Taller 1: Análisis Predictivo y Aprendizaje Estadístico Caso de Estudio: Advertising Data

Jeferson Camio Runza Olaya

Paula Andrea Blando Ramirez

Karol Dayana Bedoya Olivar

Universidad ECCI

Departamento de ingeniería

ingeniería Industrial

Sistemas Avanzados de Producción

Tabla de Contenido

Introducción 2

Fase 1: Estadística Descriptiva y Análisis Exploratorio 4

1. Caracterización Numérica: 4
2. Análisis de Distribución y Atípicos: 4
3. Evaluación de Asociación: 6
4. Interpretación de Resultados 7

Fase 2: Regresión Lineal y Diagnostico 9

1. Modelamiento Múltiple: 9
2. Interpretación de parámetros: 10
3. Análisis de Bondad de Ajuste: 11
4. Estimación Matricial: 12

Conclusiones 15

Tabla de Ilustraciones

Ilustración 1. Histogramas y diagramas de caja de las variables de inversión publicitaria y ventas 5
Ilustración 2. Matriz de Correlación de Pearson y de spearman. 7
Ilustración 3 Diagramas de dispersión entre inversión publicitaria y ventas 10

Tabla de Tablas

Tabla 1. Estadísticas descriptivas de la inversión publicitaria y las ventas 4
Tabla 2. Interpretación del sesgo (asimetría) de las variables del modelo 6

Introducción

En el entorno empresarial actual, la toma de decisiones estratégicas debe estar respaldada por evidencia cuantitativa que permita optimizar la asignación de recursos. Una de las preguntas más relevantes en el ámbito del marketing es determinar en qué medida la inversión publicitaria influye en el comportamiento de las ventas y cuál de los canales disponibles genera mayor impacto sobre el desempeño comercial.

El presente estudio analiza el conjunto de datos Advertising, el cual contiene información de 200 mercados diferentes y registra la inversión en tres medios publicitarios: televisión (TV), radio y prensa (Newspaper), así como el volumen de ventas alcanzado. A partir de este conjunto de datos, se busca evaluar la relación existente entre la inversión publicitaria y el retorno en ventas, identificando la magnitud y dirección de dichas asociaciones.

El objetivo principal del análisis es determinar qué medio publicitario presenta mayor capacidad explicativa sobre las ventas y evaluar si un modelo de regresión lineal múltiple mejora significativamente el ajuste frente a modelos simples. Asimismo, se pretende comparar este enfoque paramétrico con métodos alternativos, como los árboles de decisión, con el fin de identificar cuál modelo ofrece mayor robustez e interpretabilidad para la toma de decisiones estratégicas.

De esta manera, el estudio no solo aborda el análisis estadístico de los datos, sino que también traduce los resultados obtenidos en implicaciones prácticas para la optimización de la inversión publicitaria.

Fase 1: Estadística Descriptiva y Análisis Exploratorio

1. **Caracterización Numérica:** Elabore una tabla que resuma las estadísticas descriptivas para las variables TV, Radio, Newspaper y Sales. Esta tabla debe reportar:

Variable	*Media*	*Mediana*	*Desv. Estándar*	*Mínimo*	*Máximo*	*Asimetría*	*Kurtosis*
TV	147.04	149.75	85.85	0.7	296.4	-0.070	-1.226
Radio	23.26	22.90	14.85	0.0	49.6	0.094	-1.260
Newspaper	30.55	25.75	21.78	0.3	114.0	0.895	0.650
Sales	14.02	12.90	5.22	1.6	27.0	0.408	-0.409

Tabla 1. Estadísticas descriptivas de la inversión publicitaria y las ventas

Con la anterior tabla se puede visualizar que la inversión promedio más alta se concentra en TV (147.04), además de presentar la mayor variabilidad (DE = 85.85), lo que indica diferencias significativas entre mercados. Radio muestra una distribución bastante simétrica y estable, mientras que Newspaper presenta mayor asimetría positiva (0.895), lo que sugiere presencia de valores altos extremos.

En cuanto a Sales, la media (14.02) es ligeramente mayor que la mediana (12.90), evidenciando una leve asimetría positiva y algunos mercados con ventas superiores al promedio.

