Análisis y Regresión Lineal – 2025

Lara Cellini, Bernabe Moro,
Franco Dalla Gasperina, Joaquín Gabriel Sanchez

25 de octubre de 2025

1. Regresión Lineal Simple

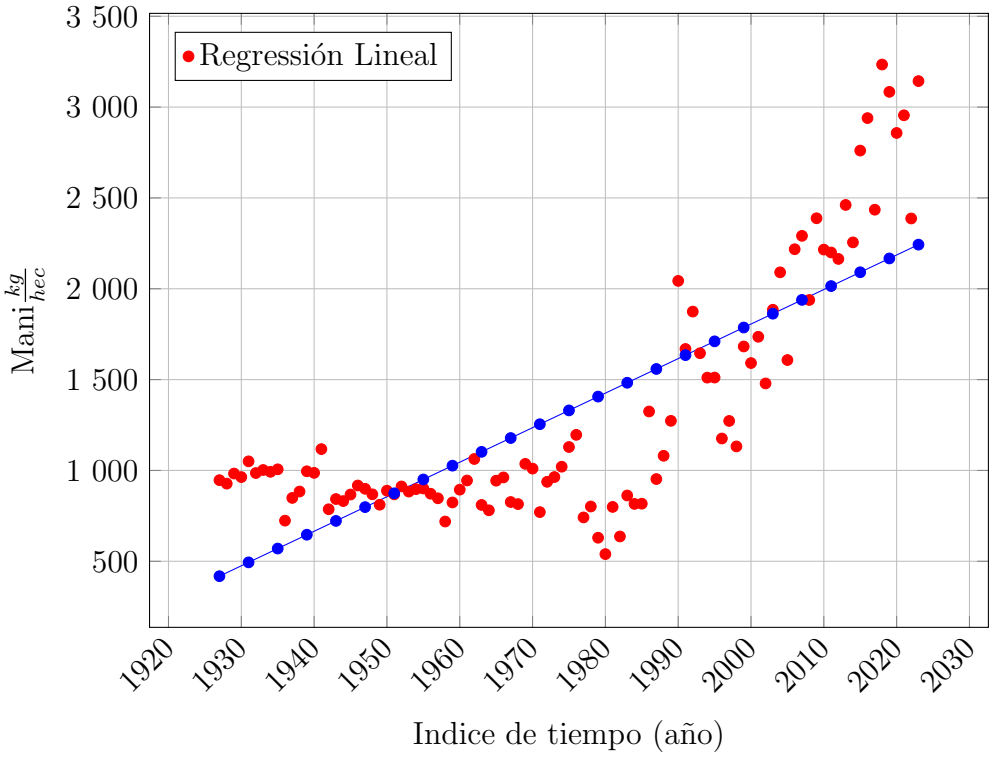
a) Se define como la variable respuesta y, representante del rendimiento de kilos por hectárea de maní producido, el cual en nuestro conjunto de datos medidos posee un intervalo que va desde $539,43\frac{kg}{hec}$ a $3234,21\frac{kg}{hec}$. Además, se definen como variables independientes: la variable x_1 que representa el indice de tiempo, la cual va del año 1927 al año 2023; y la variable x_2 que representa la superficie sembrada en hectareas; y va de 45.606 a 452.118 hectáreas.

