ausencia de un análisis explícito de linealidad en el logit. Además, la independencia entre observaciones podría verse comprometida, ya que un mismo reservorio o especie susceptible aparece en múltiples pares dentro del conjunto de datos. Estas limitaciones podrían abordarse mediante extensiones o modelos complementarios. Por ejemplo, la aplicación de una regresión logística penalizada (Ridge o Lasso) podría reducir la colinealidad y estabilizar las estimaciones. También podría considerarse un modelo logístico jerárquico o mixto que incorpore efectos aleatorios por especie, controlando así la dependencia entre pares. Finalmente, un enfoque bayesiano permitiría incluir priors débiles en los coeficientes y cuantificar de manera más completa la incertidumbre en las probabilidades estimadas.

En conjunto, el artículo ofrece una aplicación rigurosa y novedosa del modelo de regresión logística en un contexto biológico complejo, logrando integrar información filogenética y ecológica dentro de un marco estadístico interpretativo. La claridad con que se presentan los resultados, junto con su validación cuantitativa, respalda la utilidad de la regresión logística como herramienta predictiva. No obstante, un análisis más profundo de los supuestos estadísticos y una comparación con enfoques alternativos podrían fortalecer aún más la confiabilidad y generalización de los hallazgos.

3. Replicación de Resultados

3.1. Regresión Lineal

Para el caso de regresión lineal se presenta la gráfica de comunicaciones acumuladas contra menciones en los NDP (planes nacionales de desarrollo, por sus siglas en inglés) a lo largo de todo el periodo analizado. Los datos proporcionados son el listado de todas las publicaciones en paginas presidenciales que mencionan el cambio climático por mes. La lista esta en PDF; usando excel podemos fácilmente pasarlo a una hoja de calculo y guardarlo como csv. Contabilizando de manera apropiada, tenemos los mismo números reportados en el artículo.

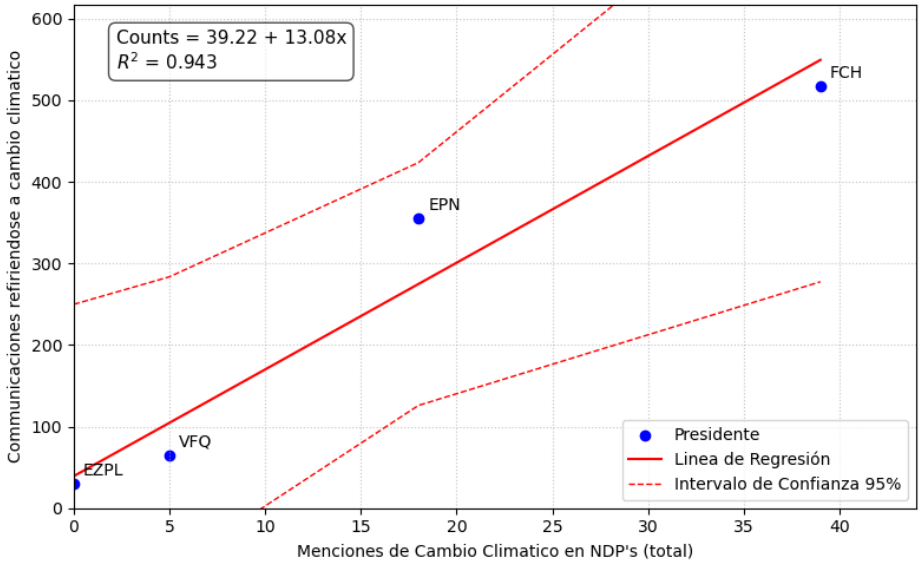

Figura 1: Regresión lineal, menciones de cambio climático en planes de desarrollo nacional en México contra comunicaciones refiriéndose a cambio climático publicadas en paginas web presidenciales.

Para contabilizar el número de menciones en los NDP's, se usan herramientas manuales, pues es mas fácil buscar en los documentos la palabra çlima"que hacer un programa que busque instancias o referencias al cambió climático en el documento. De esta manera tenemos igual el mismo número de menciones en los NDP's para los cuatro sexenios. En el código se muestran gráficos adicionales concurrentes con los vistos en el artículo, el de interés lo adjuntamos en la Figura 1.

