UNIVERSIDAD NACIONAL DE JAÉN

UNIVERSIDAD NACIONAL DE JAÉN

DEPARTAMENTO ACADÉMICO DE CIENCIAS BÁSICAS Y APLICADAS

INGENERÍA CIVIL

MONOGRAFÍA
REGRESIÓN LINEAL SIMPLE

Autores:
Dra. Rosario Yaqueliny Llauce Santamaria
Dra. Marcela Yvone Saldaña Miranda
Ing. Mario Félix Olivera Aldana

Ciclo Académico: 2024_II

Jaén – Perú, noviembre 2024

INDICE

I. Introducción

En la presente investigación analizaremos las definiciones, veremos las aplicaciones analizaremos los supuestos y estudiaremos su interpretación de la regresión lineal simple, así como también determinaremos algunos ejemplos que fundamenten y aclaren más dicho tema.

El método de regresión lineal simple se emplea para examinar la relación que existe entre dos variables: una que actúa como independiente y otra que depende de la primera. Se asume que, en este modelo, los cambios en la variable dependiente (Y) pueden ser explicados, al menos en parte, por los cambios en la variable independiente (X) mediante una relación lineal. La ecuación básica es la siguiente:

$$\hat{Y} = a + bX + \varepsilon$$

donde (a) es la intersección o intercepto, (b) indica la pendiente, y (ε) es el que indica el error. Este modelo se utiliza ampliamente en diversas áreas para hacer predicciones y comprender tendencias entre dos variables relacionadas.

II. Objetivos

2.1. Objetivo general:

Analizar y comprender la conexión entre dos variables a través del modelo de Regresión Lineal Simple, aplicando conceptos estadísticos para prever el comportamiento que tiene una variable en función de otra.

2.2. Objetivos específicos:

- Explicar los fundamentos teóricos de la Regresión Lineal, incluyendo la formulación de su modelo y sus componentes, para establecer una base conceptual sólida.
- Calcular el modelo de Regresión Lineal Simple mediante el método de mínimos cuadrados, mostrando su aplicación en un ejemplo práctico.
- Examinar los supuestos necesarios para la efectividad del modelo de regresión lineal simple, como la linealidad, homogeneidad de la varianza, normalidad e independencia, y su impacto en la precisión de las predicciones.
- Aplicar e interpretar los coeficientes de regresión (pendiente e intercepto) y el Índice de Determinación (R^2), analizando cómo se adapta la gráfica a la información y la correlación entre las variables.
- Realizar anticipaciones basadas en el modelo de regresión ajustado, utilizando tanto interpolación como extrapolación, para ilustrar el uso práctico de la regresión lineal en la estimación de valores no observados.

III. Marco teórico

3.1. Definición Regresión Lineal Simple

La regresión lineal simple es un técnica estadística empleada para analizar cómo se relacionan dos variables cuantitativas, la variable dependiente Y (variable de respuesta) y la variable independiente X (variable predictora) (Alea et al., 2015).

El principal objetivo del modelo de regresión lineal simple es determinar una ecuación lineal que se adapte adecuadamente a los datos. Esto facilita la estimación del valor de la variable dependiente basado en los valores de la variable independiente. (Walpole et al., 2012).

3.2. Ecuación de regresión lineal simple

Según Barreno et al., (2013), el modelo de la regresión lineal simple está dado a continuación:

$$\boldsymbol{Y_i = b_0 + b_1 X_i + \varepsilon_i}$$

Donde:

$b_0\ y\ b_1$: son los coeficientes de la regresión que deben estimarse. Aquí, β_0 se conoce como el intercepto o término constante, mientras que β_1 es la pendiente de la recta.

Y_i: representa la variable de respuesta o dependiente para la i-ésima observación, es decir, el valor que se desea predecir.

X_i: se refiere a la variable independiente o predictora para la i-ésima observación. También se le conoce como variable explicativa o regresora.

ε_i: es un residual o error en el ajuste de la ecuación.

3.2.1. Condiciones para modelar una regresión lineal

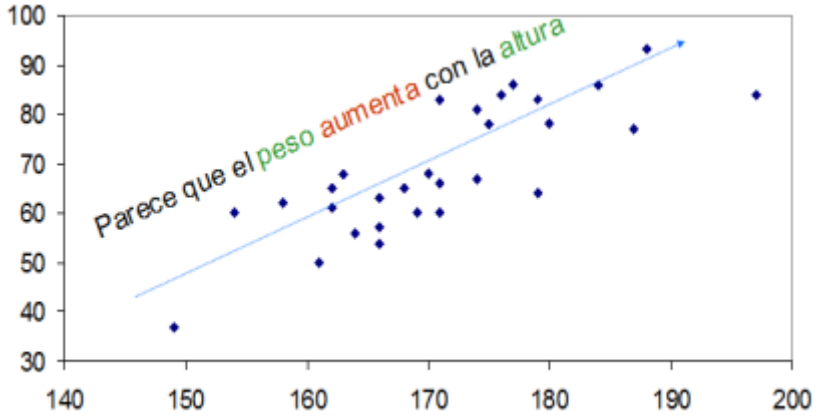
Lagunas (2014), determina las siguientes condiciones:

- **Relación linealidad:** El promedio de la variable dependiente Y se debe relacionar de manera lineal con la variable X. Esto implica que, independientemente del nivel de X, variaciones de magnitud constante en esta variable están vinculadas a un cambio uniforme en el valor promedio de Y.
- **Homogeneidad de la varianza:** La dispersión de la variable dependiente Y debe permanecer constante sin tener en cuenta los valores que asuma X. En otras palabras, la variabilidad de Y no debe mostrar dependencia de X.
- **Distribución normal:** Cuando se mantiene constante un valor de la variable explicativa X, los valores de Y se distribuyen de manera que se ajustan a la forma característica de la distribución normal.
- **Independencia:** Es importante que cada medición de la variable Y debe ser independiente de las otras observaciones.

3.3. Diagrama de dispersión

El diagrama de dispersión, también conocido como una nube de puntos, facilita la representación gráfica del tipo de relación entre las variables X e Y, y es útil para identificar posibles valores inusuales o extremos en los datos. (Laguna, 2014)

Figura 1: Diagrama de dispersión de dos variables X e Y.