La curtosis negativa en TV, Radio y Sales indica distribuciones relativamente planas, mientras que Newspaper presenta colas más pesadas, lo que podría generar mayor inestabilidad en modelos predictivos.

En síntesis, TV domina en magnitud y variabilidad, Newspaper muestra mayor irregularidad y las ventas presentan ligera concentración hacia valores altos.

2. **Análisis de Distribución y Atípicos:** Construya un histograma y un diagrama de caja (Box Plot) para cada variable:

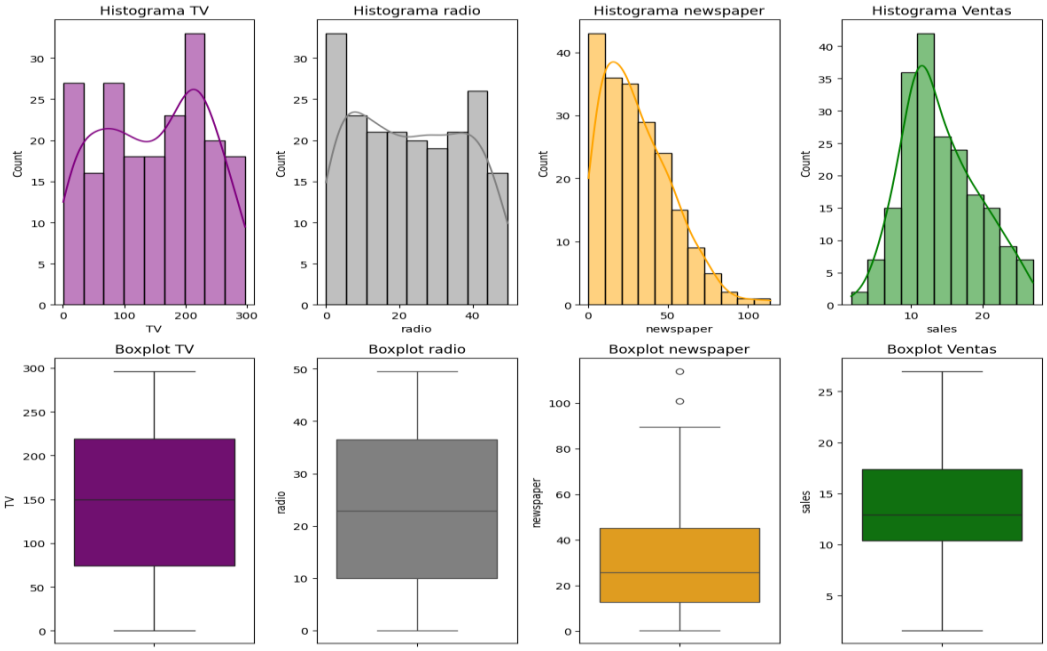


Ilustración 1. Histogramas y diagramas de caja de las variables de inversión publicitaria y ventas

Variable	***Valor (Sesgo)***	***Interpretación del Sesgo***	***¿Qué significa en la práctica?***
TV	-0.070	Negativo ligero (Casi simétrica)	La distribución es casi perfecta. Existe una levísima tendencia a tener más valores altos, pero los datos están muy bien repartidos en todo el rango.
Radio	0.094	Positivo ligero (Casi simétrica)	Al igual que la TV, su sesgo es despreciable. La inversión en radio es constante y no presenta valores extremos que deformen la campana significativamente.
Newspaper	0.895	Positivo moderado (Sesgo a la derecha)	Es la variable más sesgada. Significa que la mayoría de las inversiones son bajas, pero hay unos pocos casos con inversiones muy altas que "jalar" la media hacia arriba.

Sales	0.408	Positivo leve (Sesgo a la derecha)	Hay una ligera concentración de ventas en los valores bajos/medios, con algunas ventas excepcionalmente altas que crean una cola hacia la derecha.

Tabla 2. Interpretación del sesgo (asimetría) de las variables del modelo

¿Se identifican datos atípicos o observaciones que se alejan significativamente de la masa de los datos?

Sí. En la variable Newspaper se observan valores atípicos en la parte superior del boxplot, lo que indica que en algunos mercados la inversión en prensa es mucho mayor que en la mayoría de los casos.