Vemos que tenemos los mismo parámetros obtenidos en el artículo, y se muestra claramente que a más atención se le ponga al tema de cambio climático, mas acción se ve reflejada en las NDP's. Cabe mencionar que aunque la iniciativa se tome para afrontar el cambio climático, se menciona en el artículo que en varias ocasiones no necesariamente se llevaron a cabo las propuestas.

3.2. Regresión Logística

Los datos, junto con los resultados, están disponibles en la página de donde se consiguió el artículo. El propósito es ajustar un modelo de regresión logística multivariable para identificar si un par de especies reservorio-susceptible es un HSE. La herramienta usada en el artículo para este análisis es R, nosotros usamos Python pues nos es mas familiar. Usamos las variables predictivas reportadas para el análisis, y para ajustar el modelo usamos LogisticRegression de la librería Sklearn.linear_model. Tenemos resultados similares al artículo, pero hay diferencias notables. Del resumen del modelo, tenemos un $R^2 = 0.33$, sin embargo encontramos un umbral de alto riesgo diferente, 0.001, y esto nos lleva a resultados diferentes, una precisión de $88\,\%$ contra $86\,\%$ (calculada como (precisión + sensitividad)/2), la precisión de Leave One Out Cross Validation de un $92\,\%$. Como tenemos un umbral de riesgo diferente, tenemos también una razón de riesgo (RR) diferente, para los casos conocidos de HSE tenemos una media de RR de 68.61, mientras que para los casos en los cuales no son HSE tenemos una media de 1.36. Se presenta el boxplot de las RR separadas por HSE en la Figura 2.

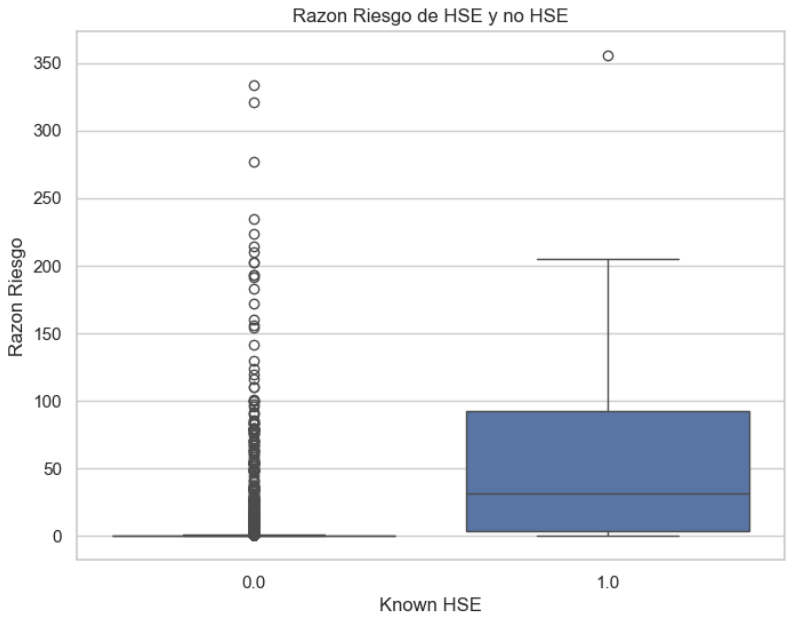

Figura 2: Boxplot de la Razon Riesgo de los datos separados entre HSE

En el código se incluye una manera de imprimir los 20 casos con mayor RR por continente, presentamos aquí el caso para América Central y del Norte:

Common Name	Variant	Known HSE	Razon Riesgo
Coyote	Canine-associated	1.0	355.526466
Arctic Fox	TX/MX Coyote	0.0	234.739848
Arctic Fox	Canine-associated	1.0	193.538212
Bush dog	TX/MX Coyote	0.0	159.804932
Coyote	Oregon gray fox	0.0	141.659470
Gray wolf	Canine-associated	1.0	141.244198
Bush dog	Canine-associated	0.0	129.529663
Striped skunk	Eastern raccoon	0.0	119.970788
Arctic Fox	Oregon gray fox	0.0	100.307738
Swift fox	Oregon gray fox	0.0	97.879275
Coati	Eastern raccoon	0.0	95.590652
Gray Fox	TX/MX Coyote	0.0	91.656786
Red fox	TX/MX Coyote	0.0	90.705400
Kit fox	Oregon gray fox	0.0	90.022130
Swift fox	Arctic fox	0.0	82.859733
Coyote	Arctic fox	0.0	79.820382
Coyote	Maine gray fox	0.0	78.766347
Coyote	California gray fox	0.0	78.766347
Coyote	Texas gray fox	0.0	78.766347
Coyote	Ontario red fox	0.0	78.766347

Cuadro 1: Riesgo de transmisión susceptible-reservorio, se excluyen casos donde ambos la especie y la variante son zorrillos.

Es difícil ver en que difieren ambos modelos, puse la tabla generada no corresponde con la reportada en el artículo, un posible culpable sería diferencias numéricas entre R y Python, después de todo R es dedicado a aplicaciones numéricas.

Exploración de alternativas

En esté apartado, vamos a explorar alternativas al análisis de regresión lineal y logística usadas en sus respectivos artículos con el fin de dar una exploración alternativa y más completa.Para ello, se emplearán métodos de regresión alternativos como Ridge, Lasso y regresión robusta para el caso lineal, y regularización L2 o modelos bayesianos para el caso de regresión logística. Primero veamos sus principales características para identificar como y cuando la usamos.

1. **Regresión robusta**. Es una familia de métodos diseñados para ser menos sensibles a los valores atípicos. En lugar de dar el mismo peso a todas las observaciones, asigna un peso menor a los puntos que se alejan mucho de la tendencia general. En particular, vamos a usar la regresión Huber que es un tipo de regresión robusta, diseñada para ser menos sensible a los valores atípicos.

2. **Ridge**. Reduce la magnitud de los coeficientes de las variables correlacionadas. Es útil cuando tenemos muchas variables y creemos que todas son relevantes en cierta medida.

3. **Lasso**. Puede reducir los coeficientes de algunas variables a cero, eliminándolas efectivamente del modelo.

Esto lo convierte en una excelente herramienta para la selección de variables, es decir, para identificar cuáles son los predictores más importantes.

4. **Regularización**. Añade una ”penalización.ªl modelo por tener coeficientes muy grandes. Esto obliga al modelo a ser más simple y a generalizar mejor.

    - L1: De manera similar a Lasso, su principal ventaja es que puede eliminar variables no relevantes al reducir su coeficiente a cero.
    - L2: De manera similar a Ridge, reduce la magnitud de todos los coeficientes, siendo muy útil para manejar la multicolinealidad.

5. **Modelos Bayesianos**. El modelo consideran los parámetros como variables aleatorias. Combina el conocimiento previo (si lo hay) con los datos observados para obtener una distribución posterior que representa nuestro conocimiento actualizado.

The systemic and governmental agendas in presidential attention to climate change in Mexico (1994–2018) — Nature Communications (2020).

El articulo utiliza regresión lineal para mostrar una correlación entre la atención al cambio climático en los Planes Nacionales de Desarrollo (PND) y el número de comunicaciones presidenciales sobre el tema. Sin embargo, existen otras técnicas que podrían ofrecer una compresión más profunda de los datos.

La regresión lineal es un método que se basa en ciertos supuestos que no siempre se cumplen. En particularidad, podemos notar tres aspectos del articulo que hacen que la regresión lineal no sea la mejor opción. Veamos los tres puntos:

- El análisis se basa en el número de comunicaciones mensuales sobre el cambio climático, sin embargo, como vemos en la tabla, hay picos de atención muy pronunciados, como a finales de 2010 coincidiendo con la COP 16 en Cancún. Estos valores extremos pueden tener una influencia desproporcionada en una regresión lineal. Por lo que hay que considerar los datos atípicos o **outliers**.

