UNINPAHU

MACHINE-LEARNING-I

Autores:

Angie López

Marlon Peña

Marlon Albarracín

Bogotá D.C.

11 de junio de 2025

Análisis Profundo de Regresión Lineal Simple: Horas de Trabajo vs. Unidades Producidas

1. Introducción

En el presente estudio se analiza la relación entre el tiempo dedicado al trabajo (en horas) y la cantidad de unidades producidas. La hipótesis subyacente es que a mayor número de horas trabajadas, mayor será la producción. Este tipo de análisis es fundamental para áreas como administración de operaciones, productividad laboral y gestión del rendimiento en fábricas, talleres u oficinas. Para ello, se emplea un modelo de regresión lineal simple, que permite cuantificar y predecir el comportamiento de la variable dependiente (producción) a partir de la variable independiente (horas trabajadas).

El análisis se ha realizado utilizando tres enfoques diferentes para estimar el modelo:

- Gradiente descendente (método iterativo de optimización).
- `scikit-learn` (biblioteca estándar de machine learning en Python).
- `statsmodels` (enfoque estadístico clásico con pruebas de hipótesis).

Además, se validan los supuestos clásicos del modelo lineal (normalidad, homocedasticidad e independencia de los errores) para asegurar la validez del modelo desde el punto de vista inferencial.

2. Exploración de los Datos

El conjunto de datos contiene dos variables numéricas:

- **Trabajo (x)**: horas de dedicación laboral.
- **Producción (y)**: número de unidades fabricadas en ese tiempo.

El primer paso fue realizar una inspección visual a través de un diagrama de dispersión, el cual mostró una relación positiva aproximadamente lineal. Esto sugiere que puede ser apropiado aplicar un modelo lineal simple.

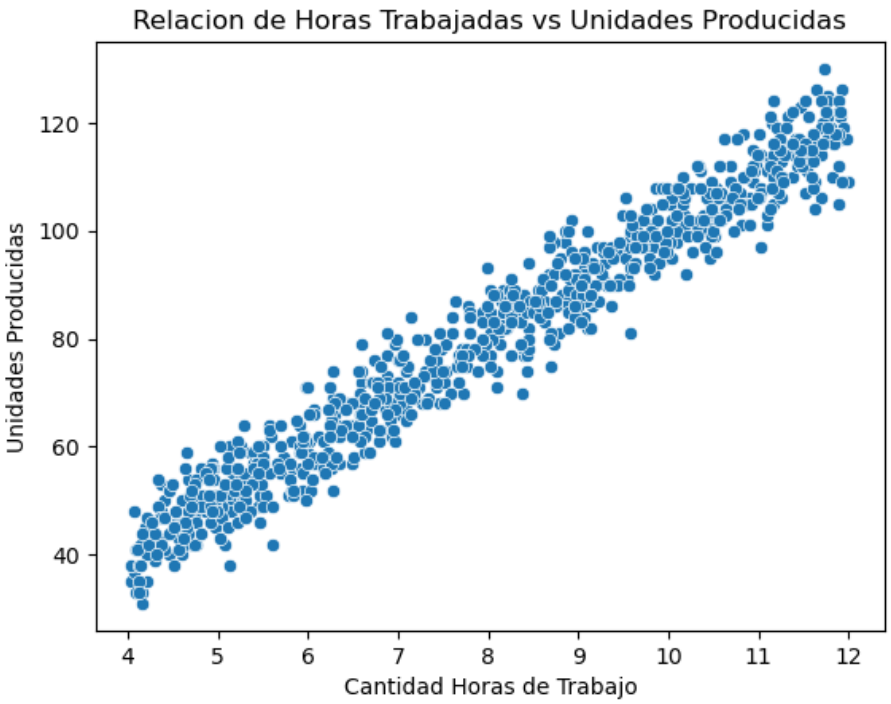

Figura 1: Relación entre Horas de Trabajo y Producción

3. Modelo de Regresión Lineal Simple

Se busca ajustar un modelo de la forma:

$$y = \theta_0 + \theta_1 x + \varepsilon$$

donde:

- y: unidades producidas (variable dependiente).

- x: horas trabajadas (variable independiente).
- θ_0: intercepto (producción cuando el tiempo de trabajo es cero).
- θ_1: pendiente (incremento de la producción por cada hora extra trabajada).
- ε: término de error aleatorio.

