Estudio para determinar los principales factores que influyen sobre el salario promedio de los científicos de datos

Jordy Alfaro Brenes, Priscilla Angulo Chaves, Dylan Benavides Castillo, Michelle Gutiérrez Muñoz[1]

jordy.alfarobrenes@ucr.ac.cr , priscilla.angulo@ucr.ac.cr,

dylan.benavides@ucr.ac.cr, michelle.gutierrezmunoz@ucr.ac.cr

RESUMEN

En la actualidad, el análisis de datos ha tenido un gran auge, llevando a muchas personas a dedicarse profesionalmente en esta área. Este artículo se centra en analizar los salarios de los científicos de datos en función de diversas variables demográficas y profesionales. El objetivo es identificar patrones y relaciones significativas entre el salario y factores como el nivel de experiencia, el tamaño de la empresa y la residencia del empleado. Se utilizó un conjunto de datos de salarios de científicos de datos para realizar un análisis de regresión lineal múltiple. Se calcularon el error estándar residual (RSE), el coeficiente de determinación (R^2) y el error cuadrático medio (MSE). Además, se realizaron pruebas de hipótesis para evaluar la significancia de los coeficientes de correlación entre el salario y las variables seleccionadas. Los resultados del análisis de regresión lineal múltiple mostraron un RSE de 58260, un R^2 de 0.269 y un MSE de 3386834788. Las pruebas de hipótesis revelaron que las variables de nivel de experiencia, el nombre del puesto, el tamaño de la empresa y la modalidad de trabajo tienen coeficientes de correlación significativamente diferentes de cero al nivel de significancia de 0.05. Estos hallazgos indican que estos factores influyen de manera significativa en los salarios de los científicos de datos. Por otra parte, las variables de tipo de empleo, lugar de residencia y el año de trabajo resultaron ser no significativas con respecto a su influencia en el salario. Los resultados pueden ayudar a empresas y profesionales a entender mejor las dinámicas salariales en esta industria y a ajustar estrategias de contratación y negociación salarial. Futuros estudios podrían ampliar este análisis incluyendo más variables y utilizando métodos estadísticos adicionales para mejorar la precisión de las predicciones salariales.

PALABRAS CLAVE: Ciencia de datos, Regresión lineal múltiple, Pruebas de hipótesis.

INTRODUCCIÓN

En los últimos años, la ciencia de datos se ha establecido como un campo esencial para extraer información valiosa de grandes bases de datos. En esta línea, analizar los salarios de los científicos de datos es un tema de interés creciente y de gran importancia, puesto que los profesionales de esta área tienen un papel fundamental en el manejo e interpretación de datos para la toma de decisiones estratégicas en gran variedad de sectores laborales.

[1] Estudiantes de la Maestría Profesional en Métodos Matemáticos y Aplicaciones, Universidad de Costa Rica.

Investigar específicamente el salario de los científicos de datos es pertinente debido al crecimiento y demanda de competencias en este campo. La digitalización de la información y la mayor disponibilidad de datos hacen que las empresas u otras organizaciones tengan el reto de contratar y mantener personas con perfiles profesionales en análisis de datos. En este punto, entender los elementos que influyen en la remuneración es vital para profesionales, empleadores y también para orientar estrategias de desarrollo académico y profesional.

El presente artículo se enfocó en identificar cuáles son los factores determinantes en los salarios de los científicos de datos a nivel mundial, haciendo uso de una base de datos que dispone de observaciones en el periodo de 2020 hasta 2024. Se examinaron variables como: nivel de experiencia, ubicación, tamaño de la empresa y tipo de empleo para analizar cuáles tienen mayor influencia en la remuneración de esta profesión.

Este proyecto es crucial debido a su impacto en la competitividad empresarial en la era digital. El estudio contribuye al campo de la ciencia de datos con nuevos conocimientos al descubrir patrones y tendencias en la muestra seleccionada. Entender a fondo los factores que influyen en el salario de los científicos de datos también tiene implicaciones sociales y políticas, como orientar programas educativos para atender las necesidades actuales y futuras de diferentes sectores económicos y sociales.