- Para construir un modelo más completo que nos ayude a explicar la atención presidencial, podríamos incluir otras variables como la opinión pública, la cobertura mediática, eventos climáticos relevante o indicadores económicos. De hecho, notemos que podría ocurrir que varias de estas variables estén correlacionados entre si, por lo que podríamos tener datos con **multicolinealidad** y puede hacer que los resultados de la regresión lineal sean inestables y difíciles de interpretar.

- Una nota que vale la pena recalcar: El aumento en las menciones en el PND es un aumento proporcional en las comunicaciones, o al menos así se interpreta en el análisis. Pero, podría haber un punto de saturación en el que por más que se mencione el tema, el número de comunicaciones no aumente al mismo ritmo, por lo que podríamos tener relaciones no lineales.

Una alternativa para resolver el problema de los datos atípicos, es usar la regresión robusta. Al aplicarla, obtendríamos una estimación de la relación entre los planes de desarrollo PND y la atención presidencial que no esté ta influenciada por los picos de atención de eventos específicos como las COPs. El resultado sería una visión más fiable y descubrir si la relación base es más fuerte o más débil de lo que la regresión lineal sugiere.

Para entender qué factores, además de los PND, influyen en la atención presidencial, podríamos usar regresión Lasso, ya que nos ayuda a encontrar de todos los factores posibles y cuáles son los que realmente impulsan la atención presidencial sobre el cambio climático. Lo que nos daría un modelo más complejo sin tener problemas con multicolinealidad.

Para completar el análisis alternativo, se tomaron los datos e hicimos una serie temporal de comunicados presidenciales, donde marca 1 si hay menciones a palabras clave ambientales y 0 en caso contrario. Con ello, construimos una variable dependiente que representa el nivel de atención mensual al cambio climático, luego aplicamos modelos de regresión lineal robusta para explicar su evolución en función de eventos o factores políticos.

Cuadro 2: Comparación de Modelos de Regresión

Modelo	**R^2 (train)**	**RMSECV**	**R^2 (media)**
OLS	0.351	—	-1.02 ± 2.06
Ridge	0.097	2.99	-0.155 ± 0.266
Lasso	0.000	3.14	-0.077 ± 0.085
Regresión robusta	0.229	2.76	-0.197 ± 0.267

Analizando la tabla, vemos que el modelo OLS tiene un 35.1 en entrenamiento moderado, lo que muestra un sobreajuste significativo, pero un rendimiento muy pobre en validación cruzada (algo esperable, porque las menciones presidenciales ocurren en apariciones espontaneas). Además, la alta desviación estándar que llega a ± 2.06 indica una gran inestabilidad. Sin embargo, Lasso simplifica en exceso la relación entre las variables, de hecho las anula en su capacidad predictiva, por lo que no logra capturar la variabilidad de los datos. Por otro lado, la regresión robusta obtuvo el mejor balance entre el ajuste y estabilidad (lo que se esperaba) aunque con menor ajuste en los datos de entrenamiento.

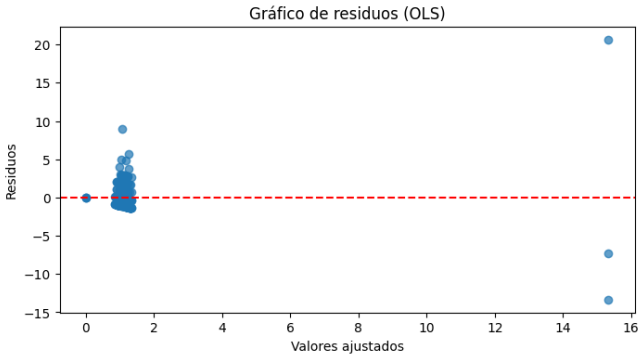

Figura 3: Gráfico de Residuos vs Valores Ajustados .

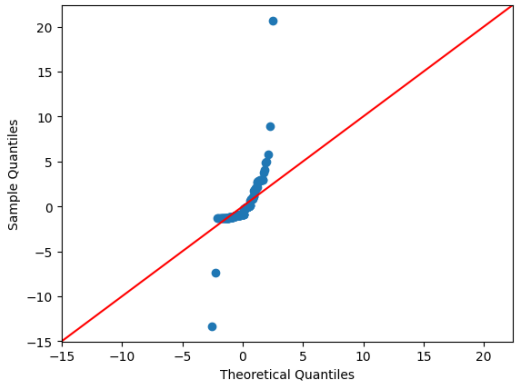

Figura 4: Gráfico Q-Q.