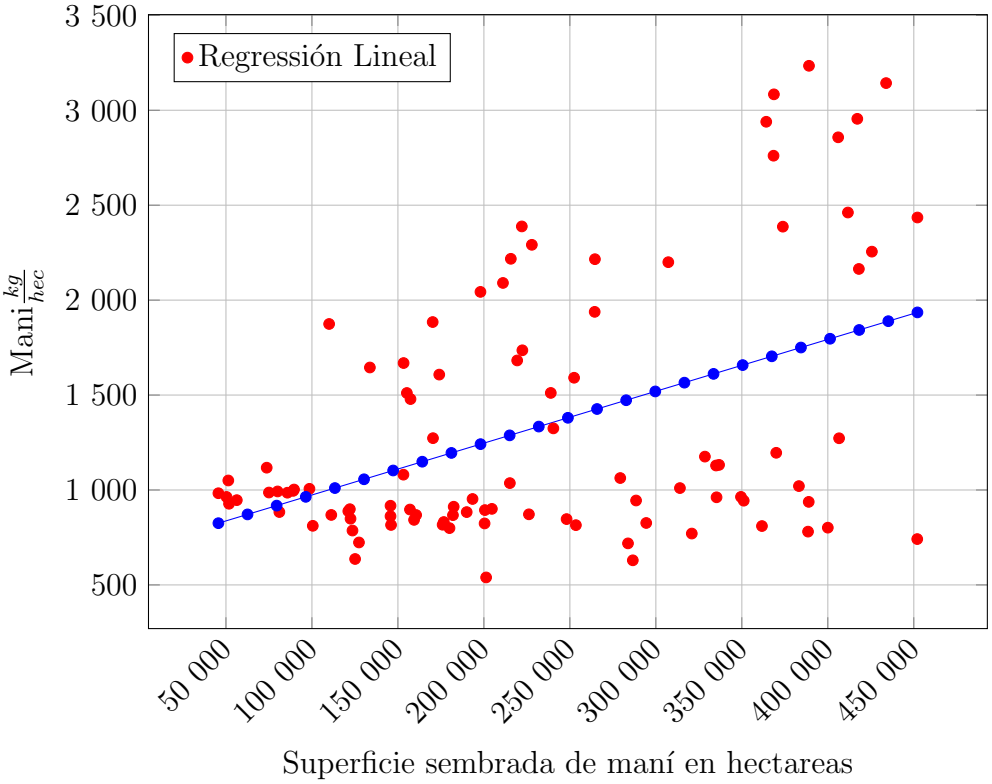
El razonamiento detrás de la elección de la variable independiente "Indice de tiempo" es el de visualizar la relación entre los avances tecnológicos de producción en el rendimiento de las cosechas. Por otro lado, el motivo por el cual se puede elegir la "Superficie sembrada en hectáreas" es que tiene una mayor correlación con la variable dependiente ya que esta es uno de los factores que la componen; además, es notable que la relación entre la superficie sembrada y el rendimiento de producción no forman una relación 1 a 1, esto a juzgar por como el valor máximo de hectáreas sembradas es diez veces mayor a el mínimo observado, mientras que el rendimiento máximo medido es 6 veces mayor que el mínimo.

Finalmente, la variable independiente fue seleccionada para este estudio debido a que nos interesa tener la capacidad de predecir el rendimiento de las cosechas anuales en base a las tecnologías utilizadas y/o la disponibilidad de tierras.

b)