Los objetivos de la investigación fueron examinar los factores más influyentes en los salarios de los científicos de datos y medir la relación entre variables como el nivel de experiencia, la ubicación, el tipo de empleo, entre otras. Inicialmente, se llevó a cabo un análisis exploratorio de la base de datos, seguido de la implementación en R Studio de un modelo de regresión lineal múltiple para medir las relaciones, efectuar pruebas de hipótesis y de esta forma poner a prueba los coeficientes de correlación encontrados, fijando *salario* como la variable de interés.

Se destaca la importancia de una planificación cuidadosa en la selección de métodos estadísticos y la comprensión de los datos en la investigación. La necesidad de un análisis exploratorio y la selección precisa de variables fueron fundamentales para ajustar las expectativas del trabajo y definir con precisión las hipótesis sometidas a prueba, como menciona Dagnino (2014).

Los hallazgos del trabajo muestran la necesidad de un enfoque integral y riguroso de la investigación estadística para captar de forma precisa las diversas dinámicas que influyen en los salarios de los científicos de datos, enfatizando en la importancia de considerar múltiples factores y evitar interpretaciones que simplifiquen erróneamente el problema de investigación.

METODOLOGÍA

Para el desarrollo del presente análisis, se utilizó el conjunto de datos titulado *Latest data science job salaries 2024*, disponible en la plataforma Kaggle (Badole, 2024). Dicha base de datos contenía un total de 14838 individuos con datos de los años 2020 hasta el 2024. La variable a predecir era el salario (salary_in_usd) y las variables independientes fueron el año de trabajo (work_year), el nivel de experiencia (experience_level), el tipo de empleado (employment_type), el nombre del puesto (job_title), el país de residencia del individuo (employee_residence), la locación de la empresa (company_location), el tamaño de la empresa (company_size) y la modalidad de trabajo (remote_ratio).

Primero, se realizó un análisis descriptivo de los datos para entender mejor las características y la distribución de cada variable donde se incluyeron medidas de tendencia central y de dispersión, así como visualizaciones gráficas para identificar posibles patrones o anomalías en los datos.

Se exploraron además las correlaciones entre las variables involucradas. De acuerdo con Lahura (2003), se puede medir la correlación entre dos variables, mediante el coeficiente de correlación poblacional, que se define como:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

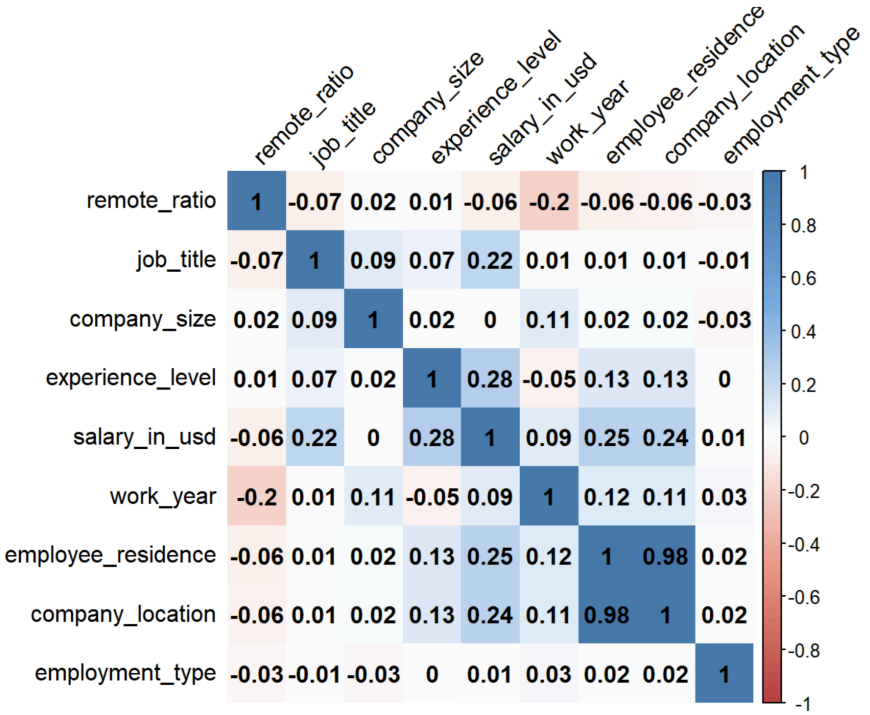
cuyo rango es: $-1 \leq Corr(X,Y) \leq 1$. En esta fórmula se utiliza la varianza y la covarianza, que están dadas por:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