Fuente: (Laguna, 2014)

3.4. Ajuste de recta por mínimos cuadrados

Es un método o herramienta utilizada en la regresión lineal simple para identificar la recta que se adapte más a un conjunto de datos. Esto implica determinar los valores de b_0 y b_1, que hacen que la línea se acerque lo máximo posible a los datos observados. El enfoque consiste en reducir la suma de los cuadrados de los errores. En este contexto, un error (e_i) se refiere a la diferencia dentro del valor registrado Y_i y el valor predicho $\widehat{Y}_i$ (Walpole et al., 2012, pag.395).

$$e_i = Y_i - \widehat{Y}_i \ , \quad i = 1,2, \dots ., n$$

La idea es minimizar:

$$SCRes = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Donde:

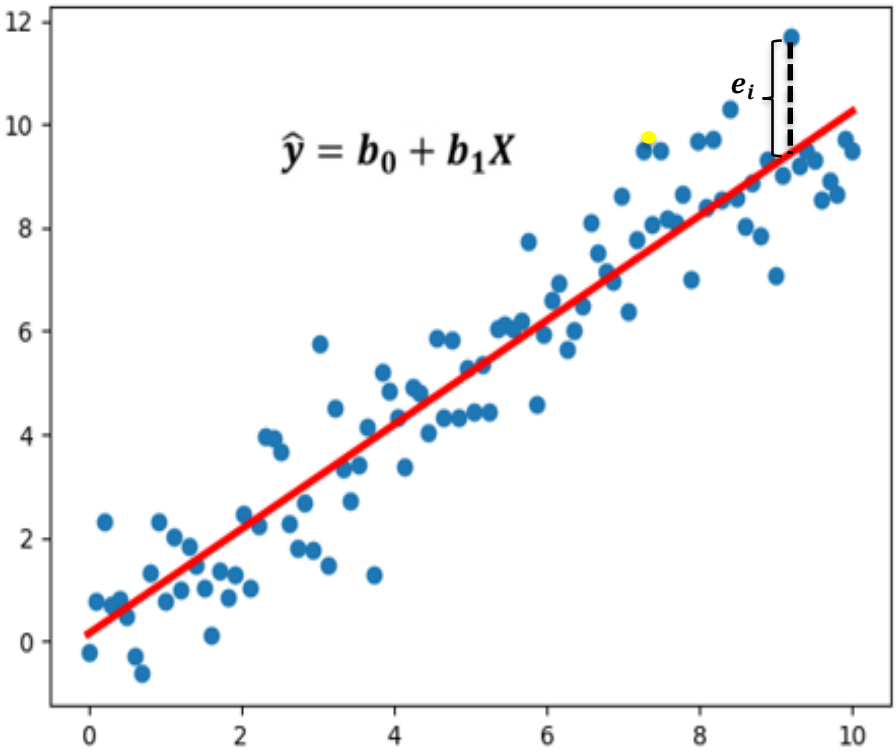
$SCRes$: Sumatoria de los errores elevados al cuadrado.

La recta de mínimos cuadrados se utiliza para aproximar el conjunto de puntos: $(x_1;\ y_1)$, $(x_2;\ y_2), (x_3;\ y_3), \dots, (x_n;\ y_n)$. Su ecuación se expresa como $\hat{y}_i = \beta_0 + \beta_1 X_i$, y se determina al resolver las siguientes ecuaciones (Estuardo, 2012).

$$b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

$$b_0 = \bar{Y} - \beta_1 \bar{x}$$

Figura 2: Gráfica de la línea de regresión lineal simple



Fuente: (Merchan, 2024)

Interpretación de la ecuación de recta de la gráfica

Laguna (2014), explica la interpretación de dos elementos clave en la ecuación:

- **Ordenada en el origen (b_0):**
  Este valor solo estima el resultado de Y cuando X es cero.
- **Pendiente (b_1):**
  Este elemento es clave porque señala el cambio en $\hat{Y}$ por cada unidad que se modifica en X, además ilustra la conexión existente entre las dos variables, revelando cómo los valores de $\hat{Y}$ fluctúan cuando X incrementa de uno en uno. Asimismo, tanto el coeficiente de regresión b_1 como el coeficiente de correlación (r) siempre tienen el mismo signo.
  Para $b_1 > 0$, cualquier incremento en X indica que $\hat{Y}$ aumenta.
  - Para $b_1 < 0$, $\hat{Y}$ disminuye cuando X crece.

3.5.Interpolación y extrapolación

Laguna (2014), define de la siguiente manera:

- La interpolación se refiere al proceso de calcular valores dentro del intervalo de datos disponibles.
- La extrapolación, se trata de calcular valores que están fuera del rango de los datos disponibles, empleando la misma línea de ajuste para anticipar un valor de x que se encuentra más allá del intervalo dado. No obstante, este método presenta un mayor riesgo de error, ya que asume que la tendencia observada se extenderá más allá de las fronteras de los datos disponibles.

Uno de los principales objetivos de la regresión es utilizar el modelo para anticipar el resultado de Y (la variable dependiente) a partir de un valor de X (la variable independiente) que no se encuentra en los datos observados (Laguna, 2014).

3.6. Bondad de un Ajuste

La bondad de ajuste permite evaluar si el modelo es apropiado para representar la conexión entre las variables y evaluar la fiabilidad de sus predicciones. Un modelo con buena bondad de ajuste indica que la línea de regresión se aproxima a la mayoría de los datos, lo que permite obtener una ilustración excelente de la relación que existe entre las variables. (Laguna, 2014).

3.6.1. El coeficiente de determinación R^2

La medida principal para analizar qué tanto se adapta el modelo es el coeficiente de determinación, R^2. Este valor nos muestra qué tan bien la línea de regresión representa información de la muestra. R^2 se interpreta como el porcentaje de la variación total de la variable dependiente (Y) que el modelo logra explicar a través de la línea de regresión (Laguna, 2014).

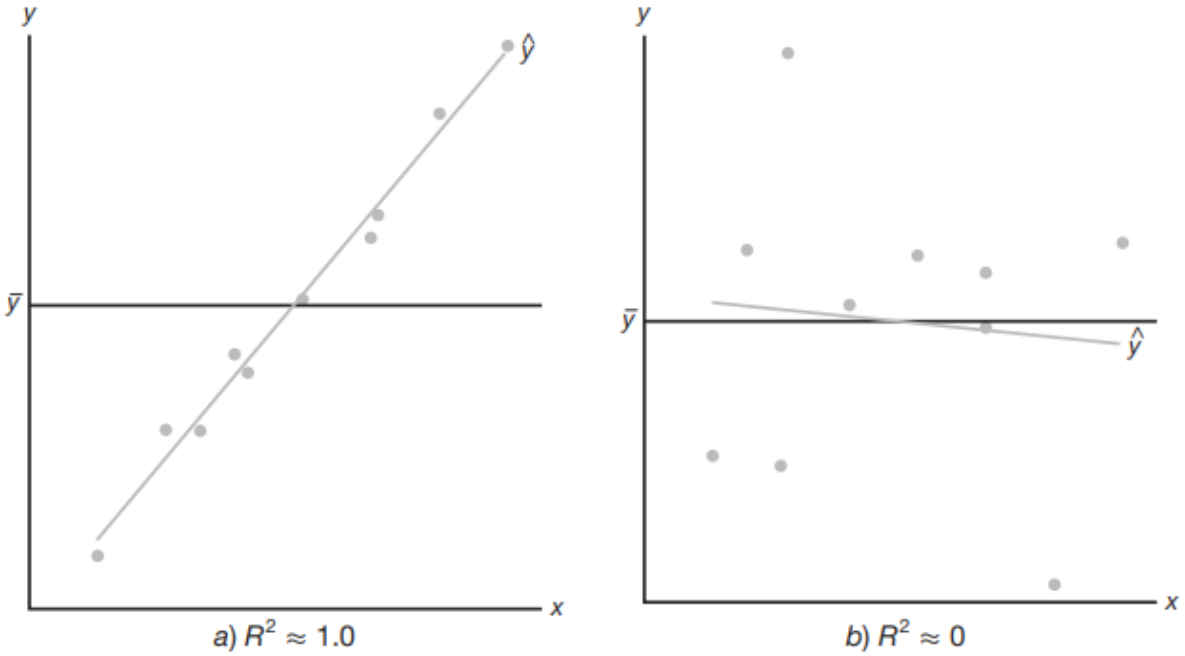
$$R^2 = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum(Y - \bar{Y})^2}$$

Donde.