Y	$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$	$\hat{\sigma}^2$
x_1	$\hat{y}_i = 19,010073638 x_i -$ $36214,433785565$	$178259,14602654814$
	R^2	r
	$0,618736208$	$0,786597869$
	$IC(\beta_1)$	$IC(\beta_0)$
	$[15,970684698;$ $22,049462578]$	$[-42217,830173978;$ $-30211,037397152]$
	$ICM(Y)$	$IP(y)$
	$[-36214,433785565 + 19,010073638 - 1,9852$ $\sqrt{178259,14602654814(\frac{1}{97} + \frac{(x^* - 1975)^2}{76048})};$ $-36214,433785565 + 19,010073638 + 1,9852$ $\sqrt{178259,14602654814(\frac{1}{97} + \frac{(x^* - 1975)^2}{76048})}]$	$[\hat{Y^*} - 1,9852\sqrt{178259,14602654814} \times$ $\sqrt{(1 + \frac{1}{97} + \frac{(x^* - 1975)^2}{76048})};$ $\hat{Y^*} + 1,9852\sqrt{178259,14602654814} \times$ $\sqrt{(1 + \frac{1}{97} + \frac{(x^* - 1975)^2}{76048})}]$
Y	$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$	$\hat{\sigma}^2$
x_2	$\hat{y}_i = 0,002731115 x_i +$ $700,57505221$	$369930,5192375846$
	R^2	r
	$0,208786109$	$0,456931186$
	$IC(\beta_1)$	$IC(\beta_0)$
	$[-0,001647406;$ $0,003814824]$	$[422,359852161;$ $978,790252259]$
	$ICM(Y)$	$IP(y)$
	$[700,57505221 + 0,002731115 - 1,9852$ $\sqrt{369930,5192375846(\frac{1}{97} + \frac{(x^* - 230633,494845361)^2}{1243287228824,25})};$ $700,57505221 + 0,002731115 + 1,9852$ $\sqrt{369930,5192375846(\frac{1}{97} + \frac{(x^* - 230633,494845361)^2}{1243287228824,25})}]$	$[\hat{Y^*} - 1,9852\sqrt{369930,5192375846} \times$ $\sqrt{(1 + \frac{1}{97} + \frac{(x^* - 230633,494845361)^2}{1243287228824,25})};$ $\hat{Y^*} + 1,9852\sqrt{369930,5192375846} \times$ $\sqrt{(1 + \frac{1}{97} + \frac{(x^* - 230633,494845361)^2}{1243287228824,25})}]$


Regressión Lineal
3 500
3 000
2 500
2 000
1 500
1 000
500
Mani $\frac{kg}{hec}$
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030
Indice de tiempo (año)


Regressión Lineal
3 500
3 000
2 500
2 000
1 500
1 000
500
Mani $\frac{kg}{hec}$
50 000
100 000
150 000
200 000
250 000
300 000
350 000
400 000
450 000
Superficie sembrada de maní en hectareas

c) Como se puede observar en las tablas, el valor de R^2 en el modelo que utiliza como variable independiente la superficie sembrada es mucho menor que el valor de R^2 el modelo que toma en cuenta los años. Mientras más alto es el valor de R^2, mejor será nuestro modelo de regresión lineal simple para explicar la variación del rendimiento. Esto significa que es factible negar el hecho de que la variación de la superficie sembrada con maní explique el aumento en el rendimiento de las cosechas.

Por otro lado, r indica la dirección de la relación entre las dos variables. Al ser r un valor dentro del intervalo de 0 a 1, se puede suponer que existe una correlación positiva fuerte entre el pasar de los años y el aumento del rendimiento de las cosechas. Al aumentar x (los años), aumenta también y (el rendimiento). Esto es visible en la gráfica que muestra el modelo con variable independiente x_1.

En cambio, al considerar el r que corresponde al modelo que toma en cuenta las hectáreas sembradas como variable predictiva, se observa que la relación también sería positiva, mas no tan fuerte. Por lo tanto, es posible suponer que no está tan fuertemente relacionado el aumento en la cantidad de hectáreas sembradas con el aumento en el rendimiento de las cosechas, aunque tampoco es posible afirmar que la correlación sea nula.

Con $x_1 : r = \sqrt{R^2} = \sqrt{0,618736208} = 0,786597869$.

Con $x_2 : r = \sqrt{R^2} = \sqrt{0,208786109} = 0,456931186$.

2. Regresión Lineal Múltiple

Para lograr una aproximación mas precisa se intenta hacer una regresión lineal múltiple utilizando todas las variables independientes disponibles en el dataset.

Las variables son:

- Independientes:
  1. Año.
  2. Superficie sembrada en hectáreas.
  3. Superficie cosechada en hectáreas.
  4. Producción en toneladas.
- Dependiente:

- Rendimiento.

a) Se define el modelo de regresión lineal múltiple por descenso de gradiente:

El objetivo de la regresión lineal es encontrar el vector de parámetros $\theta = [\theta_0, \theta_1, \theta_2, \theta_3, \theta_4]^T$ que minimiza la función de costo $J(\theta)$, la cual mide el error cuadrático medio (MSE) entre los valores reales (y_i) y las predicciones ($\hat{y}_i$).

