26

ANÁLISIS ESTADÍSTICO PARA VALIDAR PARÁMETROS DE MODELOS MATEMÁTICOS POR MEDIO MÉTODO DE MÍNIMOS CUADRADOS.

STATISTICAL ANALYSIS TO VALIDATE PARAMETERS OF MATHEMATICAL MODELS BY MEANS OF THE LEAST SQUARES METHOD.

Duver Madroñero Madroñero [1]
Universidad de Nariño

Eduardo Ibarguen Mondragón [2]

Mawency Vergel-Ortega [3]
Departamento de Matemáticas y Estadística, Universidad Francisco de Paula Santander

[1] Departamento de Matemáticas y Estadística, Universidad de Nariño, Pasto, Colombia. Correo: duver1996@udenar.edu.co Orcid: https://orcid.org/0000-0003-1041-3223

[2] Departamento de Matemáticas y Estadística, Universidad de Nariño, Pasto, Colombia. Correo: edbargun@udenar.edu.co Orcid: https://orcid.org/0000-0001-6308-1344

[3] Departamento de Matemáticas y Estadística, Universidad Francisco de Paula Santander, Norte de Santander, Cúcuta, Colombia, correo: mawencyvergel@ufps.edu.co Orcid: https://orcid.org/0000-0001-8285-2968

RESUMEN

En el presente artículo se aborda la estimación de parámetros en estadística frecuentista; por medio de regresiones lineales se estiman parámetros asociados a modelos matemáticos determinísticos, para dicho proceso se utiliza el método de mínimos cuadrados. Además, para la validación de los parámetros estimados mediante el método de mínimos cuadrados, se realiza un análisis estadístico en el cual se incluyen el coeficiente de determinación, desviación estándar, la prueba de t-student, entre otros. En otras palabras, este artículo pretende dar al lector una base de partida para aplicar un método sencillo y muy eficaz para estimar parámetros en modelos matemáticos.

PALABRAS CLAVE : Estimación de parámetros, regresión lineal, mínimos cuadrados, desviación estándar.

ABSTRACT

This paper focus on the estimation of parameters in frequentist statistics. Through linear regressions, parameters associated with deterministic mathematical models are estimated. To this end, the least squares method is used. In addition, for the validation of the estimated parameters by means of the least squares method, a statistical analysis is carried out in which the coefficient of determination, standard deviation, the t-student test are included. In other words, this article aims to give the reader a starting point to apply a simple and very efficient method to estimate parameters in mathematical models.

KEYWORDS : Parameter estimation, linear regression, least squares, standard deviation.

INTRODUCCIÓN

La estimación de parámetros de la estadística frecuentista es habitualmente utilizada en el ámbito profesional como académico. Dado que en ocasiones no se cuenta con un software estadístico para realizar dicho proceso, por tanto, es necesario conocer un método sencillo y eficaz, que permita realizar dicho proceso de manera manual. En este artículo se ha abordado la estimación de los parámetros de la regresión lineal simple, utilizando el método de mínimos cuadrados, el cual, determina la función continua que mejor se ajusta a un conjunto de puntos.

En este caso, la regresión lineal está definida por una recta y los parámetros a estimar mediante el método de mínimos cuadrados son la pendiente y el punto de intersección con el eje y.

Además, la validación de dichos parámetros se realiza por medio de un análisis estadístico, en el cual se incluye algunos conceptos como suma de cuadrados debida al error (SCE), suma total de cuadrados (STC), suma de cuadrados debida a la regresión (SCR), coeficiente de determinación, coeficiente de correlación, desviación estándar, Error cuadrado medio (ECM), error estándar de estimación y una prueba de hipótesis con la t-student, los cuales son de vital importancia para determinar el grado de significancia de los parámetros.

Las estadísticas a las que estamos acostumbrados se denominan estadísticas frecuentistas,

las cuales, se desarrollan a partir de los conceptos de probabilidad y comparación de hipótesis. Su principal objetivo es siempre sacar conclusiones en el marco de la investigación en curso, ya sea en base a la significación estadística o en base a la aceptación y rechazo de hipótesis (Carreño, 2006). Para el análisis estadístico diseñado con el propósito de comparar la eficacia

de nuevos tratamientos frente a otros conocidos, se utiliza la información obtenida en el experimento. Dado que los criterios de decisión son determinados a priori y permanecen estáticos o inalterados durante todo el proceso de investigación, entonces se establece que no existe subjetividad respecto a los parámetros (Carreño, 2006).

A continuación se muestra cómo estimar los parámetros de la regresión lineal, es decir, la intersección y la pendiente de la ecuación lineal .

Regresión lineal: aunque simple, la regresión lineal es una herramienta poderosa para analizar datos. Existen dos tipos de regresión lineal, la menos compleja es donde solo hay dos variables de interés y la segunda tiene un poco más de complejidad, ya que, tiene más de dos variables de interés (Carrasquilla et al, 2016). En este trabajo solo se abordara el caso de la regresión lineal simple.

En estadística, la regresión lineal o ajuste lineal es un modelo matemático utilizado para aproximar la relación de dependencia entre una variable dependiente y, la variable independiente x y un término aleatorio ε (Carrasquilla et al, 2016). El modelo se puede expresar así:

$$y = \beta_0 + \beta_1 x + \varepsilon, \quad (1)$$

donde β_0 y β_1 son los parámetros del modelo, y ε es una variable aleatoria llamada error, que representa la variabilidad de y que no puede ser explicada por la relación lineal entre x y y. A partir de la ecuación (1) se obtiene la ecuación de la regresión lineal simple dada por.

$$E(y) = \beta_0 + \beta_1 x. \quad (2)$$

La grafica de la ecuación (2) es una línea recta, cuya intersección con el eje Y es β_0 y su pendiente es β_1. E(y) es la valor medio o valor esperado de y para un valor de x. Por medio del método de mínimos cuadrados se obtiene la ecuación de la regresión lineal simple estimada dada por

$$\hat{y} = b_0 + b_1 x, \quad (3)$$

donde $\mathbf{b_0}$ y $\mathbf{b_1}$ son valores estimados de $\boldsymbol{\beta_0}$ y $\boldsymbol{\beta_1}$, respectivamente. La gráfica de la ecuación (3) es llamada *recta de regresión estimada*. Estos valores se estiman mediante el método de mínimos cuadrados, el cual se presenta a continuación.

¿Qué es el método de mínimos cuadrados?