3.1. Estimación por Gradiente Descendente

El gradiente descendente es un algoritmo de optimización numérica que permite encontrar los valores óptimos de los parámetros del modelo al minimizar la función de coste, típicamente el error cuadrático medio (ECM).

Tras múltiples iteraciones, los valores convergieron a:

$$\hat{\theta}_0 = 3{,}9177, \quad \hat{\theta}_1 = 17{,}142$$

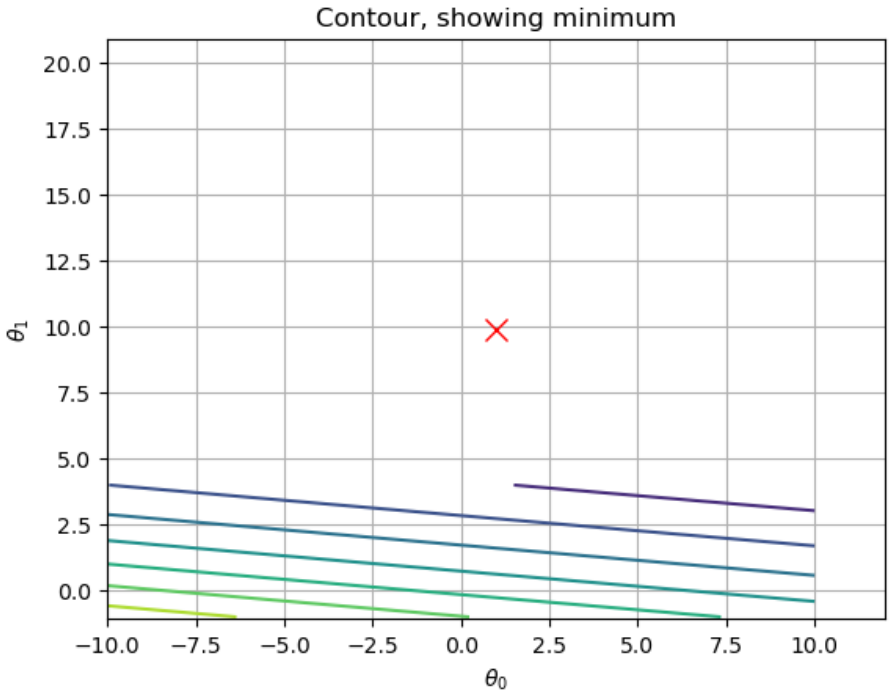

Figura 2: Curvas de nivel de la función de coste en gradiente descendente

Esto indica que incluso sin trabajar, se estima que se producen aproximadamente 3.92 unidades. Cada hora adicional de trabajo incrementa la producción en unas 17.14 unidades.

3.2. Estimación con `scikit-learn`

El modelo ajustado mediante la función `LinearRegression()` de `scikit-learn` arrojó los mismos coeficientes, validando el resultado anterior desde otra perspectiva:

$$\hat{\theta}_0 = 3{,}9177, \quad \hat{\theta}_1 = 17{,}142$$

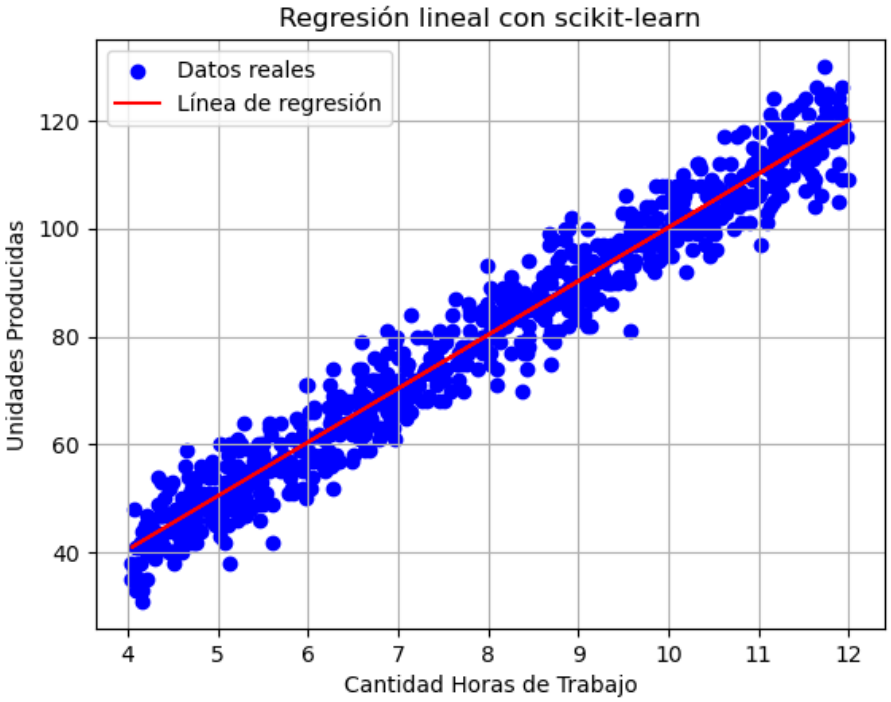

Figura 3: Modelo de regresión ajustado con `scikit-learn`

3.3. Estimación con `statsmodels`

El uso de `statsmodels` permite realizar una estimación estadística con pruebas de hipótesis e intervalos de confianza. A continuación, se presentan los resultados:

Parámetro	**Coeficiente**	**Error Est.**	***p*-valor**	**IC 95 %**
Intercepto (θ_0)	3.9177	0.460	¡0.001	[3.01, 4.82]
Pendiente (θ_1)	17.142	0.992	¡0.001	[16.81, 17.47]

Cuadro 1: Estimaciones obtenidas con `statsmodels`

$$R^2 = 0{,}956$$

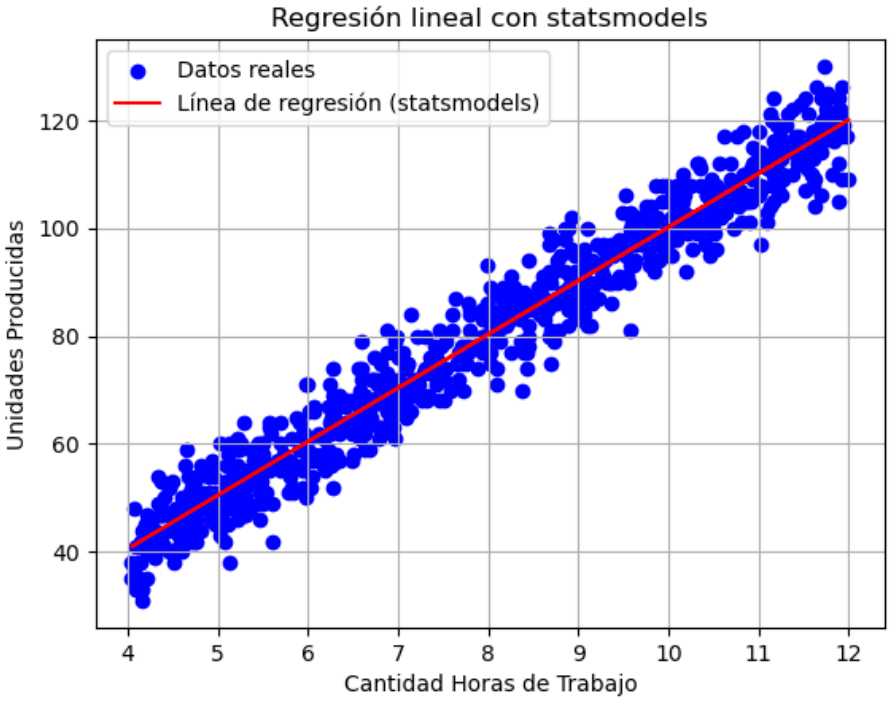

Figura 4: Línea de regresión ajustada con `statsmodels`

4. Evaluación de Supuestos del Modelo

4.1. Normalidad de los errores

La prueba de Shapiro-Wilk entregó un $p = 0{,}160$, lo que indica que no hay evidencia suficiente para rechazar la hipótesis nula de normalidad de los residuos.