En las variables TV, Radio y Sales no se evidencian valores extremos muy marcados, lo que sugiere un comportamiento más estable y homogéneo.Esto muestra que la inversión en prensa es la más irregular entre los tres medios analizados.

¿Cómo podrían influir estas observaciones en la trayectoria de una línea de regresión?

Los valores atípicos pueden afectar la forma en que se ajusta la recta de regresión. Si un punto está muy alejado del comportamiento general, puede "jalar" la línea hacia él y modificar la pendiente estimada. En este caso, los valores altos de Newspaper podrían hacer que el modelo sobreestime o subestime su impacto real sobre las ventas, afectando la precisión del análisis.

Por eso es importante identificarlos antes de construir el modelo, ya que pueden influir en la interpretación final de los resultados.

3. **Evaluación de Asociación:** Calcule y presente las matrices de correlación de Pearson y Spearman.

    - El coeficiente de Pearson evaluará la fuerza de la asociación lineal.
    - El coeficiente de Spearman permitirá identificar relaciones monótonas basadas en el rango de los datos

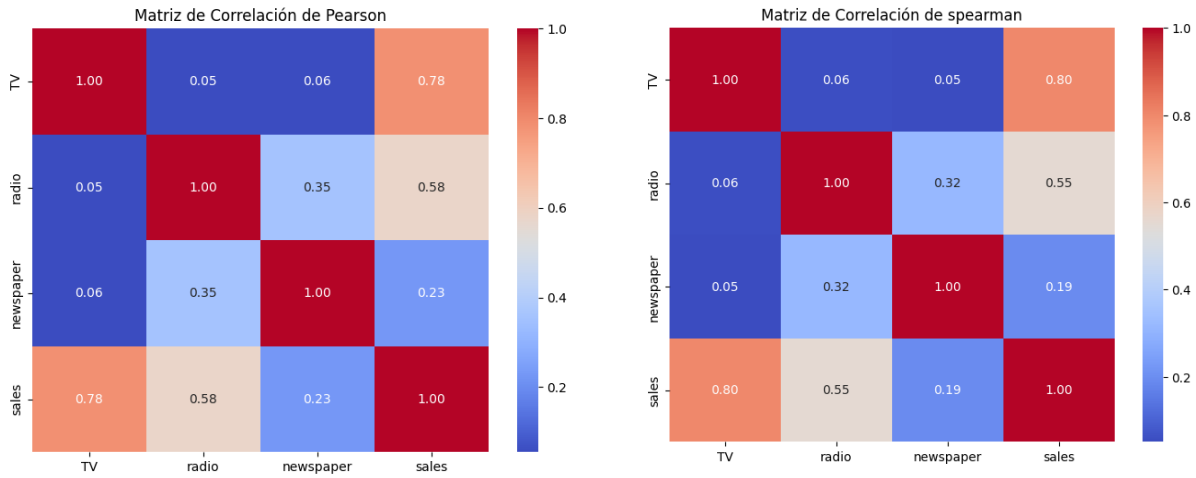


Ilustración 2. Matriz de Correlación de Pearson y de spearman.

4. Interpretación de Resultados

¿Qué significa un coeficiente cercano a 1 o −1 en el contexto de la inversión publicitaria y el retorno en ventas?

- Un coeficiente cercano a 1 indica una relación positiva fuerte entre la inversión publicitaria y las ventas. Es decir, a medida que aumenta la inversión en un medio, las ventas tienden a aumentar de manera consistente.
- Por el contrario, un coeficiente cercano a −1 indicaría una relación negativa fuerte, lo que significaría que al aumentar la inversión, las ventas disminuyen, situación que en este contexto no tendría sentido estratégico.
- En los resultados observados, la relación más fuerte se presenta entre TV y Sales (≈ 0.78–0.80), lo que sugiere que la inversión en televisión tiene una asociación positiva considerable con el volumen de ventas.

¿Cómo se interpreta un valor de correlación cercano a 0?