Este es un método de análisis numérico que en su forma más simple, intenta minimizar la suma de cuadrados de las diferencias ordenadas, llamadas residuos. Dependiendo del problema a resolver, estas diferencias surgen a partir de los puntos de la función estimada y los datos que se han proporcionado (Molina,2020). Por ejemplo, dado un conjunto de datos (pares ordenados), se busca determinar la función continua que mejor se ajuste o aproxime a los datos, mostrando así visualmente la relación entre los puntos (la función puede ser una recta, una curva cuadrática, cubica, entre otras).

En cualquier problema que surja en la ciencia, es muy conveniente y en ocasiones necesario escribir la relación entre diferentes variables a través de algunas expresiones matemáticas. Por ejemplo, en economía, el costo (C), Los ingresos (I) y las utilidades (U) se pueden relacionar mediante la siguiente formula

$$U=I-C.$$

En física, se puede relacionar la aceleración causada por la gravedad, el tiempo en que un objeto ha estado cayendo y la altura del objeto, esto es:

$$y_f = y_i + v_0 t + \frac{1}{2} g t^2.$$

En la expresión anterior y_i es la altura inicial de dicho objeto y v_0 es la velocidad inicial.

Sin embargo, encontrar una fórmula así, no es tarea fácil. Los profesionales generalmente necesitan procesar grandes cantidades de datos y realizar varios experimentos repetidamente para encontrar la relación entre los diferentes datos. Una forma común de lograr esto es usar los puntos para representar los datos obtenidos en el plano y buscar una función continua que se aproxime de manera óptima a dichos puntos. Uno de los métodos para encontrar la función más cercana a los datos dados es el método de mínimos cuadrados.

Este es un método en el cual se usan los datos muéstrales para hallar la ecuación de la regresión estimada (3). Para que la recta de regresión estimada se ajuste bien a los datos, la diferencia entre el valor observado y el valor estimado debe ser pequeña.

El método de mínimos cuadrados utiliza los datos de muestra para obtener los valores b_0 y b_1, que minimizan la suma de cuadrados de la desviación estándar (diferencias) entre valor el observado de la variable dependiente y_i y el valor estimado de la variable dependiente $\hat{y}_i$. El criterio que se utiliza para el método de mínimos cuadrados es el siguiente:

$$\min \sum (y_i - \hat{y}_i)^2, \qquad (5)$$

Los valores de $\mathbf{b_0}$ y $\mathbf{b_1}$ que minimizan la expresión (2) están dados por.

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad (6)$$

y

$$b_0 = \bar{y} - b_1 \bar{x}. \qquad (7)$$

donde

x_i = Valor de la variable independiente en la observación i

$\bar{x}$ = Media de la variable independiente

$\bar{y}$ = Media de la variable dependiente

n = Número total de observaciones.

La ecuación para encontrar la media para cada variable de datos son las siguientes:

$$\bar{x} = \frac{\sum x_i}{n}, \tag{8}$$

y

$$\bar{y} = \frac{\sum y_i}{n}, \tag{9}$$

En el siguiente ejemplo se utiliza el método de mínimos cuadrados para encontrar los valores de b_0 y b_1.

Ejemplo 1. Encontrar la recta que mejor se ajusta a los datos de la Tabla 1:

Tabla 1. Datos para la estimación de b_0 y b_1.

i	x_i	y_i
1	7	2
2	1	9
3	10	2
4	5	5
5	4	7
6	3	11
7	13	2
8	10	5
9	2	14
	$\sum x = 55$	$\sum y = 57$

El diagrama de dispersión de estos datos es:

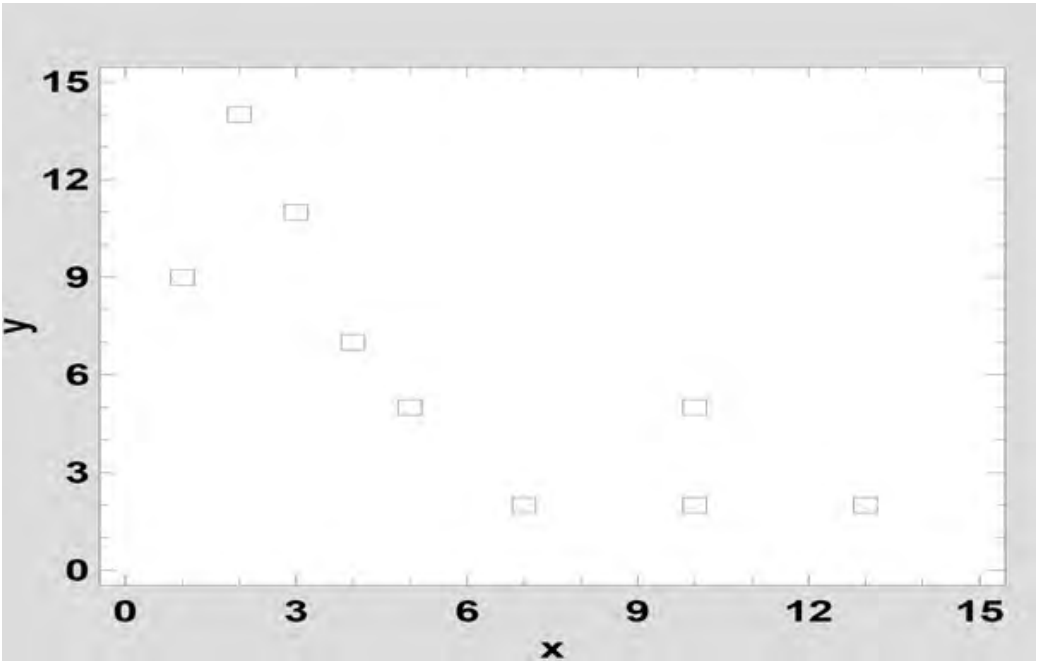

Figura 1. Diagrama de dispersión para los datos de la tabla 1.

Se necesita encontrar la ecuación de la recta (3), donde se aplicará el método de mínimos cuadrados para encontrar b_0 y b_1. Primero, deben usarse las formulas (8) y (9) para calcular la media o promedio de las dos variables. Por lo tanto se tiene.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{7+1+10+5+4+3+13+10+2}{9} = \frac{55}{9} = 6.11111$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{2+9+2+5+7+11++2+5+14}{9} = \frac{57}{9} = 6.33333$$

Además, con el propósito de utilizar el método de mínimos cuadrados se construye la Tabla 2.