Figura 1

Matriz de correlaciones entre las variables



Cabe destacar que dependiendo del valor del coeficiente de correlación entre las variables X y Y, existen variables aleatorias que tienen correlación positiva siempre que $Corr(X,Y) \geq 0$, correlación negativa si $Cor(X,Y) \leq 0$ y que no tienen correlación si $Cor(X,Y) = 0$, por ejemplo, en la Figura 1 se observa que company_location y employee_residence tienen una alta correlación positiva, mientras que work_year y remote_ratio poseen una correlación negativa.

A partir de este análisis, se determinó que existían elementos representativos en cada variable y otros que más bien eran datos atípicos por lo que se decidió agrupar dentro de las distintas variables los datos con características similares. En el caso de la variable job_title, se conservaron solamente las 5 categorías con más individuos y se clasificaron en ellas las otras categorías que tuvieran más de 100 individuos, el resto se omitió de la tabla, para conservar un total de 13089 filas. Para las variables employee_residence y company_location, los países se clasificaron en seis regiones: Norteamérica, Latinoamérica, Europa, Asia, África y Oceanía.

A continuación, se utilizó un modelo de regresión lineal múltiple para examinar la relación entre los salarios y las variables independientes seleccionadas. Este modelo se eligió debido a su capacidad para cuantificar el efecto individual de cada variable mientras se controla por las otras variables en el modelo tal como plantea Peláez (2016). Para este modelo se consideraron todas las variables mencionadas y también se ejecutó calculando las interacciones dos a dos de las variables.

Llinás (2017) define la recta de regresión lineal de n observaciones mediante la ecuación $y_i = \delta + \beta x_i + \varepsilon_i$ para $i = 1,\ldots,n$ donde X es la variable independiente, Y la variable dependiente y ε_i es el error, por lo que la recta es el conjunto de pares (x_i, y_i) donde $x_i \in X$ y $y_i \in Y$.

Un modelo optimizado de una recta de regresión es el que minimiza el error cuadrático y se conoce como recta de regresión lineal de mínimos cuadrados. Aquí, los parámetros del modelo están dados por:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad , \quad \hat{\delta} = y^{-} - \hat{\beta} x^{-}$$

donde

$$S_{xx} = \sum_{i=1}^{n} \left(x_i - x^{-}\right)^2 y \quad S_{xy} = \sum_{i=1}^{n} \left(x_i - x^{-}\right)\left(y_i - y^{-}\right)$$

Cuando se tienen más de dos variables predictoras, se pueden incluir en el modelo, de forma que se crea un hiperplano de regresión lineal múltiple, que se define como:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \quad para \quad i = 1, \ldots, n$$

donde X_j con $j = 1, \ldots, p-1$ son $p-1$ variables linealmente independientes, Y_i es la variable dependiente y x_{ij} es el i-ésimo elemento de la variable X_j.

En el caso de una muestra en la que se aplica un modelo de regresión lineal, se emplea el coeficiente de correlación muestral, que también se conoce como coeficiente de Pearson y se calcula de la siguiente forma:

$$R = \hat{\beta}\sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$

donde $\hat{\beta}$ es la estimación por mínimos cuadrados del parámetro β de la regresión lineal. En añadidura, R^2 se llama coeficiente de determinación muestral.

Se calculó el error estándar residual (RSE) para medir la precisión de las predicciones del modelo. Además, se utilizó el coeficiente de determinación (R^2) para evaluar la proporción de la variabilidad en los salarios que puede ser explicada por las variables incluidas en el modelo. El error cuadrático medio (MSE) también se calculó como una medida adicional de la precisión del modelo.