La función de costo utilizada es el Error Cuadrático Medio (MSE) dividido por $2m$:

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{y}^i - y^i)^2 = \frac{1}{2m} (X\theta - y)^T (X\theta - y)$$

donde m es el número de muestras.

El algoritmo fue programado y ejecutado con una tasa de aprendizaje $\eta = 0{,}01$, un valor de tolerancia $\epsilon = 0{,}0001$ y un criterio de convergencia de cambio mínimo en el costo de 0,00000001.

https://github.com/FrancoDalla/matematicas/blob/main/DescensoGradiente/Descenso_con_error.py

Este algoritmo dio como resultados: $R^2 = 0{,}9120$, indicador de que el modelo explica cerca del 91,20 % de la variabilidad del rendimiento. Como es un valor alto, es decir cercano a 1, muestra que este modelo se ajusta correctamente.

b) Se define el modelo de regresión lineal múltiple por mínimos cuadrados de la siguiente forma:

$$f(b_0, b_1, b_2, b_3, b_4) = \sum \left[y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} + b_4 x_{4i}) \right]^2$$

A partir de ella se sacan las derivadas parciales de cada variable independiente y b_0, el cual representa nuestra ordenada al origen:

$$\frac{\partial f}{\partial b_0} = -2\sum\left[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i})\right]$$
$$\frac{\partial f}{\partial b_1} = -2\sum x_{1i}\left[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i})\right]$$
$$\frac{\partial f}{\partial b_2} = -2\sum x_{2i}\left[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i})\right]$$
$$\frac{\partial f}{\partial b_3} = -2\sum x_{3i}\left[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i})\right]$$
$$\frac{\partial f}{\partial b_4} = -2\sum x_{4i}\left[y_i - (b_0 + b_1x_{1i} + b_2x_{2i} + b_3x_{3i} + b_4x_{4i})\right]$$

Tras ser igualadas a cero se obtiene el siguiente sistema de ecuaciones normales:

$$\sum y_i = nb_0 + b_1\sum x_{1i} + b_2\sum x_{2i} + b_3\sum x_{3i} + b_4\sum x_{4i}$$
$$\sum x_{1i}y_i = b_0\sum x_{1i} + b_1\sum x_{1i}^2 + b_2\sum x_{2i}x_{1i} + b_3\sum x_{3i}x_{1i} + b_4\sum x_{4i}x_{1i}$$
$$\sum x_{2i}y_i = b_0\sum x_{2i} + b_1\sum x_{1i}x_{2i} + b_2\sum x_{2i}^2 + b_3\sum x_{3i}x_{2i} + b_4\sum x_{4i}x_{2i}$$
$$\sum x_{3i}y_i = b_0\sum x_{3i} + b_1\sum x_{1i}x_{3i} + b_2\sum x_{2i}x_{3i} + b_3\sum x_{3i}^2 + b_4\sum x_{4i}x_{3i}$$
$$\sum x_{4i}y_i = b_0\sum x_{4i} + b_1\sum x_{1i}x_{4i} + b_2\sum x_{2i}x_{4i} + b_3\sum x_{3i}x_{4i} + b_4\sum x_{4i}^2$$