Tabla 2. Datos para estimar $\mathbf{b_0}$ y $\mathbf{b_1}$ mediante minimos cuadrados

i	x	y	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	7	2	0.88889	-4.33333	-3.85185	0.79013
2	1	9	-5.11111	2.66667	-13.62964	26.12345
3	10	2	3.88889	-4.33333	-16.85184	15.12347
4	5	5	-1.11111	-1.33333	1.48148	1.23457
5	4	7	-2.11111	0.66667	-1.40741	4.45679
6	3	11	-3.11111	4.66667	-14.51852	9.67901
7	13	2	6.88889	-4.33333	-29.85183	47.45681
8	10	5	3.88889	-1.33333	-5.18517	15.12347
9	2	14	-4.11111	7.66667	-31.51852	16.90123
	55	57			−115.3333	136.88893

A partir de los datos mostrados en la Tabla 2 se verifica que:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-115.3333}{136.88893} = -0.842532$$

$$b_0 = \bar{y} - b_1\bar{x} = 6.33333 - (-0.842532 * 6.11111) = 11.4821.$$

Por tanto, la ecuación de regresión estimada por el método de mínimos cuadrados es:

$$\hat{y}_i = 11.4821 - 0.842532x \qquad (10)$$

La Figura 2 muestra la gráfica de la ecuación (10).

(Ver figura 2)

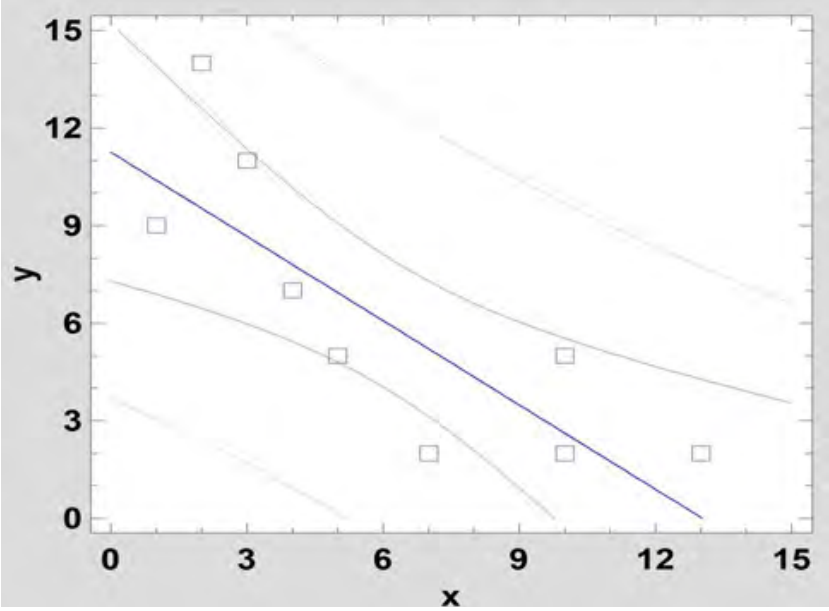

Figura 2. Grafica de la ecuacion de la regresion estimada.

A continuación se presenta un ejemplo de aplicación donde se utiliza el método de mínimos cuadrados para estimar los parámetros de la ecuación de regresión lineal simple.

Ejemplo 2. La ***Tabla 3*** muestra datos sobre las ventas de ropa. Por medio del método de mínimo cuadrados se analizara el comportamiento de las ventas en los dos próximos años.

Tabla 3. Datos de las cantidades vendidas de ropa por cada año.

i	Años (x)	Cantidades vendidas (y)
1	1	220
2	2	245
3	3	250
4	4	258
5	5	273.5
	∑15	1246.5

El diagrama de dispersión de los datos de la ***Tabla 3.***

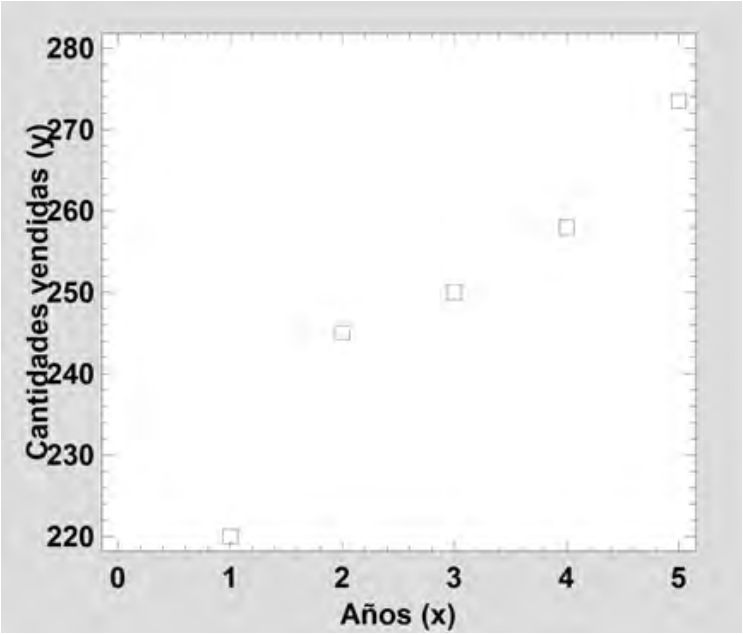

Figura 3. Diagrama de dispersion de las cantidades vendidas versus años.

Los promedios de x y y están dados por.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3.$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{220+245+250+258+273.5}{5} = \frac{1246.5}{5} = 249.3.$$

En la siguiente tabla se presentan los cálculos requeridos para utilizar el método de mínimos cuadrados

Tabla 4. Datos para estimar los parametros b_0 y b_1 de la ecuación de regresión lineal simple, mediante mínimos cuadrados.

i	Años (x)	Cantidades vendidas (y)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	1	220	-2	-29.3	58.6	4
2	2	245	-1	-4.3	4.3	1
3	3	250	0	0.7	0	0
4	4	258	1	8.7	8.7	1
5	5	273.5	2	24.2	48.4	4
$\sum$	15	1246.5			120	10

Los datos de la Tabla 4 se utilizan para estimar b_0 y b_1, los cuales están dados por

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{120}{10} = 12$$

$$b_0 = \bar{y} - b_1\bar{x} = 249.3 - 12 * 3 = 213.3$$

Por tanto, la ecuación de regresión estimada por el método de mínimos cuadrados es

$$\hat{y}_i = 213.3 + 12x \qquad (11)$$

La Figura 4 muestra la gráfica de la ecuación (11).