$\hat{Y}$: Valor predicho de la recta.

$\bar{Y}$: Media de los valores de la variable dependiente.

Figura 3: Gráficas que ilustran los ajustes de recta de acuerdo conR^2.



Fuente: (Walpole et al., 2012, pag.408)

Características del coeficiente de determinación R^2:

- El valor de R^2 es un valor numérico y solo puede estar entre 0 y 1.
- Si el ajuste del modelo es bueno, R^2 se acercará a uno, indicando una asociación fuerte entre las dos variables.
- En cambio, si el ajuste es deficiente, R^2 estará cerca de cero, lo que significa que la línea de regresión no explica una correlación entre x e y.

3.6.2. Coeficiente de Correlación lineal de Pearson (r)

El coeficiente "r" mide la magnitud de la relación lineal que existe entre dos variables, X e Y. Por lo tanto, los valores que puede asumir "r" oscilan entre [-1 ,1] (Walpole et al., 2012):

- $\boldsymbol{r = 1}$: Muy buena correlación en sentido positivo (si X aumenta, Y también aumenta de forma lineal).
- $\boldsymbol{r = -1}$: Muy buena correlación en sentido negativo (si X aumenta, Y disminuye de forma lineal).
- $\boldsymbol{r = 0}$**:** No hay una correlación lineal clara entre X e Y.

Este coeficiente, puede parecer complicado al observar su fórmula matemática, pero en realidad representa una idea sencilla: Si "r" se aproxima a 1 (en términos absolutos) existe bastante

relación entre las variables X e Y, eso quiere decir que ambas aumentan o disminuyen. Este tipo de cambio simultáneo entre ambas variables es lo que se denomina covarianza. (Laguna,2014)

La fórmula para calcular r es:

$$r = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

Donde:

- $Cov(X,Y)$: Covarianza entre X e Y.
- σ_X: Desviación estándar de X, se calcula como:

$$\sigma_X = \sqrt{\frac{\sum(X_i - \bar{X})}{n}}$$

- σ_Y: Desviación estándar de Y, se calcula como:

$$\sigma_Y = \sqrt{\frac{\sum(Y_i - \bar{Y})}{n}}$$

Relación entre r y R^2:

Según Laguna (2014), es fundamental comprender las diferencias entre el coeficiente de correlación (**r**) y el coeficiente de determinación ($\mathbf{R^2}$) (Laguna,2014).

- $\boldsymbol{R^2}$**:** Este coeficiente indica el porcentaje de ajuste que existe entre la variable dependiente Y y la variable independiente X.
- $\boldsymbol{r}$: Este coeficiente evalúa el grado de relación o asociación existente entre variable X e Y.

Por ello se cumple la siguiente igualdad:

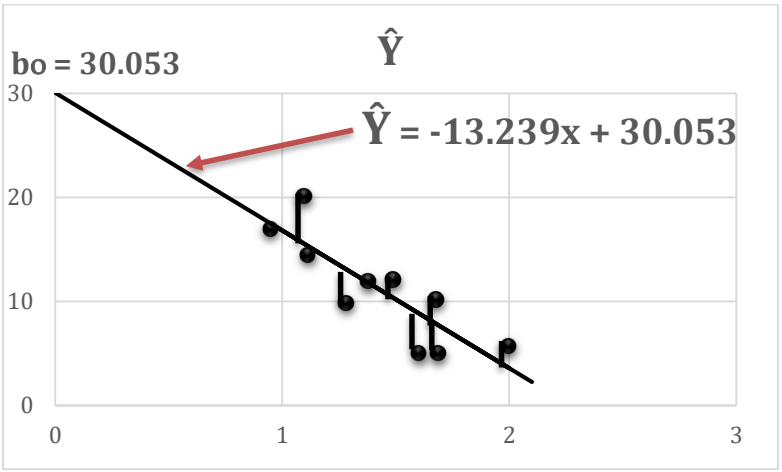
$$R^2 = r^2$$

3.7.Cálculo de los Residuos

Las distancias en dirección vertical entre la línea de la ecuación lineal y los puntos se denominan residuos

$$Observación = Ajustes + Residuio$$
$$\boldsymbol{Y} = \boldsymbol{\hat{Y}} + (\boldsymbol{Y} - \boldsymbol{\hat{Y}})$$

Figura 4: Gráficas que ilustran las constantes



Debemos entender que:

- Mientas mayor sea el valor de los residuos, mayor dispersión hay en la recta, por lo tanto, la recta menos se ajusta.
- Mientras menos se ajuste la recta, tenemos menor capacidad de predecir.

3.8. Error estándar de estimación

Se trata de una medida que refleja, en promedio, la magnitud de la desviación entre las cifras observadas y aquellas que son estimadas por la ecuación de regresión. Esta medida evalúa la exactitud de las predicciones generadas por el modelo. (Estuardo, 2012)

$$SEE = s_{y.x} = \sqrt{\frac{\sum\left(Y - \hat{Y}\right)^2}{n-2}}$$

$$SEE = s_{y.x} = \sqrt{\frac{\sum Y^2 - b_0\sum Y - b_1\sum XY}{n-2}}$$

Donde:
$SEE = s_{y.x}$: Es el error estándar de estimación o también conocido como desviación estándar.

3.9.Error estándar de estimación para un valor en específico

Se determina de la siguiente manera:

$$s_f = SEE * \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X - \bar{X})^2}}$$

Donde:

s_f: Error estándar para una predicción específica.

SEE: Error estándar residual.

$\bar{X}$: Promedio de la variable independiente.

3.10. Ejemplos de Aplicación

3.10.1. Ejemplo 1:

Se tienen datos sobre la relación en el tiempo de curado (en días) y la capacidad del concreto para soportar fuerzas de compresión (en MPa). Usando una regresión lineal, estima la capacidad del concreto para soportar fuerzas de compresión después de 18, 36, 47 días de curado.

Variables:

Independiente: Tiempo de curado (días).

Dependiente: Resistencia a la compresión (MPa).

TIEMPO DE CURADO (X)	RESISTENCIA MPa (Y)
5	10
10	15
20	22
25	25
30	28
40	32
50	35
180	167

SOLUCIÓN:

Se debe calcular la ecuación de la recta de regresión "$\hat{Y} = b_1 X + b_0$" para predecir la resistencia a compresión del concreto después de 18, 36, 47 días.