Análogamente, es posible expresarlo de forma matricial:

$$\underbrace{\begin{pmatrix} n & \sum x_{1i} & \sum x_{2i} & \sum x_{3i} & \sum x_{4i} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{2i}x_{1i} & \sum x_{3i}x_{1i} & \sum x_{4i}x_{1i} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \sum x_{3i}x_{2i} & \sum x_{4i}x_{2i} \\ \sum x_{3i} & \sum x_{1i}x_{3i} & \sum x_{2i}x_{3i} & \sum x_{3i}^2 & \sum x_{4i}x_{3i} \\ \sum x_{4i} & \sum x_{1i}x_{4i} & \sum x_{2i}x_{4i} & \sum x_{3i}x_{4i} & \sum x_{4i}^2 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \\ \sum x_{3i}y_i \\ \sum x_{4i}y_i \end{pmatrix}}_{Y}$$

$$Y = X\beta \mid n = 97$$

Variable	Valor	Descripción
$\sum x_{1i}$	191575	Suma de los valores de la variable independiente años
$\sum x_{2i}$	22,371,449	Suma de los valores de la variable independiente superficie sembrada en hectáreas
$\sum x_{3i}$	21,446,480	Suma de los valores de la variable independiente superficie cosechada en hectáreas
$\sum x_{4i}$	36,439,841.4	Suma de los valores de la variable independiente producción en toneladas
$\sum x_{1i}^2$	378,436,673	Suma de los valores al cuadrado de años
$\sum x_{2i}^2$	6,402,892,696,449	Suma de los valores al cuadrado de superficie sembrada
$\sum x_{3i}^2$	5,972,973,986,094	Suma de los valores al cuadrado de superficie cosechada
$\sum x_{4i}^2$	24,719,672,959,067.2	Suma de los valores al cuadrado de producción
$\sum y_i$	129,054.78	Suma de los valores de la variable dependiente rendimiento

Sustituyendo los valores en la matriz:

$$\begin{pmatrix} 97 & 191575 & 22371449 & 21446480 & 36439841{,}4 \\ 191575 & 378436673 & 22371449 \times 191575 & 21446480 \times 191575 & 36439841{,}4 \times 191575 \\ 22371449 & 191575 \times 22371449 & 6402892696449 & 21446480 \times 22371449 & 36439841{,}4 \times 22371449 \\ 21446480 & 191575 \times 21446480 & 22371449 \times 21446480 & 5972973986094 & 36439841{,}4 \times 21446480 \\ 36439841{,}4 & 191575 \times 36439841{,}4 & 22371449 \times 36439841{,}4 & 21446480 \times 36439841{,}4 & 24719672959067{,}2 \end{pmatrix} \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 129054{,}78 \\ 191575 \times 129054{,}78 \\ 22371449 \times 129054{,}78 \\ 21446480 \times 129054{,}78 \\ 36439841{,}4 \times 129054{,}78 \end{pmatrix}$$

Se tiene entonces una ecuación matricial donde la intención es despejar la matriz β:

Se sabe que:

$$X^T Y = X^T X \beta$$

Entonces para despejar $\hat{\beta}$ se debe hacer:

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

Para hacer la estimación de los betas se realizó un script en Python el cual utiliza las librerías "panda" para leer y procesar el dataset en formato *csv* y "numpy" para los cálculos.

https://github.com/FrancoDalla/matematicas/blob/main/MinimosCuadrados/minimos_cuadrados.py

Los resultados obtenidos fueron los siguientes:

$$\begin{aligned} b_0 \text{ (intercepto)} &= -16988{,}052990805078 \\ b_1 \text{ (Año)} &= 9{,}257050531336022 \\ b_2 \text{ (Superficie Sembrada)} &= 0{,}0031599268599484262 \\ b_3 \text{ (Superficie Cosechada)} &= -0{,}0066507792825439335 \\ b_4 \text{ (Producción)} &= 0{,}002069716111442946 \\ R^2 \text{ del modelo} &= 0{,}9120 \end{aligned}$$