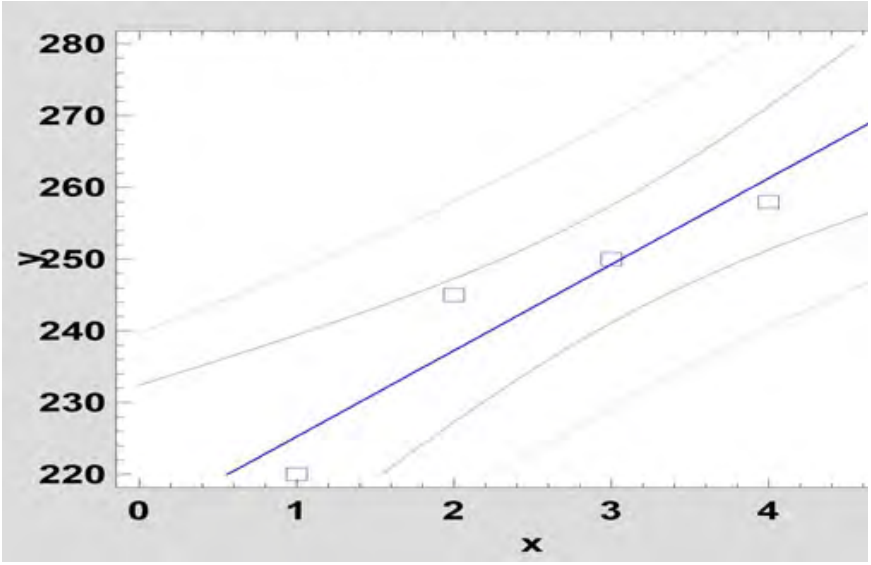

Figura 4. Grafica de la ecuacion de regresion estimada (11)

A partir de la ecuación (11) se estima que para el próximo año; es decir, $x = 6$, el comportamiento de las ventas está dado por

$$\hat{y}_i = 213.3 + 12(6) = 285.3.$$

De manera similar se verifique para el año 7 el comportamiento está dado por

$$\hat{y}_i = 213.3 + 12(7) = 297.3$$

Como podemos observar que la pendiente es positiva se concluye que las ventas tendrán un comportamiento creciente.

En una situación en la que todos los puntos de un diagrama de dispersión estuvieran ubicados en una recta, no habría que preocuparse por encontrar la recta que mejor se ajuste a los puntos del diagrama de dispersión. Solo bastaría con unir lo puntos entre ellos y se obtendría la recta con un buen ajuste. Mas sin embargo, la realidad muestra que esto poco ocurre, entonces en una nube de puntos no solo pasa una recta sino muchas, por tanto, encontrar la recta que mejor se ajuste a los datos es el proceso que se hace con la utilización del método de mínimos cuadrados.

Hasta el momento hemos presentado ejemplos en los cuales se han estimado los valores de los parámetros. Sin embargo, es importante determinar el error en la estimación. Para este fin, introduciremos nuevos conceptos.

Suma de cuadrados debida al error (SCE): La ecuación de la SCE permite calcular el error obtenido cuando se utiliza la ecuación de regresión estimada. La SCE es

$$SCE = \sum(y_i - \hat{y}_i)^2. \qquad (12)$$

El error es

$$\varepsilon = y_i - \hat{y}_i \qquad (13)$$

A continuación se presenta un ejemplo donde se calcula el error y error al cuadrado.

Ejemplo 3: el error (ε) y el error al cuadrado (ε^2) del ejemplo 3 se resumen en la Tabla 7

i	Años (x_i)	Cantidades vendidas (y_i)	$\hat{y}_i = 213.3 + 12x_i$	ε	ε^2
1	1	220	225.3	-5.3	28.09
2	2	245	237.3	7.7	59.29
3	3	250	249.3	0.7	0.49
4	4	258	261.3	-3.3	10.89
5	5	273.5	273.3	0.2	0.04
	$\sum 15$	1246.5			$SCE = 98.8$

Tabla 7. Cálculo del error y error al cuadrado

De la ***Tabla 7*** se concluye que la SCE es 98.8.

Suma total de cuadrados: La suma total de cuadrados (STC) nos permite medir la variabilidad total de una variable dependiente, es decir, mide tanto la parte explicada por el modelo como la parte no explicada por este. La suma total de cuadrados es, de forma muy simple, la variabilidad total de una variable que estamos intentando explicar o estimar.

$$STC = \sum(y_i - \bar{y})^2. \qquad (14)$$

A continuación se presenta un ejemplo para indicar como se calcula STC.

Ejemplo 4: A partir de los datos de la Tabla 3 se calcula la STC, sabiendo que $\bar{y}$=249.3. Por lo tanto, la nueva tabla se ve así:

Tabla 8. Calculo de la suma total de cuadrados.

i	Años (x_i)	Cantidades vendidas (y_i)	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	1	220	-29.3	858.49
2	2	245	-4.3	18.49
3	3	250	0.7	0.49
4	4	258	8.7	75.69
5	5	273.5	24.2	585.64
	15	1246.5		STC = 1538.8

De la Tabla 8 se concluye que la STC para los datos de la Tabla 3 es 1538.8.

Suma de cuadrados debida a la regresión: Para medir cuanto se desvían los valores estimados del promedio $\bar{y}$ se calcula la SCR dada por

$$SCR = \sum(\hat{y} - \bar{y})^2. \quad (15)$$

A continuación se presenta un ejemplo donde se calculará la SCR.

Ejemplo 5: con base en los datos de la Tabla 3, se calcula la SCR, sabiendo que $\bar{y} = 249.3$. Por lo tanto, la nueva tabla se ve así:

(Ver tabla 9)

Tabla 9. Calculo de la SCR

i	Años (x)	Cantidades vendidas (y)	$\hat{y}_i = 213.3 + 12x_i$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
1	1	220	225.3	-24	576
2	2	245	237.3	-12	144
3	3	250	249.3	0	0
4	4	258	261.3	12	144
5	5	273.5	273.3	24	576
$\sum$	15	1246.5			SCR = 1440

Existe una relación entre SCE, STC y SCR la cual está dada por

$$STC = SCR + SCE \tag{16}$$

Esta relación es útil para encontrar una de las tres sumas cuando se conoce las otras dos. A continuación se presenta un ejemplo donde comprueba la ecuación 16

Ejemplo 6: A partir de los valores de SCE, STC y SCR presentados en la Tablas 7, 8 y 9, respectivamente, se verifica la ecuación 15. En efecto SCR = 1440, SCE = 98.8 y STC = 1538.8, luego.

$$STC = SCR + SCE = 1440 + 98.8 = 1538.8.$$

.A continuación se presenta el coeficiente de determinación, el cual permite ver si la relación entre las variables es estadísticamente significativa.