La justificación para la selección del modelo de regresión lineal múltiple se basa en su capacidad para manejar múltiples variables explicativas y proporcionar una estimación clara de sus efectos individuales. Este enfoque es adecuado para entender las complejas interacciones entre las diferentes variables que afectan los salarios en el sector tecnológico. Además, se aseguraron los principales supuestos del modelo de regresión lineal, como la linealidad, la independencia de los errores, la homocedasticidad y la normalidad de los errores, a través de pruebas diagnósticas y análisis de residuos.

Posteriormente, se realizaron pruebas de hipótesis para evaluar la significancia estadística de los coeficientes de las variables independientes. Estas pruebas permitieron determinar si las relaciones observadas entre los salarios y las variables seleccionadas eran estadísticamente significativas, utilizando un nivel de significancia del 0.05.

Una prueba de hipótesis es un método inferencial que como menciona Rice (2007) se utiliza para evaluar la validez de una afirmación sobre una población basada en una muestra de datos. Linás (2017) define la hipótesis nula (H_0), como la proposición que se debe comprobar y se puede plantear de diferentes formas y la hipótesis alternativa (H_1) como el complemento de la hipótesis nula y su validez se demuestra al rechazar H_0. En este caso, se definió $H_0: \beta_i = 0$ y $H_1: \beta_i \neq 0$, donde los β_i corresponden a los estimadores del modelo de regresión lineal.

Una vez que se plantea la hipótesis nula, se utilizan los valores que arroja el modelo de regresión para cada variable: el estimador, el error estándar, el t valor y el p valor, donde el p valor es la probabilidad de obtener los mismos resultados al tomar la hipótesis nula como verdadera. De esta forma, si $p < 0.05$ se rechaza la hipótesis nula indicando que el coeficiente es significativamente diferente de cero, bajo el nivel de significancia definido; mientras que si $p > 0.05$ se acepta la hipótesis nula.

Todos estos análisis estadísticos se llevaron a cabo mediante el lenguaje de programación R (R Core Team, 2024), utilizando la versión 4.4.0. y los paquetes stats (R Core Team, 2024) y ggplot2 (Wickham, 2023).

RESULTADOS

A partir del análisis descriptivo de los datos se encontró que existen correlaciones negativas débiles y cercanas a cero para varias variables, se destaca una mayor correlación para la variable de nivel de experiencia, seguido de lugar de residencia y ubicación de la compañía.

Al realizar el modelo de regresión lineal múltiple se obtuvo un valor de 0.27 para el coeficiente de determinación (R^2), esto quiere decir que el modelo puede explicar la variabilidad del salario en un 27%. Se obtuvo además un error residual estándar (RSE) de 58260 y un valor de 3386834788 para el MSE. El RSE es cercano al valor de desviación estándar, la cual corresponde a 68068.18, como es esperado. Por otro lado, un valor tan alto de MSE concuerda con que el modelo no puede explicar en su mayoría la variabilidad del salario, tal como lo indica el coeficiente de determinación.

Con el fin de mejorar el resultado anterior, se ejecutó una prueba del modelo en donde se ajustaron los salarios de cada año de acuerdo con la inflación. Sin embargo, este cambio no generó un aumento significativo del coeficiente de determinación, por el contrario, lo bajó en menos de un 1%, por esta razón no se tomó en cuenta este ajuste en la tabla para el análisis final.

En el Cuadro 1 se observan los coeficientes obtenidos con el modelo de regresión lineal. En la cuarta columna se observan los p valores utilizados en las pruebas de hipótesis. A partir de este análisis se demostró que las variables de nivel de experiencia, el nombre del puesto, el tamaño de la empresa y la modalidad de trabajo tienen un impacto significativo en los salarios. Por otra parte, las variables de tipo de empleo, lugar de residencia y el año de trabajo resultaron ser no significativas con respecto a su influencia en el salario.

La relación positiva entre el nivel de experiencia y el salario indicaría que a medida que los trabajadores adquieren más años de experiencia, sus salarios tienden a aumentar. Este hallazgo es consistente con las expectativas del mercado laboral, donde según lo planteado por Landon-Murray (2016), la experiencia adicional suele traducirse en habilidades más avanzadas y una mayor productividad, justificando salarios más altos.