Un valor cercano a 0 indica que no existe una relación lineal clara entre las variables. En este caso, la correlación entre Newspaper y Sales (≈ 0.19–0.23) es baja, lo que sugiere que la inversión en prensa no tiene una asociación lineal significativa con el incremento en ventas. Esto podría indicar que este canal tiene menor impacto o que su efecto no es directo ni proporcional.

Comparación entre Pearson y Spearman

¿Las relaciones son estrictamente lineales?

Los coeficientes de Pearson y Spearman son muy similares en todas las variables. Por ejemplo:

- TV–Sales: 0.78 (Pearson) vs 0.80 (Spearman)
- Radio–Sales: 0.58 vs 0.55
- Newspaper–Sales: 0.23 vs 0.19

Cuando ambos coeficientes son cercanos, significa que la relación no solo es monótona, sino también aproximadamente lineal.

Además, al observar las nubes de puntos:

- TV vs Sales muestra una tendencia lineal positiva clara.
- Radio vs Sales presenta relación positiva moderada.
- Newspaper vs Sales muestra dispersión amplia y poca estructura lineal.

En conclusión, los resultados sugieren que las relaciones con TV y Radio pueden modelarse adecuadamente con un enfoque lineal, mientras que Newspaper no presenta una relación lineal fuerte y podría no ser un predictor relevante en un modelo simple.

Fase 2: Regresión Lineal y Diagnostico

1. Modelamiento Múltiple:

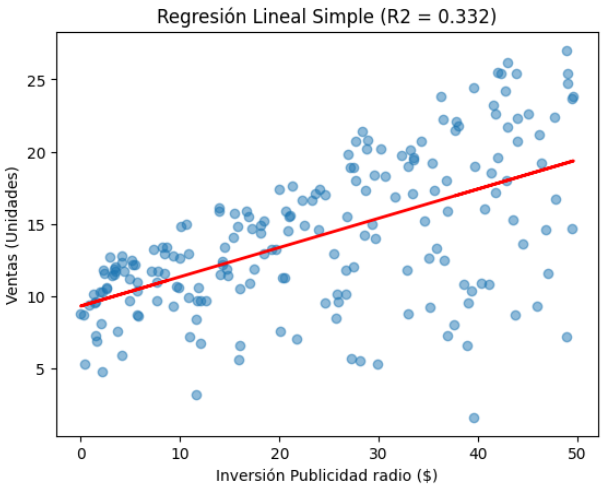
El modelo publicitario está traccionado principalmente por la TV, cuya relación lineal con las ventas (r = 0.78$) es altamente confiable debido a su distribución simétrica; en contraste, la inversión en Newspaper presenta ineficiencias estadísticas, mostrando un sesgo pronunciado hacia valores altos que no logran impactar significativamente en el retorno de ventas (r = 0.23$).

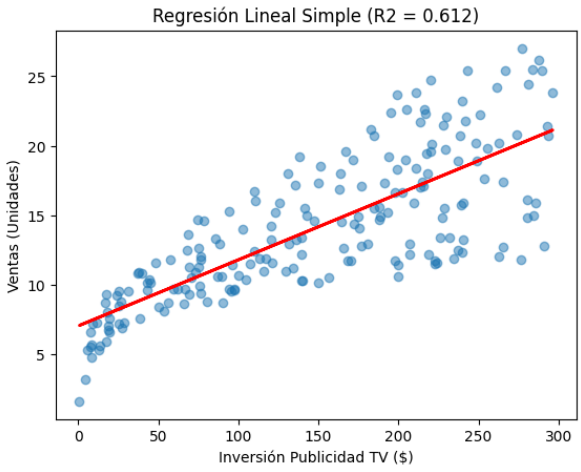
Coeficientes del modelo

- Intercepto (beta_0): 2.9389
- Pendiente (beta_1): 0.0458
- Pendiente (beta_2): 0.1885
- Pendiente (beta_3): -0.0010
- Intercepto (beta0 = 2.9389$)

Es el valor esperado de las ventas si las inversiones en TV, Radio y Periódico fueran cero.

- Interpretación: En un escenario sin publicidad, se estima que las ventas base serían de aproximadamente 2.94 unidades.