Coeficiente de determinación: se llama coeficiente de determinación al cociente de la suma cuadrada debida a la regresión entre la suma total de cuadrados y se denota por r^2.

$$r^2 = \frac{SCR}{STC}. \tag{17}$$

Este resultado también se puede expresar en forma de porcentaje, luego r^2 se puede interpretar como el porcentaje de la suma total de cuadrados (STC), que se explica mediante el uso de la ecuación de regresión estimada.

Valores grandes de r^2 implican que la recta que encontrada a través del método de mínimos cuadrados se ajusta bien a los datos, es decir, los datos observados están más cerca de la recta de mínimos cuadrados. Sin embargo, el valor de r^2 no es suficiente para garantizar que la relación entre x y y sea estadísticamente significativa, por lo que se debe tener en cuenta otras

consideraciones (Franco, Reyes, & Cuadrado, 2017)

En seguida se presenta un ejemplo donde se calcula el coeficiente de correlación .

Ejemplo 8: con los valores del Ejemplo 2 se calculó STC y SCR, donde se obtuvo los siguientes valores 1538.8 y 1440 respectivamente. Calcular el coeficiente de determinación r^2.

$$r^2 = \frac{SCR}{STC} = \frac{1440}{1538.8} \approx 0.936.$$

Por tanto, del ejemplo anterior se puede concluir que el 93.6% de la variabilidad en las ventas se explica por la relación lineal que existe entre los años y la cantidad de ventas.

Este resultado aunque útil no es suficiente para concluir que las dos variables tiene una relación estadísticamente significativa.

Coeficiente de correlación: El rango del coeficiente de correlación es [-1,1], cuando $r_{xy}=1$, x y y tiene correlación positiva perfecta, es decir, todos los datos se encuentran en una línea recta que tiene pendiente positiva. Si $r_{xy}=-1$, x y y tiene correlación negativa perfecta, es decir, todos los datos se encuentran en una línea recta que tiene pendiente negativa. Los valores del coeficiente de correlación cercano o igual a cero indican que x y y no están relacionadas linealmente. Se calcula mediante la siguiente ecuación.

$$r_{xy} = (\text{signo de } b_1)\sqrt{r^2}. \qquad (17)$$

El signo del coeficiente de regresión muestral es positivo si la ecuación de regresión estimada tiene pendiente positiva ($b_1>0$); y es negativo si la ecuación de la regresión estimada tiene pendiente negativa ($b_1<0$).

Un método alternativo para el análisis de significancia es la prueba de hipótesis, la cual se presenta a continuación.

Prueba de significancia: en una ecuación de regresión lineal simple, la media aritmética de y es una función lineal de x:

$$E(y) = \beta_0 + \beta_1 x$$

Las variables x y y satisfacen las siguientes propiedades.

1 Si $\beta_1 = 0$, entonces x y y no están relacionadas linealmente.
2 Si $\beta_1 \neq 0$, entonces x y y están relacionadas linealmente.

Si se cumple la segunda condición, se debe probar que existe una relación de regresión significante, por lo tanto, el proceso a seguir es realizar una prueba de hipótesis para determinar si el valor de β_1 es diferente de cero. Una de las pruebas que se puede utilizar es la de t-student. Cabe aclarar que otras pruebas también permiten este proceso. En este trabajo se utiliza la t-student. Por tanto, es

necesario estimar la varianza σ^2 del error ε en el modelo de regresión.

A continuación se presenta como estimar la varianza σ^2 del error.

Estimación de σ^2: la varianza del error ε, también representa la varianza de los valores de y respecto a la recta de regresión estimada. Dado que las desviaciones de los valores de y con respecto a la recta de regresión estimada son residuos, entonces SCE es una medida de variabilidad de las observaciones reales respecto a la recta de regresión estimada.

Error cuadrado medio (ECM) (Estimación de σ^2): el ECM de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima.

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n-2}. \qquad (18)$$

En la ecuación 18 se divide entre $n-2$ debido a los grados de libertad que tiene SCE, ya que para calcular SCE se necesita estimar dos parámetro (β_0 y β_1). Como el valor de ECM proporciona una estimación de σ^2, se puede emplear la notación de s^2.

A continuación se da un ejemplo, en el cual se calcula el error cuadrado medio.
Ejemplo 9: Con el resultado del Ejemplo 4 para SCE y $n = 5$, calcular el error cuadrado medio.

$$s^2 = \text{ECM} = \frac{\text{SCE}}{n-2} = \frac{98.8}{3} = 32.9.$$

Este valor no se interpreta.

A partir de s^2 se estima σ, usando el ECM.

Ejemplo 10: Calcular el error estándar de estimación a partir de s^2 del Ejemplo 9.

$$s = \sqrt{\text{ECM}} = \sqrt{\frac{\text{SCE}}{n-2}} = \sqrt{32.9} = 5.74.$$

El error estándar de estimación se utiliza en la prueba de significancia de la relación entre x y y.

Prueba t-Student: Dado que el modelo de regresión lineal simple es $y = \beta_0$ $\beta_1 x + \varepsilon$, donde x y y se relacionan linealmente, entonces $\beta_1 \neq 0$. El propósito de us la prueba t-student es determinar si realmente se puede concluir que $\beta_1 \neq 0$.

Para este fin se define la hipótesis nula H_0 y la hipótesis alternativa H_a dadas por

$$H_0: \beta_1 = 0.$$

$$H_a: \beta_1 \neq 0.$$

Si no se acepta la hipótesis nula, se puede concluir que $\beta_1 \neq 0$, además que x y y tienen una relación estadísticamente significativa. El fundamento principal para esta prueba de hipótesis se deducen de las propiedades de la distribución muestral de b_1, que es el estimador de β_1, el cual obtiene en este caso mediante el método de mínimos cuadrados (Sánchez, 2015). A continuación se presenta las propiedades de la distribución muestral de b_1.