CALCULAMOS CADA VALOR DE LA ECUACIÓN GENERAL:

PROMEDIO ($\bar{X}$):

$$\bar{X} = \frac{\sum_{i=1}^{n} \boldsymbol{X}}{\boldsymbol{n}}$$

$$\bar{X} = \frac{5 + 15 + 20 + 25 + 30 + 35 + 50}{7}$$

$$\bar{X} = \frac{180}{7} \qquad \Longrightarrow \quad \bar{X} = 25.7143$$

PROMEDIO ($\bar{Y}$):

$$\bar{Y} = \frac{\sum_{i=1}^{n} \boldsymbol{Y}}{\boldsymbol{n}}$$

$$\bar{Y} = \frac{10 + 15 + 22 + 25 + 28 + 32 + 35}{7}$$

$$\bar{Y} = \frac{167}{7} \quad \Longrightarrow \quad \bar{Y} = 23.8571$$

PARA LOS DEMÁS VALORES USAMOS EXCEL:

TIEMPO DE CURADO **(X)**	RESISTENCIA MPa **(Y)**	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	XY	X^2
5	10	-20.7142857	-13.8571429	287.0408163	429.0816327	50	25
15	15	-10.7142857	-8.85714286	94.89795918	114.7959184	225	225
20	22	-5.71428571	-1.85714286	10.6122449	32.65306122	440	400
25	25	-0.71428571	1.14285714	-0.816326531	0.510204082	625	625
30	28	4.28571429	4.14285714	17.75510204	18.36734694	840	900
35	32	9.28571429	8.14285714	75.6122449	86.2244898	1120	1225
50	35	24.2857143	11.1428571	270.6122449	589.7959184	1750	2500
180	167	0	0	755.7142857	1271.428571	5050	5900

CALCULAMOS EL VALOR DE LA PENDIENTE (b_1):

$$b_1 = \frac{\sum (X - \bar{X})\,(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

REEMPLAZAMOS LOS DATOS EN LA PRIMERA ECUACIÓN:

$$b_1 = \frac{\sum(X - \bar{X})\,(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

$$b_1 = \frac{287.0408163 + 94.89795918 + 10.6122449 + -0.816326531 + 17.75510204 + 75.6122449 + 270.6122449}{429.0816327 + 94.89795918 + 32.65306122 + 0.510204082 + 18.36734694 + 86.2244898 + 589.7959184}$$

$$b_1 = \frac{755.7142857}{1271.428571}$$

$$b_1 = 0.594382022$$

$$b_1 = 0.5944$$

CALCULAMOS EL VALOR DEL INTERCEPTO (b_0):

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = \frac{\sum \bar{Y}}{n} - \frac{b_1 \sum X}{n}$$

REEMPLAZAMOS LOS DATOS EN LA PRIMERA ECUACIÓN:

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_0 = 23.8571 - (0.594382022)(25.7143)$$

$$b_0 = 23.8571 - 15.284117628$$

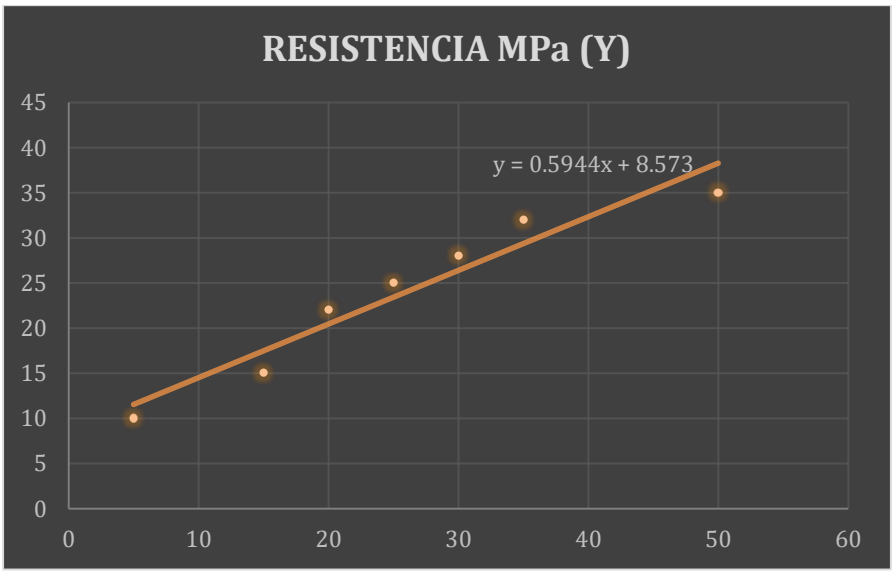
$$b_0 = 8.572982372$$

$$b_0 = 8.5730$$

$\therefore$ *LA FUNCIÓN QUEDA DE LA SIGUIENTE FORMA*:

$$\hat{Y} = 0.5944X + 8.573$$

LA REPRESENTACIÓN GRÁFICA ES LA SIGUIENTE:



AHORA PARA ESTIMAR LA RESISTENCIA EN LOS DÍAS QUE NOS PIDEN, SE USARÁ LA ECUACIÓN DE REGRESIÓN "$\widehat{\boldsymbol{Y}} = \mathbf{0.5944}\boldsymbol{X} + \mathbf{8.573}$":

Para el día 18, $\boldsymbol{X} = \mathbf{18}$:

$$\widehat{\boldsymbol{Y}} = (\mathbf{0.5944})(\mathbf{18}) + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{10.6992} + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{19.2722}\ \boldsymbol{MPa}$$

INTERPRETACIÓN: La capacidad del concreto para soportar fuerzas de compresión después de **18** días se estima que es de $\widehat{\boldsymbol{Y}} = \mathbf{19.2722}\ \boldsymbol{MPa}$

Para el día 36, $\boldsymbol{X} = \mathbf{36}$:

$$\widehat{\boldsymbol{Y}} = (\mathbf{0.5944})(\mathbf{36}) + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{21.3984} + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{29.9714}\ \boldsymbol{MPa}$$

INTERPRETACIÓN: La capacidad del concreto para soportar fuerzas de compresión después de **36** días se estima que es de $\widehat{\boldsymbol{Y}} = \mathbf{29.9714}\ \boldsymbol{MPa}$

Para el día 47, $\boldsymbol{X} = \mathbf{47}$:

$$\widehat{\boldsymbol{Y}} = (\mathbf{0.5944})(\mathbf{47}) + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{27.9368} + \mathbf{8.573}$$

$$\widehat{\boldsymbol{Y}} = \mathbf{36.5098}\ \boldsymbol{MPa}$$

INTERPRETACIÓN: La capacidad del concreto para soportar fuerzas de compresión después de **18** días se estima que es de $\widehat{\boldsymbol{Y}} = \mathbf{36.5098}\ \boldsymbol{MPa}$

3.10.2. Ejemplo 2:

El rendimiento(y) de los trabajadores de una constructora después de una quincena de trabajo ha estado disminuyendo debido a los minerales, mientras más edad(x) tengan los trabajadores mayores será la disminución de su rendimiento. Hallar la recta de regresión lineal simple de acuerdo a la siguiente información:

EDAD (X)	27	29	31	35	36	39	43
RENDIMIENTO (Y)	42	49	50	53	69	76	89

Solución:

PERSONAS	EDAD (X)	DISMINUCIÓN DE RENDIMIENTO EN (%) (Y)	X^2	Y^2	XY
1	27	42	729	1764	1134
2	29	49	841	2401	1421
3	31	50	961	2500	1550
4	35	53	1225	2809	1855
5	36	69	1296	4761	2484
6	39	76	1521	5776	2964
7	43	89	1849	7921	3827
Total	**240**	**428**	**8422**	**27932**	**15235**