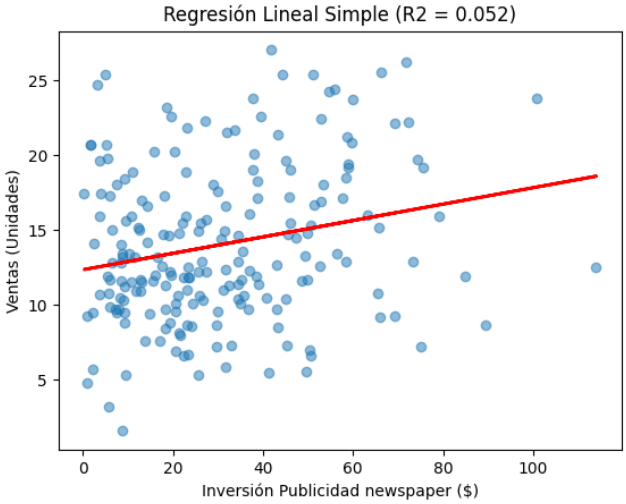


Ilustración 3 Diagramas de dispersión entre inversión publicitaria y ventas

2. Interpretación de parámetros:

Interpretación del intercepto ($\hat{\beta}_0 = 2.9389$)

El intercepto representa el nivel esperado de ventas cuando la inversión en TV, Radio y Newspaper es igual a cero. En este caso, el modelo estima que, sin inversión publicitaria, las ventas serían aproximadamente 2.94 mil unidades. Desde el punto de vista conceptual, sí tiene sentido que sea diferente de cero, ya que el producto podría generar ventas mínimas por reconocimiento de marca, clientes recurrentes o demanda orgánica, incluso sin publicidad activa.

Interpretación de los coeficientes (ceteris paribus)

Los coeficientes indican cuánto cambia el valor promedio de las ventas ante un aumento unitario en cada variable, manteniendo las demás constantes.

$\hat{\beta}_1 = 0.0458$ (TV)

Por cada incremento de 1 unidad monetaria en inversión en TV (mil dólares), las ventas aumentan en promedio 0.0458 mil unidades, manteniendo constante Radio y Newspaper. Esto confirma una relación positiva entre televisión y ventas.

$\hat{\beta}_2 = 0.1885$ (Radio)

Por cada incremento de 1 unidad monetaria en Radio, las ventas aumentan en promedio 0.1885 mil unidades, manteniendo constantes las demás variables. Este coeficiente es mayor que el de TV, lo que sugiere que, en términos marginales, la radio tiene un mayor impacto por unidad invertida.

$\beta_3 = -0.0010$ (Newspaper)

Este coeficiente es prácticamente cero y ligeramente negativo. Indica que, manteniendo constantes TV y Radio, un aumento en Newspaper no genera un impacto positivo relevante en ventas. Su efecto es casi nulo dentro del modelo.

3. **Análisis de Bondad de Ajuste:**

```
# Supongamos que df ya contiene los datos cargados de Advertising.csv
X = df[['TV', 'radio','newspaper']]
y = df['sales']

# Ajustar el modelo de Regresión Múltiple
model_multiple = LinearRegression()
model_multiple.fit(X, y)

# 3. Obtener e imprimir resultados
intercepto = model_multiple.intercept_
coeficientes = model_multiple.coef_
r_cuadrado = model_multiple.score(X, y)

print(f"--- Coeficientes del Modelo ---")
print(f"Intercepto (beta_0): {intercepto:.4f}")
print(f"Coeficiente TV (beta_1): {coeficientes[0]:.4f}")
print(f"Coeficiente Radio (beta_2): {coeficientes[1]:.4f}")
print(f"Coeficiente Newspaper (beta_3): {coeficientes[2]:.4f}")
print(f"\n--- Bondad de Ajuste ---")
print(f"R-cuadrado (R2): {r_cuadrado:.4f}")
```

```
--- Coeficientes del Modelo ---
Intercepto (beta_0): 2.9389
Coeficiente TV (beta_1): 0.0458
Coeficiente Radio (beta_2): 0.1885
Coeficiente Newspaper (beta_3): -0.0010
```

El modelo de regresión lineal múltiple presenta un $R^2 = 0.8972$.