- $E(b_1) = \beta_1$
-
- $\sigma_{b_1} = \frac{\sigma}{\sqrt{\sum(x_i - \bar{x})^2}}$
- Los datos se distribuyen normalmente (20)

Dado que el valor esperado de $b_1$es igual a β_1, se puede concluir que $b_1$es un estimador insesgado de β_1.
De esta manera se obtiene el estimador siguiente de σ_{b_1}.

Desviación estándar estimada de b_1: al no conocerse el valor de σ, se busca una estimación de σ_{b_1}, que se denota por s_{b_1} y estimando σ mediante s en la ecuación (20), para obtener la siguiente ecuación.

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} \quad (21)$$

Ejemplo 11: En el Ejemplo 10 calculo $s = 5.74$, y de la Tabla 4 se conoce $\sum(x_i - \bar{x})^2 = 10$, luego la desviación estándar para b_1 es.

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{5.74}{10} = 0.574.$$

Ahora, para verificar el nivel de significancia se utiliza la prueba t de significancia para la regresión lineal simple.

El estadístico de prueba se lo calcula con la siguiente ecuación.

La regla para rechazar o aceptar una de las dos hipótesis dice lo siguiente.

$$H_0 \text{ si } t = \frac{b_1}{s_{b_1}} \leq t_{\alpha/2} \text{ o si } t \geq t_{\alpha/2}. \quad (22)$$

Donde $t_{\alpha/2}$ se toma de la distribución t con $n-2$ grados de libertad
Además se busca el intervalo de confianza para β_1, el cual se lo calcula por.

$$b_1 \pm t_{\alpha/2} s_{b_1}.$$

El estimador puntual es b_1 y el margen de error es $t_{\alpha/2} s_{b_1}$. El intervalo de confianza para este valor es $1-\alpha$ y $t_{\alpha/2}$ es el valor t, que genera un área de $\alpha/2$ en la cola superior de la distribución t con $n-2$ grados de libertad (Montgomery et al, 2021).

Ejemplo 12: Con el valor $b_1 = 12$ del Ejemplo 2 se desea obtener una estimación de β_1 con un intervalo de confianza de 99%. Además, se tiene que $\alpha = 0.01$ y $n-2 = 5-2 = 3$ grados de libertad, entonces a partir de la distribución t-student se obtiene $t_{0.005} = 5.841$. Luego.

$$b_1 \pm t_{\alpha/2} s_{b_1} = 12 \pm 5.841(0.574) = 12 \pm 3.35.$$

En consecuencia, el intervalo de confianza es [8.65 ,15.35].

Empleando $\alpha = 0.01$ como nivel de significancia, Dado que $\beta_1 = 0$, es el valor hipotético, no está comprendido en el intervalo de confianza [8.65, 15.35], se rechaza H_0. Por lo tanto, se acepta la hipótesis alternativa H_a , lo cual implica que existe una relación estadísticamente significativa entre los años y cantidades vendidas.

Aplicación de los resultados en un modelo que describe un proceso de inactivación de virus.

En esta sección se utilizan la teoría estudiada hasta el momento para estimar la tasa de inactivación del virus MS2 y realizar el análisis estadístico para establecer el nivel de la estimación de la constante. Para este fin se utilizan datos presentados en (Ibarguen et al, 2020), ver Tabla 10.

El modelo a utilizar es el siguiente:

$$\frac{dN}{dt} = -kN.$$

Utilizando la regresión lineal simple se pretende estimar k. Se tiene la siguiente tabla

Tabla 10. Datos de inactivación de colifagos de MS2.

i	Tiempo (t,min) t	MS2-10 ($x10^{-3}$) N
0	0	0.975
1	10	0.540
2	20	0.390
3	30	0.237
4	40	0.137
5	50	0.109
6	60	0.0645
7	70	0.0192
Σ	280	2.4717

El diagrama de dispersión para estos datos es.

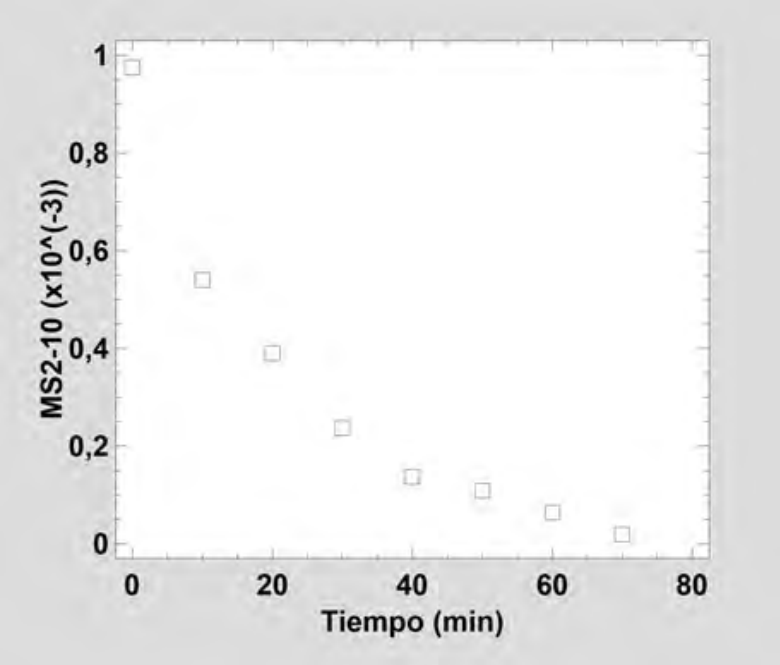

Figura 7. Grafico de dispersion de MS2-10 vs tiempo

Se necesita encontrar la ecuación de la recta (**3**), donde se aplicará el método de mínimos cuadrados para encontrar $\mathbf{b_0}$ y $\mathbf{b_1}$. Primero, deben usarse las formulas (8) y (9) para calcular la media o promedio de las dos variables. Por lo tanto se tiene.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{0 + 10 + 20 + 30 + 40 + 50 + 60 + 70}{8} = 35$$

$$\bar{y} = \frac{\sum y_i}{n} = \frac{0.975 + 0.540 + 0.390 + 0.237 + 0.137 + 0.109 + 0.0645 + 0.0192}{8}$$

$$= \frac{2.4717}{8} = 0.30896$$

Luego le aumentamos cuatro columnas a la Tabla 10, de donde se va sacar lo datos necesarios para aplicar el método de mínimos cuadrados.