En el gráfico Figura 3, nos ayuda a evaluar la varianza de los errores y la presencia de outliers. Podemos ver que la magnitud de los errores del modelo aumenta a medida que los valores predichos crecen y existen putnos claramente alejados del resto, lo que indica observaciones con influencia desproporcionada sobre el modelo.

En el gráfico Figura 4, compara los cuantiles de los residuos con los cuantiles de una distribución normal. Podemos notar que los puntos se desvían de la línea, especialmente en los extremos. Esto puede indican que las colas de tu distribución son más pesadas, por lo que en efecto hay valores atípicos extremos que hacen que la distribución de los errores no sea normal.

Los resultados muestran que el artículo sobreestima la capacidad explicativa de la regresión lineal y subestima la irregularidad temporal. Mientras que el modelo de regraón robusta representa mejor la realidad, reflejando picos de atención altos en momentos clave, pero sin tendencia sostenida ni efectos duraderos de eventos internacionales.

A logistic regression model to predict the next rabies virus host-shift event — scientific reports (2025).

El artículo emplea una regresión logística para predecir la probabilidad de un Host- Shift event (HSE) de virus de la rabia, basándose en características biológicas y ecológicas. Sin embargo, existen alternativas que pueden ofrecer ventajas importantes en este contexto.

La regresión logística tiene algunas limitaciones que podrían ser relevantes en el articulo. Nuevamente notamos dos aspectos del articulo que hacen que la regresión logística no sea la mejor opción. Veamos los dos puntos:

- Los autores utilizaron un método de eliminación hacia atrás para decidir qué variables incluir en el modelo final, pero pequeños cambios en los datos podrían llevar a un modelo final diferente. Además, con varias variables predictoras (parentesco, temperatura, tamaño de camada, peso), siempre existe el riesgo de que el modelo se **sobreajuste** a los datos existentes y no generalice tan bien a nuevos casos en el futuro.
- Notemos que el peso de una especie podría estar relacionado con el tamaño de su camada, por lo que podríamos tener un problema de colinealidad. Aun que en el articulo se verifico que no hubiera una **multicolinealidad** excesiva con VIF< 10, podría ocurrir que al hacer las estimaciones de los coeficientes del modelo sean inestables y difíciles de interpretar.

Una forma de hacer mejor la selección de variables es con regularización L1, ya que podría identificar de manera más objetiva cuáles son los factores biológicos y ecológicos verdaderamente críticos para predecir un cambio de huésped. En lugar de depender de un proceso de eliminación por pasos, Lasso seleccionaría el subconjunto de variables más predictivo.

También utilizaremos un modelo bayesiano utilizanod MCMC ya que nos ayuda a cuantificar explicitamente la incertidumbre, lo cual es información crucial para la toma de decisiones, ya que permite a las autoridades de salud pública entender el rango completo de los posibles resultados. La interpretación de los resultados se basa en los Intervalos de Alta Densidad (HDI), que representan el rango donde se encuentra el verdadero valor del parámetro con una cierta probabilidad (en este caso, $94\,\%$).

Veamos como se hizo exploración alternativa. Utilizando Python, la variable categórica Weight Difference fue convertida a un formato numérico mediante codificación one-hot encoding. Esto crea columnas binarias separadas para cada categoría, permitiendo su inclusión en el modelo de regresión.

El conjunto de datos fue dividido en un $70\,\%$ para entrenamiento y un $30\,\%$ para prueba . Se utilizó un muestreo estratificado para asegurar que la proporción de eventos HSE (la clase minoritaria) fuera la misma en ambos conjuntos, lo cual es crucial para datos desbalanceados. Además, todas las variables predictoras numéricas fueron estandarizadas utilizando. Este se usa para los modelos de regularización, ya que asegura que la penalización se aplique de manera justa a todos los coeficientes.