Esto significa que el 89.72% de la variabilidad total de las ventas es explicada conjuntamente por las variables TV, Radio y Newspaper. Es un nivel de explicación bastante alto, lo que indica que el modelo tiene una capacidad predictiva fuerte.

En la Fase 1 se observaron los siguientes R^2 aproximados:

- TV → $R^2 \approx 0.612$
- Radio → $R^2 \approx 0.332$
- Newspaper → $R^2 \approx 0.052$

Comparando estos resultados con el modelo múltiple (0.8972), se evidencia una mejora sustancial en el ajuste al incluir simultáneamente las variables Radio y Newspaper junto con TV.

El incremento de 0.612 (modelo solo con TV) a 0.8972 (modelo múltiple) indica que la combinación de medios permite explicar mucho mejor el comportamiento de las ventas que cualquier variable por separado.

4. **Estimación Matricial:**

Dada las matrices de diseño X y el vector de respuesta Y:

$$X = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}, \quad Y = \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix}$$

Calcule la matriz transpuesta X′.

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}$$

Calcule el producto X′X y su respectiva matriz inversa (X′X) −1

$$X'X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{pmatrix}$$

Producto X’X

$$X'X = \begin{pmatrix} 5 & 15 \\ 15 & 55 \end{pmatrix}$$

Inversa (X′X) −1

$$A^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Determinante:

$$(5)(55) - (15)(15) = 275 - 225 = 50$$

$$(X'X)^{-1} = \frac{1}{50} \begin{pmatrix} 55 & -15 \\ -15 & 5 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix}$$

Calcule el producto X′Y

$$X'Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix} \begin{pmatrix} 2 \\ 4 \\ 6 \\ 7 \\ 9 \end{pmatrix}$$

$$2 + 4 + 6 + 7 + 9 = 28$$

$$1(2) + 2(4) + 3(6) + 4(7) + 5(9)$$
$$= 2 + 8 + 18 + 28 + 45 = 101$$

$$X'Y = \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

Obtenga el vector de parámetros βˆ y escriba la ecuación de la recta estimada ˆy = βˆ 0+βˆ 1x

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{\beta} = \begin{pmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{pmatrix} \begin{pmatrix} 28 \\ 101 \end{pmatrix}$$

β_0:

$$1.1(28) - 0.3(101)$$

$$30.8 - 30.3 = 0.5$$

β_1:

$$-0.3(28) + 0.1(101)$$

$$-8.4 + 10.1 = 1.7$$

$$\hat{\beta} = \begin{pmatrix} 0.5 \\ 1.7 \end{pmatrix}$$

Ecuación estimada:

$$\boxed{\hat{y} = 0.5 + 1.7x}$$

Conclusiones

El análisis exploratorio permitió identificar que la inversión en televisión es el canal con mayor asociación positiva respecto a las ventas, seguido por la radio, mientras que la inversión en prensa escrita presenta una relación débil y mayor dispersión en sus datos. Las correlaciones obtenidas evidenciaron que TV y Radio mantienen relaciones lineales claras con las ventas, mientras que Newspaper muestra una influencia limitada.

El modelo de regresión lineal múltiple presentó un coeficiente de determinación $R^2 = 0.8972$, lo que indica que aproximadamente el 89.72% de la variabilidad de las ventas es explicada por la combinación de los tres medios publicitarios. Este resultado demuestra una mejora significativa frente a los modelos simples, confirmando que el efecto conjunto de los canales proporciona una explicación más robusta del comportamiento comercial.

En términos de impacto marginal, la radio mostró el mayor efecto por unidad invertida, seguida de la televisión, mientras que la prensa escrita presentó un efecto prácticamente nulo dentro del modelo. Esto sugiere que, desde una perspectiva estratégica, la asignación eficiente del presupuesto debería priorizar televisión y radio.

Finalmente, el análisis confirma que los modelos lineales son adecuados para describir la relación entre inversión publicitaria y ventas en este conjunto de datos, dado el comportamiento aproximadamente lineal observado en las nubes de puntos y la similitud entre los coeficientes de Pearson y Spearman. En consecuencia, la regresión lineal múltiple se consolida como un modelo sólido, interpretable y útil para apoyar decisiones de inversión en marketing.