Tabla 11. Datos para estimar los parámetros de la regresión lineal.

i	Tiempo (min) t	MS2-10 ($x10^{-3}$) N	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	0	0.975	-35	0.66604	-23.3114	1225
1	10	0.540	-25	0.23104	-5.776	625
2	20	0.390	-15	0.08104	-1.2156	225
3	30	0.237	-5	-0.07196	0.3598	25
4	40	0.137	5	-0.17196	-0.8598	25
5	50	0.109	15	-0.19996	-2.9994	225
6	60	0.0645	25	-0.24446	-6.1115	625
7	70	0.0192	35	-0.28976	-10.1416	1225
$\sum$	280	2.4717			-50.0555	4200

A partir de los datos mostrados en la Tabla 11 se verifica que:

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{-50.0555}{4200} = -0.011918.$$

$$b_0 = \bar{y} - b_1\bar{x} = 0.30896 - (-0.011918) * 35 = 0.72609.$$

Por lo tanto, la ecuación de regresión estimada, por medio del método de mínimos cuadrados es la siguiente:

$$\hat{y}_i = 0.72609 - 0.011918x \quad (23)$$

La pendiente de la ecuación (23) es negativa, implica que a medida que aumenta el tiempo, disminuye MS2-10.

La Figura 8 muestra la gráfica de la ecuación (10).

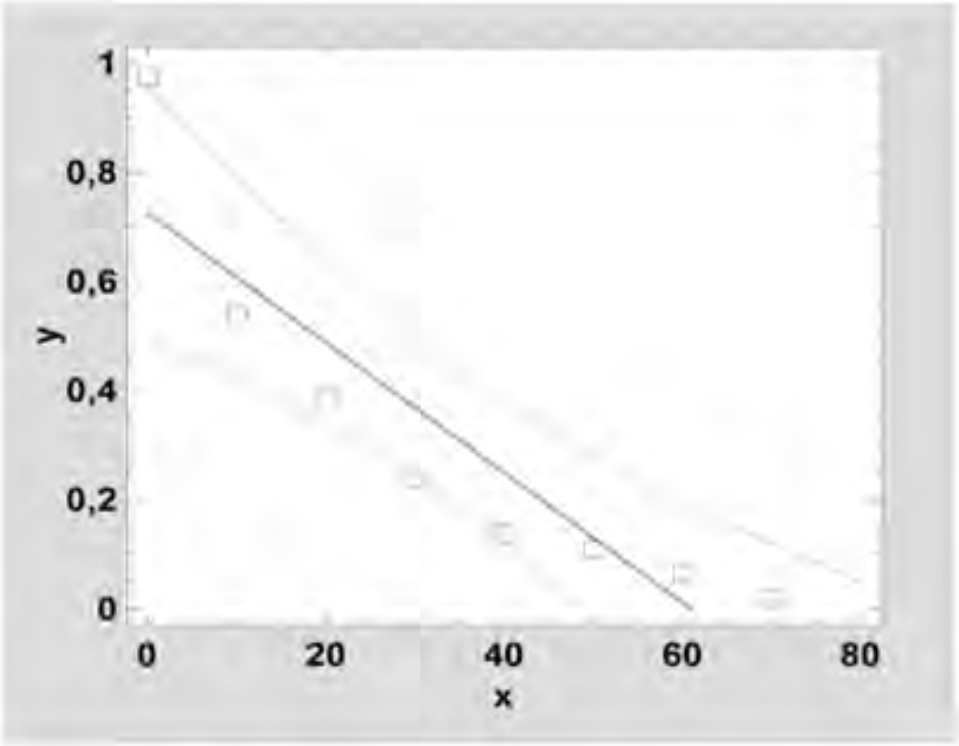

Figura 8. Grafica de la ecuación de la regresión lineal estimada.

Se calcula SCE de la Tabla 10.

(Ver tabla 12)

Tabla 12. Cálculo de la SCE

i	Tiempo (t,min) t	MS2-10 ($x10^{-3}$) N	$\hat{y}_i = 0.72609 - 0.011918x$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
0	0	0.975	0.726	0.249	0.062
1	10	0.540	0.607	-0.067	0.0045
2	20	0.390	0.488	-0.098	0.0096
3	30	0.237	0.369	-0.132	0.0174
4	40	0.137	0.249	-0.112	0.0125
5	50	0.109	0.130	-0.021	0.00044
6	60	0.0645	0.011	0.0535	0.00287
7	70	0.0192	-0.108	0.13	0.0169
$\sum$	280	2.4717			$SCE = 0.126$

La SCE= 0.123 mide el error que existe al utilizar la ecuación de regresión estimada $\hat{y}_i = 0.72609 - 0.011918x$, para predecir

A continuación se va calcular STC de la Tabla 13, dado $\bar{y} = 0.30896$.

Tabla 13. Calculo de STC.

i	Tiempo (t,min) t	MS2-10 ($x10^{-3}$) N	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
0	0	0.975	0.666	0.444
1	10	0.540	0.231	0.053
2	20	0.390	0.081	0.007
3	30	0.237	-0.072	0.005
4	40	0.137	-0.172	0.030
5	50	0.109	-0.19996	0.040
6	60	0.0645	-0.244	0.060
7	70	0.0192	-0.290	0.084
$\sum$	280	2.4717		$STC = 0.723$

Se puede concluir que STC=0.723.

Teniendo en cuenta que $\bar{y} = 0.30896$, se va calcula la SCR de los datos de la Tabla 14.

Tabla 14. Calculo de la SCR.

i	Tiempo (t,min) t	MS2-10 ($x10^{-3}$) N	$\hat{y}_i = 0.72609 - 0.011918x$	$\hat{y} - \bar{y}$	$(\hat{y} - \bar{y})^2$
0	0	0.975	0.726	0.417	0.1739
1	10	0.540	0.607	0.298	0.0888
2	20	0.390	0.488	0.179	0.0320
3	30	0.237	0.369	0.060	0.0036
4	40	0.137	0.249	-0.060	0.0036
5	50	0.109	0.130	-0.179	0.0320
6	60	0.0645	0.011	-0.298	0.0888
7	70	0.0192	-0.108	-0.417	0.1739
$\sum$	280	2.4717			SCR = 0.5966

De la Tabla 14 se obtiene que SCR = 0.5966, esta suma mide que tanto se desvían los valores de $\hat{y}$ del valor de $\bar{y}$.