El modelo Lasso demostró ser inadecuado para este conjunto de datos. La fuerte penalización, combinada con el extremo desbalance de clases (5743 casos negativos vs. 8 positivos), resultó en la anulación de todos los coeficientes de las variables predictoras.

```
                  Variable  Coeficiente_Lasso
0               Relatedness                0.0
...
```

El reporte de clasificación reveló una precisión del 100 %, la cual es engañosa. El modelo logró esta métrica al clasificar correctamente todos los casos negativos y fallar en identificar todos los casos positivos, resultando en un recall de cero para la clase de interés. Por lo tanto, el modelo carece de poder predictivo para los HSE.

Por otro lado, el modelo Bayesiano convergió exitosamente. La Tabla 3 resume las distribuciones posteriores para cada coeficiente. La principal herramienta de inferencia es el intervalo de credibilidad del $94\,\%$ (HDI).

Cuadro 3: Distribuciones posteriores de los coeficientes del modelo Bayesiano.

Variable	**mean**	**sd**	**hdi_3 %**	**hdi_97 %**
Intercept	-7.296	1.511	-10.128	-4.386
Litter_Size_log	**1.432**	**0.433**	**0.590**	**2.222**
Relatedness	**-0.020**	**0.004**	**-0.028**	**-0.013**
Weight_Difference_Low	**-2.583**	**0.911**	**-4.314**	**-0.947**
Weight_g_log	**0.349**	**0.136**	**0.110**	**0.613**
Lineage	0.114	0.567	-0.965	1.166
Temperature_Difference	-0.009	0.110	-0.207	0.200
Weight_Difference_Medium	-0.573	0.470	-1.390	0.348

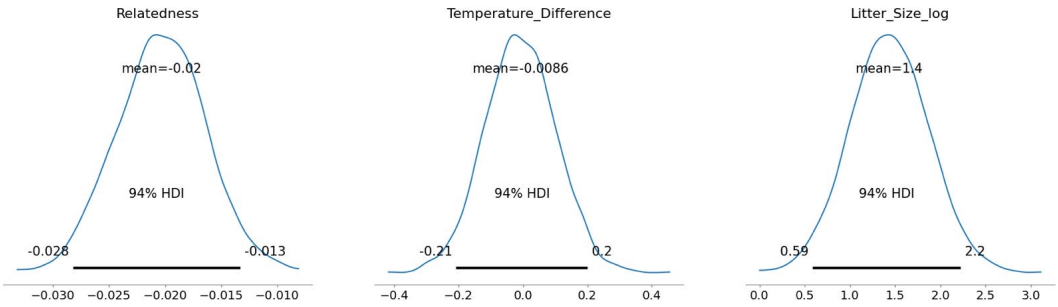

Figura 5: Distribuciones posteriores para los predictores.

Las variables con un efecto creíble sobre el riesgo de HSE son: Litter Size log, Relatedness, Weight Difference Low y Weight g log. Por otro lado, no se encontró evidencia de un efecto para Lineage, Temperature Difference y Weight Difference Medium , ya que sus intervalos de credibilidad incluyen el cero. La Figura 5 visualiza estas distribuciones. Además, notemos que las distribuciones para *Litter_Size_log* y *Relatedness* no se superponen con el cero.

Por lo que la regresión logística Bayesiana demostró ser una herramienta muy superior para este problema. No solo manejó el desbalance de datos, sino que cuantifico la incertidumbre. Los resultados permiten afirmar con un alto grado de confianza que factores como la camada o un mayor peso corporal y un parentesco genético más cercano aumentan el riesgo de un evento de cambio de huésped del virus de la rabia. También podemos concluir que no hay evidencia suficiente para afirmar que el linaje del virus o la diferencia de temperatura jueguen un papel significativo.

Referencias

Balderas Torres, A., Lazaro Vargas, P., and Paavola, J. (2020). The systemic and governmental agendas in presidential attention to climate change in mexico 1994–2018. *Nature Communications*, 11:455.

Boutelle, C., Mollentze, N., Gigante, C., et al. (2025). A logistic regression model to predict the next rabies virus host-shift event. *Scientific Reports*, 15:19306.

Balderas Torres et al. (2020) Boutelle et al. (2025)