El nombre del puesto también tendría un rol significativo de acuerdo con la prueba de hipótesis, en el cuál, dependiendo del puesto en específico se podría esperar que tenga un impacto en el salario. Esto quiere decir que dependiendo el cargo en específico el valor que el mercado le asigna es diferente aún si se encuentran dentro del espectro de Ciencia de Datos.

Otro factor que mostró una influencia significativa en los salarios es el tamaño de la empresa, específicamente la categoría de mayor tamaño. Según reportes de la página Glassdoor (2023), las empresas de mayor tamaño tienden a ofrecer salarios más altos en comparación con las empresas más pequeñas.

Este resultado puede atribuirse a varios factores, incluyendo mayores recursos financieros, políticas de compensación más competitivas y mejores oportunidades de crecimiento profesional en las empresas más grandes.

Cuadro 1

Resumen de coeficientes del modelo de regresión lineal

Coeficientes	Estimado	Error estándar	Valor t	Pr(> |t|)	Significancia
(Intercept)	-774.4	23413.8	-0.033	0.97362	
work_year2021	-19402.1	9777.3	-1.985	0.04715	*
work_year2022	-15069.7	8890.5	-1.695	0.09019	.
work_year2023	-4930.3	8043.8	-0.613	0.54011	
work_year2024	-14127.1	8848.6	-1.596	0.11066	
experience_level2	21531.1	2281.3	9.436	< 2e-16	***
experience_level3	15389.5	2231.9	6.897	6.07e-12	***
experience_level4	18409.7	2997.0	6.141	9.14e-10	***
employment_type2	19709.5	5863.7	3.361	0.00078	***
employment_type3	12074.1	1680.4	7.188	7.11e-13	***
employment_type4	32784.1	4110.8	7.975	1.66e-15	***
job_title3	10412.5	2148.2	4.847	1.31e-06	***
job_title4	40834.9	3543.1	11.529	< 2e-16	***
employee_residence2	6944.1	3642.8	1.906	0.05665	.
employee_residence3	72824.4	4374.0	16.646	< 2e-16	***

employee_residence4	11682.6	4325.1	2.702	0.00691	**
employee_residence5	34527.4	5754.0	6.001	2.02e-09	***
employee_residence6	64148.3	4514.2	14.207	< 2e-16	***
remote_ratio.L	1178.7	1100.4	1.071	0.28428	
remote_ratio.Q	-453.9	1781.9	-0.255	0.79860	
company_location2	39058.6	11297.6	3.458	0.00055	***
company_location3	8636.6	4497.2	1.920	0.05480	.
company_location4	12368.9	3310.5	3.737	0.00019	***
company_location5	41120.1	4261.3	9.652	< 2e-16	***
company_location6	12946.5	4215.0	3.071	0.00214	**
company_size2	17398.6	4407.7	3.948	8.08e-05	***
company_size3	49439.6	6350.9	7.785	6.18e-15	***

Nota: La significancia está dada por: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

La modalidad de trabajo, específicamente el trabajo remoto, también se identificó como un factor significativo que afecta los salarios, específicamente la categoría intermedia de trabajo remoto, es decir, cuando se maneja aproximadamente el 50% del tiempo en trabajo remoto y el otro 50% de forma presencial. Los empleados que trabajan de forma remota en categoría intermedia tienden a recibir salarios más altos que sus contras partes que trabajan de manera presencial. El teletrabajo podría derivar en una reducción en los costos operativos para las empresas que permiten el trabajo remoto, lo que les permite redistribuir esos ahorros en forma de salarios más altos.

CONCLUSIONES

El objetivo principal de la investigación fue encontrar los factores más importantes para determinar el salario de un científico de datos. A partir de la base de datos elegida, la cual contenía diferentes variables pertenecientes a profesionales en la ciencia de datos, se esperaba obtener resultados que indicaran una correlación fuerte de una o más variables con la variable de interés

del salario. Se llevó a cabo un análisis descriptivo de los datos, un modelo de regresión múltiple y pruebas de hipótesis con base a los coeficientes beta obtenidos del mismo.