Utilizamos las fórmulas de los mínimos cuadrados para calcular la pendiente (b_1) y el intercepto (b_0):

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7 * 15235 - 240 * 428}{7 * 8422 - 8422^2} = 2.8988$$

$$b_0 = \frac{\sum y}{n} - \frac{b_1 \sum x}{n} = \frac{428}{7} - \frac{2.8988 * 240}{7} = -38.245$$

Reemplazamos en la fórmula de la recta de mínimos cuadrados:

$$\hat{y} = b_0 + b_1 X_i$$

$$\hat{y} = -38.245 + 2.8988X$$

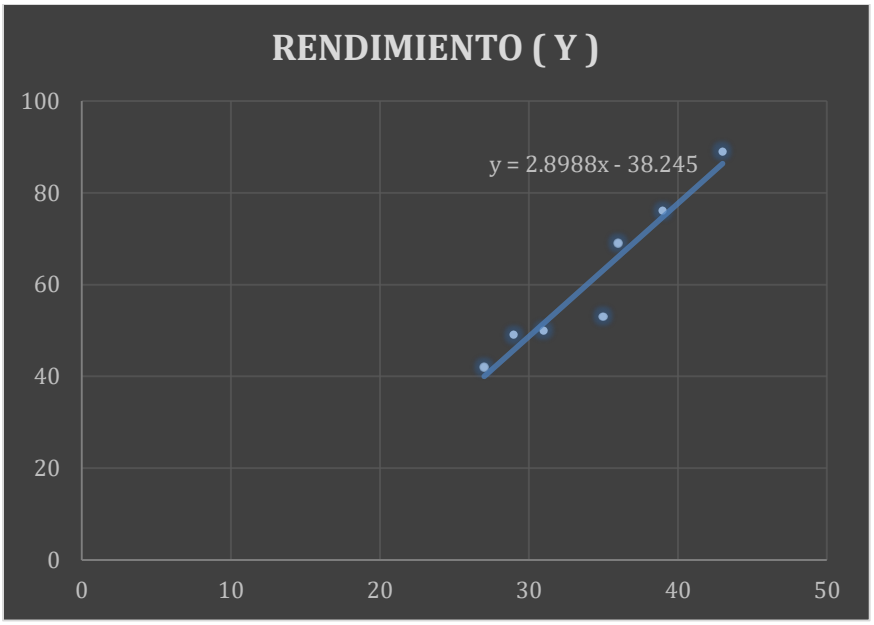
Interpretación:

La constante $b_0 = -38.245$, estima el valor esperado del rendimiento de un trabajador cuando su edad es igual a 0.

La pendiente $b_1 = 2.8988$ señala que al aumentar 1 año a la vida del trabajador el nivel medio de rendimiento aumenta en 2.8988.

La relación que se forma entre el desempeño de un trabajador (Y) y su edad (X)se representa a través de la gráfica de regresión estimada: $\hat{y} = -38.245 + 2.8988X$

REPRESENTACIÓN GRÁFICA:



Nota. La recta mostrada puede ser utilizada para pronosticar o calcular el valor esperado de la disminución del rendimiento a través de la edad del trabajador.

Por ejemplo, si la edad de un trabajador es 37, el modelo pronostica un rendimiento medio de: $\hat{y}(37) = -38.245 + 2.8988 * (37) = 69.01$

3.10.3. Ejemplo 3:

Aplicando la ecuación, la cual establece una relación entre la edad de los trabajadores y su rendimiento, es probable que necesitemos estimar el rendimiento de un trabajador de 28 años:
$\hat{y}(28) = -38.245 + 2.8988 * (28) = 42.92$

Para un valor de x = 28 se presenta un valor estimado de y=42.92

En otro caso:

Con el mismo ejemplo del rendimiento se desea calcular el rendimiento de un trabajador que tiene 50 años de edad:
$\hat{y}(50) = -38.245 + 2.8988 * (50) = 106.695$
Para un valor de x=50 se presenta un valor estimado de y = 106.695
Esto es un caso de extrapolación, ya que se está pronosticando el rendimiento de una persona que tiene una edad que está fuera del intervalo de los datos existentes

3.10.4. Ejemplo 4:

En un proyecto de construcción de carreteras, se desea analizar la relación entre el **tráfico promedio diario (TPD)** y el **nivel de deterioro del pavimento (NDP)**, medido en una escala del 0 al 100, donde 100 representa un pavimento completamente deteriorado. Los ingenieros desean predecir el nivel de deterioro basándose en el tráfico promedio diario para planificar intervenciones de mantenimiento.

Se han tomado muestras en 10 tramos diferentes de la carretera, recopilando datos del tráfico promedio diario y el nivel de deterioro del pavimento en cada tramo.

Tramo	Tráfico Promedio Diario (TPD) (X)	Nivel de Deterioro del Pavimento (NDP) (Y)
1	1200	35
2	1500	42
3	1800	50
4	2000	58
5	2200	63
6	2500	70
7	2700	75
8	3000	80
9	3200	85
10	3500	90

a) Encuentra la ecuación de regresión lineal simple que relacione el tráfico promedio diario (x) con el nivel de deterioro del pavimento (y). Interpreta los coeficientes obtenidos.
b) Calcula el Coeficiente de correlación ($\boldsymbol{r}$) para analizar qué tan fuerte se relaciona el nivel de deterioro del pavimento con el tráfico promedio diario.
c) Utiliza la ecuación de regresión para estimar el nivel de deterioro del pavimento cuando el tráfico promedio diario es de 2150 vehículos y calcula el error típico de la estimación.

SOLUCIÓN:

a) Para encontrar la recta de Regresión Lineal Simple utilizamos el modelo matemático de mínimos cuadrados.

   Calculamos los promedios de los valores de X e Y:

$$\bar{X} = \frac{\sum_{i=1}^{n} x}{n}$$

$$\bar{X} = \frac{1200 + 1500 + 1800 + 2000 + 2200 + 2500 + 2700 + 3000 + 3200 + 3500}{10}$$

$$\bar{X} = 2360$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} y}{n}$$

$$\bar{Y} = \frac{35 + 42 + 50 + 58 + 63 + 70 + 75 + 80 + 85 + 90}{10}$$

$$\bar{Y} = 64.8$$

Dada la siguiente tabla:

Tramo	**(TPD) (x)**	**(NDP) (y)**	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$
1	1200	35	-1160	-29.8	34568	1345600
2	1500	42	-860	-22.8	19608	739600
3	1800	50	-560	-14.8	8288	313600
4	2000	58	-360	-6.8	2448	129600
5	2200	63	-160	-1.8	288	25600
6	2500	70	140	5.2	728	19600
7	2700	75	340	10.2	3468	115600
8	3000	80	640	15.2	9728	409600
9	3200	85	840	20.2	16968	705600
10	3500	90	1140	25.2	28728	1299600
Total	23600	648	0	0	124820	5104000