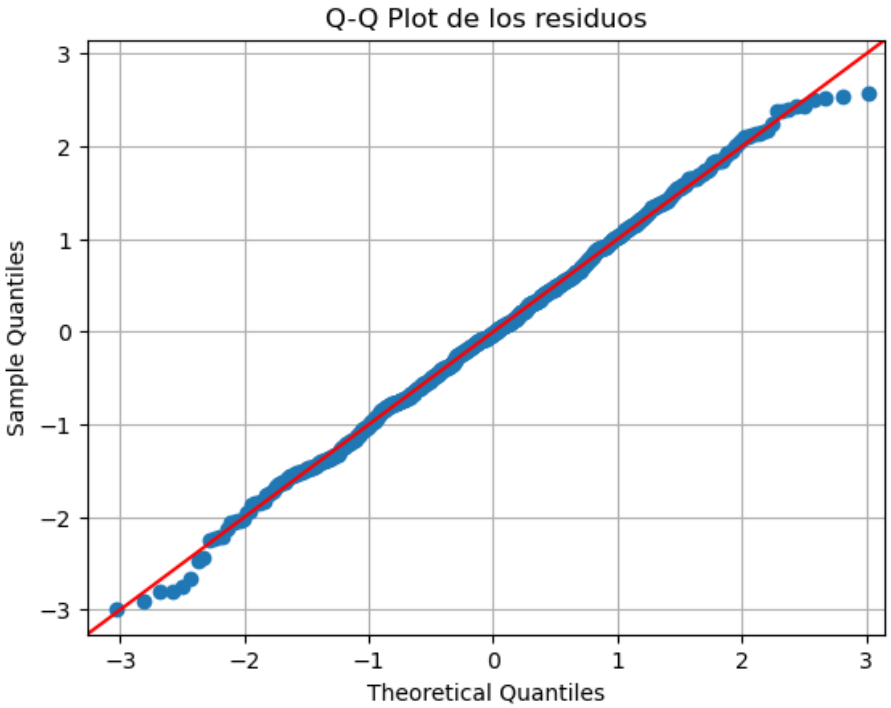

Figura 5: Gráfico Q-Q para evaluar normalidad de residuos

4.2. Heterocedasticidad

La prueba de Breusch-Pagan obtuvo un $p = 0{,}001$, por lo que se rechaza la hipótesis nula de homocedasticidad. Esto implica que hay heterocedasticidad.

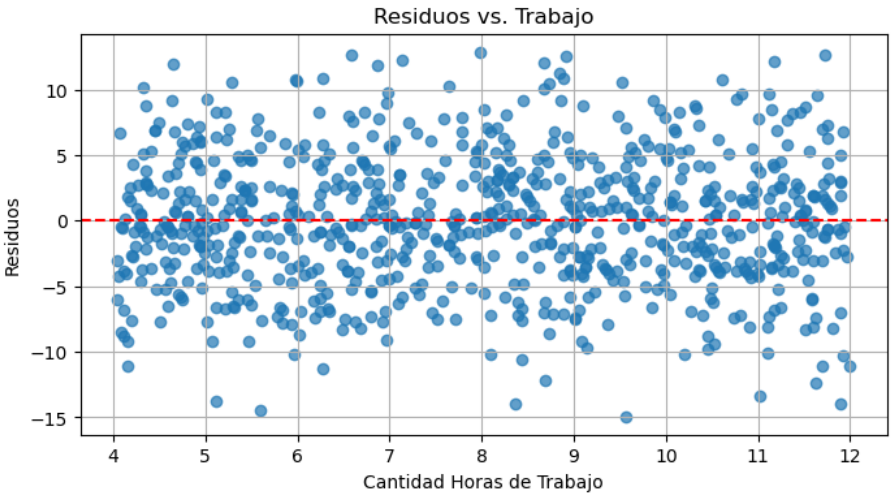

Figura 6: Gráfico de residuos vs. variable independiente

4.3. Independencia de los errores

El estadístico Durbin-Watson resultó en 1.94, indicando que no hay autocorrelación serial significativa.

5. Conclusiones

- Existe una fuerte relación lineal positiva entre las variables.
- El modelo ajustado tiene un excelente poder explicativo ($R^2 = 0{,}956$).
- Los residuos presentan distribución normal e independencia, pero se detectó heterocedasticidad.
- Se sugiere en estudios posteriores explorar modelos que manejen heterocedasticidad.