A partir del coeficiente de determinación obtenido para el modelo de regresión lineal múltiple se concluye que el modelo puede explicar la variabilidad del salario en un 27%, lo cual indica que deben existir otras variables no contempladas que influyen en la predicción, como interacciones de variables dos a dos o incluso mayor cantidad de interacciones, sin embargo, no fue viable para el análisis tomar en cuenta estos casos. Así mismo, se considera también la conclusión de que posiblemente un modelo de regresión lineal no es el óptimo para realizar este análisis en donde entran en juego tantas variables.

Con las pruebas de hipótesis realizadas a partir de los coeficientes obtenidos del modelo de regresión lineal se concluye que, para las variables de nivel de experiencia, tamaño de la empresa (categoría más grande), nombre del puesto y modalidad de trabajo remoto intermedio, no hay suficiente evidencia con un nivel de significancia del 5% para aceptar la hipótesis nula, la cual dice que los coeficientes del modelo son iguales a 0, y por lo tanto no significativos. Es decir, se puede decir que estos coeficientes obtenidos en el modelo de regresión lineal reflejan una relación entre la variable dependiente y la independiente, con excepción de las variables tipo de empleo, año de trabajo y lugar de residencia, para las cuales se concluye que si hay suficiente evidencia con un nivel de significancia del 5% para rechazar la hipótesis nula.

Entre los problemas encontrados, se encuentran las modificaciones necesarias que fueron hechas a la base de datos para poder tener variables que fueran continuas en su totalidad, y no categóricas. Esto incluye, además, factorizar algunas variables, que, si bien se representan con un número, en realidad son categóricas, porque solo hay una cantidad determinada de posibilidades, por ejemplo, la variable de trabajo remoto.

Otro desafío fue tratar de mejorar el modelo para lo cual se creó uno con interacciones dos a dos entre todas las variables independientes, sin embargo, se volvía muy extenso y en muchos casos se obtenían datos de la forma NA, lo cual volvía más complejo el análisis. Debido a esto se seleccionaron algunas variables específicas de interés para las interacciones, sin embargo, esto no aumentó el valor del coeficiente de determinación.

Como trabajo futuro se puede llevar a cabo un análisis que indique si las variables son linealmente separables o no y complementar con un análisis en componentes principales para estudiar la dispersión de los datos para analizar los tipos de correlación entre las variables involucradas (utilizando un círculo de correlaciones), con el fin de analizar el poder de discriminación de las variables.

AGRADECIMIENTOS

Agradecemos al profesor Maikol Solís Chacón por su acompañamiento y buena disposición a lo largo de esta investigación. También agradecemos a los compañeros del curso Estadística Actuarial I de la Escuela de Matemática de la Universidad de Costa Rica por sus valiosas discusiones y sugerencias durante el desarrollo de este proyecto.

BIBLIOGRAFÍA

Badole, S. (2024). *Latest data science job salaries 2024* [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/saurabhbadole/latest-data-science-job-salaries-2024

Dagnino, J. S. (2014). Inferencia estadística: Pruebas de hipótesis. *Revista Chilena de Anestesiología, 43*(2), 125–128.

Glassdoor. (2023). Reporte anual de salarios de científicos de datos en diferentes ciudades. https://www.glassdoor.com.mx/Sueldos/san-jose-costa-rica-data-scientist-sueldo-SRCH_IL.0,19_IM955_KO20,34.htm

Lahura, E. (2003). El coeficiente de correlación y correlaciones espúreas (*Documentos de Trabajo,* Vol. 218). Pontificia Universidad Católica del Perú, Departamento de Economía.

Landon-Murray, M. (2016). Big data and intelligence: Applications, human capital, and education. *Journal of Strategic Security, 9*(2), 92–121. http://dx.doi.org/10.5038/1944-0472.9.2.1514

Llinás, H. (2017). Estadística inferencial. Universidad del Norte.

Peláez, I. M. (2016). Modelos de regresión: Lineal simple y regresión logística. *Revista Seden*, *14*, 195–214.

R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.0) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Rice, J. A. (2007). *Mathematical statistics and data analysis (3rd ed.).* Thomson/Brooks/Cole.

Wickham, H. (2023). *ggplot2: Create elegant data visualisations using the grammar of graphics* (Versión 3.4.4) [Paquete R]. CRAN. https://CRAN.R-project.org/package=ggplot2