Calculamos la pendiente (b_1) y el intercepto (b_0) utilizando las fórmulas del método de mínimos cuadrados:

$$b_1 = \frac{\sum(X - \bar{X})\,(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{124820}{5104000} = 0.0245$$

$$b_0 = \bar{Y} - b_1\bar{X} = 64.8 - 0.0245 * 2360 = 7.0854$$

La ecuación es de la forma: $\hat{y} = b_0 + b_1 x$

Reemplazamos los valores obtenidos:

$$\hat{y} = 7.0854 + 0.0245x$$

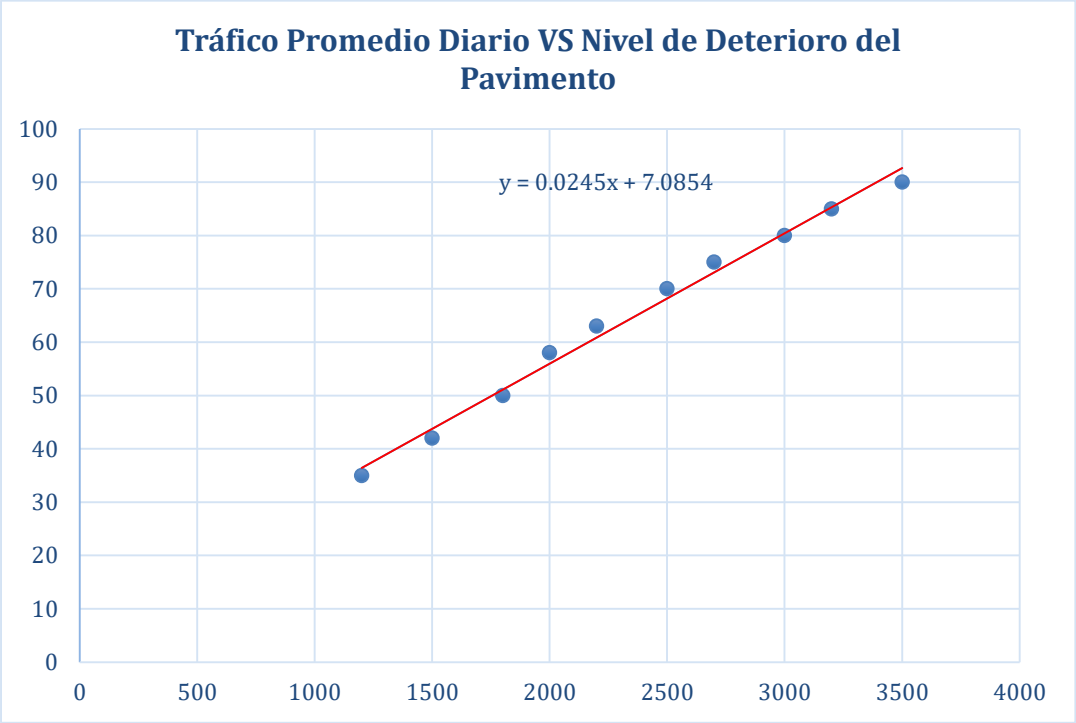
Interpretación:

La inclinación $b_1 = 0.0245$ da a conocer que si el tráfico promedio diario aumenta en una unidad el nivel de deterioro del pavimenta en 0.0245.

La constante $b_0 = 7.0854$, estima el valor esperado del nivel de deterioro cuando el tráfico promedio diario es igual a 0.

La relación que se establece entre el nivel de deterioro del pavimento (Y) y el tráfico promedio diario (X) se representa a través de la ecuación de la recta de regresión estimada: $\hat{Y} = 7.0854 + 0.0245X$

GRÁFICO DE DISTRIBUCIÓN NORMAL:



b) Dada la siguiente tabla:

Tramo	(TPD)(x)	(NDP)(y)	xy	x^2	y^2
1	1200	35	42000	1440000	1225
2	1500	42	63000	2250000	1764
3	1800	50	90000	3240000	2500
4	2000	58	116000	4000000	3364
5	2200	63	138600	4840000	3969

6	2500	70	175000	6250000	4900
7	2700	75	202500	7290000	5625
8	3000	80	240000	9000000	6400
9	3200	85	272000	10240000	7225
10	3500	90	315000	12250000	8100
Total	**23600**	**648**	**1654100**	**60800000**	**45072**

Utilizamos la fórmula del coeficiente de correlación (r):

$$\boldsymbol{r} = \frac{Cov(X,Y)}{\sigma_X * \sigma_Y}$$

Determinamos la covarianza $Cov(X,Y)$:

$$Cov(X,Y) = \frac{(23600 - 2360)(648 - 64.8)}{10} = \frac{(23600 - 2360)(648 - 64.8)}{10}$$

$$Cov(X,Y) = 1238716.8$$

Determinamos la desviación estándar σ_X y σ_Y:

$$\sigma_X = \sqrt{\frac{\sum(X_i - \bar{X})}{n}} = \sqrt{\frac{(23600 - 2360)}{10}}$$

$$\sigma_X = 6716.6778$$

$$\sigma_Y = \sqrt{\frac{\sum(Y_i - \bar{Y})}{n}} = \sqrt{\frac{(648 - 64.8)}{10}}$$

$$\sigma_Y = 184.4240$$

Remplazamos los valores en (r):

$$\boldsymbol{r} = \frac{1238716.8}{(6716.6778)(184.4240)} \Rightarrow \boldsymbol{r} = 1$$

Interpretación:

El valor de 1 determina una conexión extremadamente fuerte. Cuanto más cercano está el coeficiente de 1 o -1, más fuerte es la relación. Significa que hay una relación positiva entre las variables, es decir, cuando una variable aumenta, la otra también tiende a aumentar.

c) Utilizamos la ecuación de regresión calculada anteriormente para determinar el nivel de deterioro del pavimento cuando el tráfico promedio diario es de 4000 vehículos:

$$\hat{y} = 7.0854 + 0.0245x$$

$$\hat{y}(2150) = 7.0854 + 0.0245 * (2150) = 59.76$$

Interpretación: con un tráfico de 2150 vehículos diarios, el nivel de deterioro estimado del pavimento es de aproximadamente 59.76 %.

Calculamos el SEE de la siguiente manera:

$$SEE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

Reemplazamos los datos de acuerdo a los datos de esta tabla:

Tramo	**Tráfico Promedio Diario (TPD)(x)**	**Nivel de Deterioro del Pavimento (NDP)(y)**	$\hat{Y}$	$Y - \hat{Y}$
1	1200	35	36.4854	-1.4854
2	1500	42	43.8354	-1.8354
3	1800	50	51.1854	-1.1854
4	2000	58	56.0854	1.9146
5	2200	63	60.9854	2.0146
6	2500	70	68.3354	1.6646
7	2700	75	73.2354	1.7646
8	3000	80	80.5854	-0.5854
9	3200	85	85.4854	-0.4854
10	3500	90	92.8354	-2.8354
Total	**23600**	**648**	**649.054**	**-1.054**

$$SEE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

$$SEE = \sqrt{\frac{(-1.054)^2}{10 - 2}}$$

$$SEE = 0.373$$

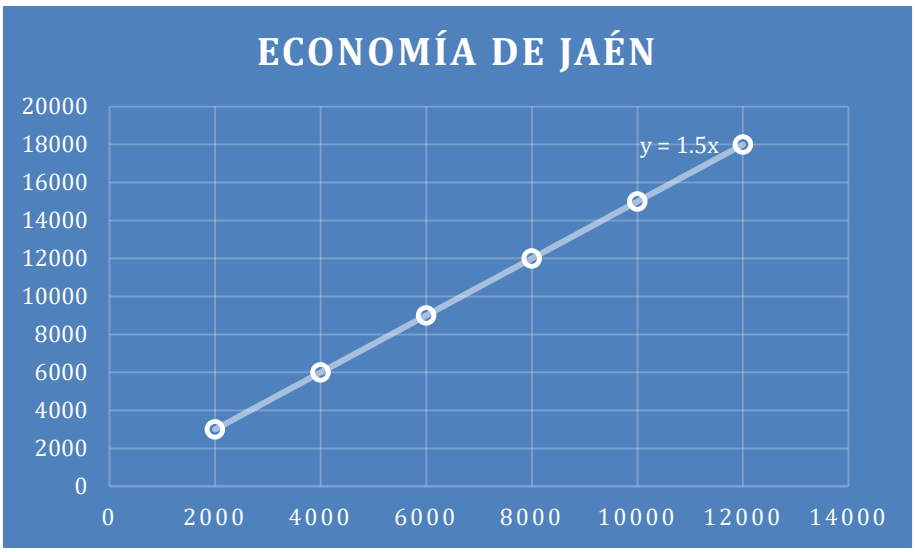
INTERPRETACIÓN: Los valores que se observan de los pronosticados por la ecuación de Regresión se desvían en promedio 0.073.

3.10.5. Ejemplo 5:

Una empresa que construye viviendas en Jaén, Perú; ha identificado de que su flujo de ingresos por proyectos de construcción depende de la economía local de Jaén. Los ingresos de la empresa y la economía local de Jaén durante los últimos 6 años se muestran en la tabla.

Ingresos (en soles) (y)	Economía de Jaén (en soles) (x)	X^2	XY	Y^2
12000	18 000	$324*10^6$	$216*10^6$	$144*10^6$
10 000	15 000	$225*10^6$	$150*10^6$	$100 * 10^6$
8 000	12 000	$144*10^6$	$96*10^6$	$64 * 10^6$
6 000	9 000	$81*10^6$	$54*10^6$	$36 * 10^6$
4 000	6 000	$36*10^6$	$24*10^6$	$16 * 10^6$
2 000	3 000	$9*10^6$	$6*10^6$	$4 * 10^6$
$\bar{Y} = 7000$	$\bar{X} = 10\ 500$	$819*10^6$	$546*10^6$	$364 * 10^6$

DIAGRAMA DE DISPERSIÓN:



CORRELACIÓN POSITIVA PERFECTA

Ecuación de la recta $Y = a + bx$

Ordenada $a = Y - bx$

Pendiente de la recta de regresión

$$\beta_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{6 * 546 * 10^6 - 63000 * 42000}{6 * 819 * 10^6 - 63000^2} = 0.67$$

Intercepto:

$$\beta_0 = \frac{\sum y}{n} - \frac{\beta_1 \sum x}{n} = \frac{42000}{6} - \frac{0.67 * 63000}{6} = -35$$

Reemplazando en la ecuación d la recta de los mínimos cuadrados:

$$y = \beta_0 + \beta_1 X_i$$

$$\hat{y} = -35 + 0.67\text{x}$$

Utilizando el modelo, estimamos $\widehat{\boldsymbol{Y}}$ y calculamos los residuos para cada valor de X dada la siguiente tabla:

Ventas (en soles) (y) (variable dependiente)	Nómina (en soles) (x) (variable independiente)	$\widehat{\boldsymbol{Y}}$**(Mpa)**	**Residuo** $\boldsymbol{e_i = Y_i - \widehat{Y}_i}$	$\boldsymbol{e^2}$
12000	18 000	12 025	-25	625
10 000	15 000	10 015	-15	225
8 000	12 000	8005	-5	25
6 000	9 000	5 995	5	25
4 000	6 000	3 985	15	225
2 000	3 000	1 975	25	625
$\bar{Y} = 7000$	$\bar{X} = 10\ 500$		$\sum(Y_i - \hat{Y}_i) = 0$	

CALCULAMOS EL ERROR ESTÁNDAR RESIDUAL:

$$Se = \sqrt{\frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}}$$

REEMPLAZAMOS LOS DATOS EN LA ECUACIÓN:

$$Se = \sqrt{\frac{0^2}{6-2}} = \sqrt{0} = 0$$

Interpretación

El error estándar de estimación es $Se \approx 0$ Esto indica que, en promedio, los valores registrados de las ventas se mantienen.

3.10.6. Ejemplo 6:

Se tiene una base de datos sobre la relación entre la humedad de un suelo en $(\%)$ que afecta su cohesión medido en (KPa) en un suelo areno. Esta propiedad es importante en estudios de estabilidad de taludes y cimentaciones, por ello usando la regresión lineal simple podemos predecir la cohesión en función de la humedad del suelo.

a) ¿Determina la deviación estándar?
b) ¿Qué porcentaje de variabilidad de Cohesión del suelo es explicado por el modelo de regresión lineal simple?

Solución:

Inciso a):

$SEE = s_{y.x}$: Es el error estándar de estimación o también conocido como desviación estándar.

- **Variable independiente (X)**: Contenido de humedad del suelo en porcentaje (%).
- **Variable dependiente (Y)**: Cohesión del suelo en kPa.

Contenido de humedad (%)	Cohesión (kPa)
5	12
10	10
15	8
20	6
25	5
30	4
35	3
40	2

SOLUCIÓN:

Paso 1: Calculamos cada valor de la ecuación general.

promedio ($\bar{X}$):

$$\bar{X} = \frac{\sum_{i=1}^{n} X}{n}$$

$$\bar{X} = \frac{5 + 10 + 15 + 20 + 25 + 30 + 35 + 40}{8} \quad \Longrightarrow \quad \bar{X} = \frac{45}{8}$$

$$\bar{X} = 22.5$$

promedio ($\bar{Y}$):

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y}{n}$$

$$\bar{Y} = \frac{12 + 10 + 8 + 6 + 5 + 4 + 3 + 2}{7} \quad \Longrightarrow \quad \bar{Y} = \frac{25}{4}$$

$$\bar{Y} = 6.25$$

Paso 2: para calcular los valores que se necesita en la ecuación usamos Excel:

Contenido de humedad (%)	Cohesión (kPa)	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	XY	X^2
5	12	-17.5	5.75	-100.625	306.25	60	25
10	10	-12.5	3.75	-46.875	156.25	100	100
15	8	-7.5	1.75	-13.125	56.25	120	225
20	6	-2.5	-0.25	0.625	6.25	120	400
25	5	2.5	-1.25	-3.125	6.25	125	625
30	4	7.5	-2.25	-16.875	56.25	120	900
35	3	12.5	-3.25	-40.625	156.25	105	1225
40	2	17.5	-4.25	-74.375	306.25	80	1600
180	50	0	0	-295	1050	830	5100