La estimación de la ecuación de regresión finalmente queda con esta forma:

$$\begin{aligned} \hat{y} &= \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \hat{b}_3 x_3 + \hat{b}_4 x_4 \\ \hat{y} &= -16988{,}052990805078 + 9{,}257050531336022 x_1 + 0{,}0031599268599484262 x_2 \\ &\quad - 0{,}0066507792825439335 x_3 + 0{,}002069716111442946 x_4 \\ \hat{y} &= -16988{,}052 + 9{,}25 \times \text{año} + 0{,}00315 \times \text{SuperficieSembrada} \\ &\quad - 0{,}00663 \times \text{SuperficieCosechada} + 0{,}00207 \times \text{Producción} \end{aligned}$$

En este caso es posible observar como el coeficiente del año (b_1) indica una tendencia del rendimiento a lo largo del tiempo posiblemente por mejoras tecnológicas o mejoras en las prácticas agrícolas.

El coeficiente de Superficie Sembrada ($b_2 > 0$) sugiere que mayores superficies tienden a incrementar levemente el rendimiento promedio.

Por otro lado, Superficie Cosechada ($b_3 < 0$) podría denotar el hecho de que cuando aumenta la superficie efectivamente cosechada, el rendimiento promedio baja, lo cual podría darse por factores climáticos, pestes, etc.

El coeficiente de Producción ($b_4 > 0$) es positivo, esto es coherente con la idea de que mayor producción total se asocia con mayores rendimientos.

El $R^2 = 0{,}9120$ indica que el modelo explica cerca del 91,02 % de la variabilidad del rendimiento. Como es un valor alto, es decir cercano a 1, muestra que este modelo se ajusta de manera razonable a los datos y permite estimar el rendimiento a partir de las variables independientes.

En conclusión, ambos métodos parecen dar un estimativo adecuado, de hecho, el R^2 conseguido por ambos métodos es el mismo.

c) El agregado de más variables predictoras mejoró enormemente la estimación en comparación a la obtenida en el inciso c). Se puede notar fácilmente comparando los R^2 obtenidos utilizando regresión lineal simple con el R^2 obtenido con regresión lineal múltiple:

Regresión lineal simple:

- Con $x_1 : r = \sqrt{R^2} = \sqrt{0,618736208} = 0,786597869$.
- Con $x_2 : r = \sqrt{R^2} = \sqrt{0,208786109} = 0,456931186$.

Regresión lineal múltiple:

- Descenso del gradiente: $R^2 = 0{,}9120$
- Mínimos cuadrados: $R^2 = 0{,}9120$

La mejoría en el estimativo utilizando regresión lineal múltiple se debe principalmente a dos motivos:

El primer motivo, y el más evidente, es que el cálculo del rendimiento para cada muestra es en base a la superficie cosechada y las toneladas producidas. Esto, sin embargo, no significa que reduciendo las variables independientes a solo estos dos datos de un mejor estimativo. En todo caso, si se ejecutan los algoritmos con solo estos datos, el estimativo empeora ($R^2 = 0{,}8621$ en ambos algoritmos), mostrando la importancia de las otras variables independientes, como lo es el índice del tiempo que al agregarlo aumenta R^2 a 0,9104 en ambos algoritmos.

El segundo motivo es que se utilizan variables independientes que no pueden ser conocidas previo al sembrado, como es el caso de la superficie cosechada y la producción en toneladas. En un caso real donde, por ejemplo, se quisiera estimar el rendimiento de las cosechas en base a un dato conocido (el año en que se siembra) y un dato que es posible afectar (cuanto se va a sembrar), el modelo terminaría siendo menos preciso que si se considerara únicamente el tiempo como variable independiente ($R^2 = 0{,}6230$ con ambos algoritmos).

Nota:

Para algunos cálculos se utilizaron scripts hechos en python, específicamente para el S_{xy} y la varianza de la variable independiente x_1 (https://github.com/FrancoDalla/matematicas/blob/main/sxy_calculatoo.py) y (https://github.com/FrancoDalla/matematicas/blob/main/varianza.py).

El resto de cálculos (especialmente los necesario para realizar la tabla) fueron en su mayoría realizados utilizando calculadora y/o hechos a mano.