Calculamos el valor de la pendiente (b_1):

$$b_1 = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

Remplazamos los valores:

$$b_1 = \frac{-295}{1050}$$

$$b_1 = -0.2810$$

Calculamos el valor del intercepto (b_0):

$$b_0 = \bar{Y} - b_1\bar{X}$$

Remplazamos los valores:

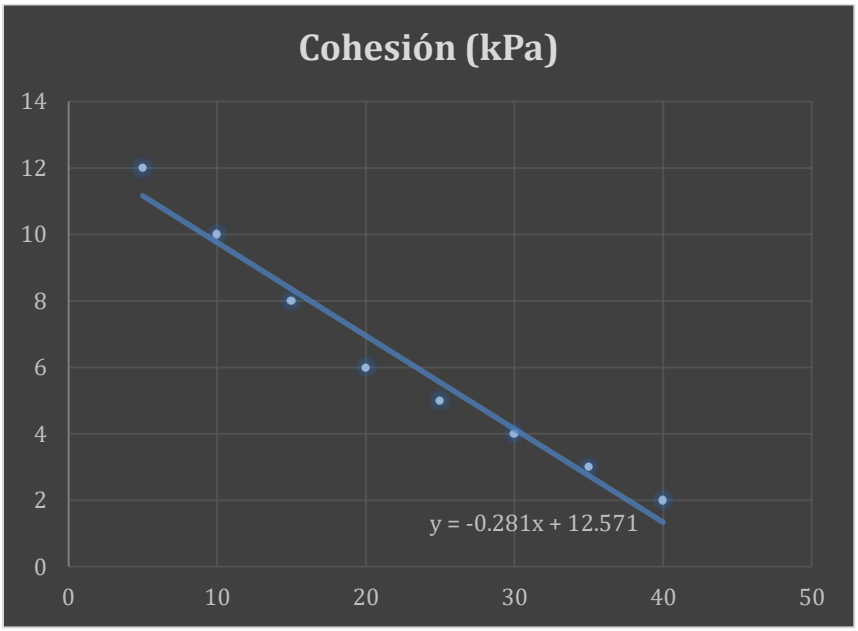
$b_0 = 6.25 - (-0.2810)(22.5)$

$b_0 = 12.5725$

La ecuación de la función queda de la siguiente forma:

$$\hat{y} = 12.5725 - 0.2810X$$

Paso 3: Representación gráfica de la ecuación:



ERROR ESTÁNDAR DE ESTIMACIÓN

$$SEE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

REALIZAMOS LOS CÁLCULOS EN EXCEL:

CONTENIDO DE HUMEDAD (%)	COHESIÓN (kPa)	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
5	12	11.1675	0.8325	0.69305625
10	10	9.7625	0.2375	0.05640625
15	8	8.3575	-0.3575	0.12780625
20	6	6.9525	-0.9525	0.90725625
25	5	5.5475	-0.5475	0.29975625
30	4	4.1425	-0.1425	0.02030625
35	3	2.7375	0.2625	0.06890625
40	2	1.3325	0.6675	0.44555625
180	50	50	0	2.61905

REEMPLAZAMOS LOS DATOS DEL CUADRO EN LA ECUACIÓN:

$$SEE = \sqrt{\frac{\sum(Y - \hat{Y})^2}{n - 2}}$$

$$SEE = \sqrt{\frac{2.61905}{8 - 2}}$$

$$SEE = \sqrt{\frac{2.61905}{6}}$$

$$SEE = \sqrt{0.43651}$$

$$SEE = 0.6607$$

INTERPRETACIÓN**:** los valores observados se desvían con un promedio de 0.06607 de los valores predichos por el modelo de Regresión Lineal.

Inciso b):

Aplicamos la siguiente fórmula:

$$R^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

Remplazamos los datos:

$$R^2 = 1 - \frac{2.61905}{1050} = 0.9975$$

$$R^2 = 99.7506\ \%$$

INTERPRETACIÓN: El 99.7506 % de Cohesión del suelo es explicado por el modelo de regresión lineal simple, es decir cercano a 1. Esto indica que existe un ajuste adecuado en el modelo entre la variable independiente relacionada con la humedad y la variable dependiente correspondiente a la cohesión.

IV. Conclusiones

La regresión lineal simple es una herramienta poderosa para examinar y prever la relación entre dos variables cuantitativas. En los ejemplos presentados, se evidencia que la ecuación de la recta de regresión facilita la predicción de los valores de la variable dependiente en función de las variaciones en la variable independiente. Esto resalta su aplicabilidad en situaciones prácticas, como en la estimación de la resistencia del concreto o el análisis del rendimiento laboral.

El uso de regresión lineal simple, demuestra su aplicación práctica en muchos campos, como para resolver problemas reales en los diferentes campos de aplicación de la ingeniería civil.

El coeficiente R^2, es clave para evaluar la precisión del modelo. En los casos presentados, R^2 se refiere a la capacidad del modelo para interpretar la variabilidad presente en la variable dependiente, lo que permite evaluar su eficacia y precisión., lo que permite validar la efectividad del modelo en los datos analizados.

El error estándar de la estimación permite evaluar la precisión del modelo, mostrando cuánto, en promedio, se desvían los valores observados de los predichos. Los resultados del informe indican que los errores de predicción son mínimos, lo que sugiere una buena capacidad predictiva del modelo en los ejemplos analizados.

V. Bibliografía

Alea, V., Jiménez, E., Muñoz C., Viladomiu N. (2015). *ESTADÍSTICA I: TEORÍA Y EJERCICIOS.* Obtenido el 10 de noviembre del 2024, de https://infolibros.org/pdfview/estadistica-i-teoria-y-ejercicios-victoria-alea-riera-ernest-jimenez-garrido-carme-munoz-vaquer-y-nuria-viladomiu-canela-223/

Barreno, E., Chue, J., Millones, R., Vásquez, F., Castillo, C. (2013). *Estadística Aplicada.* Tarea Asociación Gráfica Educativa. Obtenido el 11 de noviembre del 2024, de https://infolibros.org/pdfview/estadistica-aplicada-emma-barreno-jorge-chue-rosa-millones-felix-vasquez-y-carlos-castillo-223/

Estuardo, G. A. (2012). *ESTADISTÍCA YPROBABILIDADES. Obtenido el 12 de noviembre del 2024, de https://es.scribd.com/doc/241026548/Estadistica-y-Probabilidad-G-a-Estuardo*

Laguna, C. (2014). Correlación y regresión lineal. *Instituto Aragonés de Ciencias de la Salud*, *4*, 1-18.

Novales, A. (2010). *Análisis de Regresión.* Obtenido el 11 de noviembre del 2024, de https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf

Merchan, L. (2024). *Futuro de los trabajos más afectados por la IA ¿que pasará en 10 años?* Obtenido de AllMarket AI.: https://allmarket.ai/blog/futuro-del-trabajo-afectados-ia/

RONALD E. WALPOLE, RAYMOND H. MYERS, SHARON L. MYERS Y KEYING YE. (México). Probabilidad y estadística para ingeniería y ciencias . México: Pearson Educación de México, S.A.

Walpole, R. E., Myers, R. H., Myers, S. L., & Ye, K. (2012). *Probabilidad y estadística para ingeniería y ciencias* (8.a ed.). Pearson Educación.