Utilizando los valores de la STC y SCR previamente calculados, que son 0.723 y 0.5966 respectivamente. Para calcular r^2

$$r^2 = \frac{SCR}{STC} = \frac{0.5966}{0.723} = 0.825$$

Por tanto, se puede concluir que el 82.5% de la variabilidad en el MS2-10 se explica por la relación lineal que existe entre MS2-10 y el tiempo.

A continuación usando $r^2 = 0.825$, calcular el coeficiente de correlación

$$r_{xy} = (\text{signo de } b_1)\sqrt{r^2} = -\sqrt{0.825} = -0.908$$

Por tanto, se puede concluir que el 82.5% de la variabilidad en el MS2-10 se explica por la relación lineal que existe entre MS2-10 y el tiempo.

A continuación usando r^2=0.825, calcular el coeficiente de correlación:

$$r_{xy} = (\text{signo de } b_1)\sqrt{r^2} = -\sqrt{0.825} = -0.908$$

Dado que $r_{xy} = -0.908$, se puede concluir que las dos variables se encuentran bien relacionadas, en una relación lineal negativa.

Teniendo en cuenta que el ECM da una estimación de σ^2, calcular ECM para los datos $SCE = 0.126$ y $n - 2 = 6$.

$$s^2 = ECM = \frac{SCE}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2} = \frac{0.126}{6} = 0.021$$

No se hace una interpretación de este resultado.

Luego, para estimar σ, se calcula el ECM obteniendo el valor de s.

$$s = \sqrt{ECM} = \sqrt{\frac{SCE}{n-2}} = \sqrt{0.021} = 0.145$$

Este resultado se utilizara más adelante en la prueba de significancia de la relación entre x y y.

Ahora utilizando el resultado anterior $s = 0.145$ y de la Tabla 11 tomando el valor $\sum(x_i - \bar{x})^2 = 4200$, luego, se remplazan estos valores en la siguiente ecuación.

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{0.145}{4200} = 0.000035$$

Por tanto la desviación estándar estimada de b_1 es 0.000035

A continuación se presenta la prueba t de significancia para la regresión lineal simple, de la cual, se obtiene el siguiente resultado.

Con el valor de $b_1 = -0.011918$ del se desea tener una estimación de β_1 mediante un intervalo de confianza de 99% de confianza, además, se tiene que $\alpha = 0.01$ y $n - 2 = 8 - 2 = 6$ grados de libertad, por lo tanto, si buscamos el valor de t en la tabla de la distribución t-student, se obtiene que $t_{0.005} = 3.707$. Luego, la estimación mediante un intervalo de 99% de confianza es

$$b_1 \pm t_{\alpha/2} s_{b_1} = -0.011918 \pm 3.707(0.000035) = -0.011918 \pm 0.000129745$$

En consecuencia, el intervalo de confianza es [-0.0120 ,-0.0118.].

Además, empleando $\alpha=0.01$ como nivel de significancia, se puede usar el intervalo de 99% de confianza como alternativa para llegar a la conclusión de la prueba de hipótesis que se obtienen con los valores del ***Ejemplo 12.*** Como 0, que es el valor hipotético de β_1, no está comprendido en el intervalo de confianza -0.0120 a -0.0118 se rechaza H_0 y se concluye que y si existe una relación estadísticamente significativa.

CONCLUSIONES

Se puede ver que si bien hay varios métodos para estimar parámetros, el método de mínimos cuadrados es el más utilizado, ya que es un proceso fácil de realizar a mano cuando los datos no son muchos, mas sin embargo, si los datos son bastantes lo más adecuado es utilizar un software estadístico que haga este proceso por nosotros.

Es posible realizar una comparación entre el valor k obtenido del artículo de (Ibarguen et al, 2020) y el encontrado en este trabajo.

Los resultados son los siguientes: k=0.0443 y -k=-0.011918 entonces k=0.011918, por tanto, se mira que hay una diferencia notable entre los dos resultados, esto puede ser porque el primer valor es calculado con un modelo matemático donde su función es no lineal y para calcular el segundo valor de k se lo calculo con una función lineal.

REFERENCIAS BIBLIOGRÁFICAS

Dagnino S., J. (2014). REGRESIÓN LINEAL. Revista Chilena de Anestesia , 43 (2). https://doi.org/10.25237/revchilanestv43n02.14

Mendenhall, W., Beaver, R. y Beaver, B. (2010). Introducción a la probabilidad y estadística. Cengage Learning (p. 746). Obtenido de http://investigadores.cide.edu/aparicio/data/refs/Mendenhall_Prob_Estadistica_13.pdf%0Ahttps://riunet.upv.es/bitstream/handle/10251/84261/78536109X_TFG_1496841944831665936546568219?

Vinuesa, P. (2017). Tema 9 - Regresión lineal simple y polinomial: teoría y práctica 1 Regresión lineal simple y múltiple: teoría y práctica, 1–33. Obtenido de https://www.ccg.unam.mx/~vinuesa/R4biosciences/docs/Tema9_regresion.html

Molina Arias, M. (2020). La distancia más corta. El método de los mínimos cuadrados. Obtenido de https://anestesiar.org/2020/la-distancia-mas-corta-el-metodo-de-los-minimos-cuadrados/

Carrasquilla-Batista, A., Chacón-Rodríguez, A., Núñez-Montero, K., Gómez-Espinoza, O., Valverde-Cerdas, J., & Guerrero-Barrantes, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. Revista Tecnología En Marcha, 29(8), 33. https://doi.org/10.18845/tm.v29i8.2983

Carreño (Ed.). (2006). La estadística frecuentista y la estadística inferencial. El Teorema de Bayes. En MÉTODOS ESTADÍSTICOS PARA ENFERMERÍA NEFROLÓGICA (pp. 99-106). https://www.revistaseden.org/files/7-CAP%207.pdf

Laguna, C. (2009). CORRELACIÓN Y REGRESIÓN LINEAL Autor: Clara Laguna 4.1 INTRODUCCIÓN. Instituto Aragones De Ciencia De La Salud (pp. 1–18). Retrieved from http://www.ics-aragon.com/cursos/salud-publica/2014/pdf/M2T04.pdf

Weisberg, S. (2005). Applied linear regression (Vol. 528). John Wiley & Sons.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). Introduction to linear regression analysis. John Wiley & Sons.

Ibarguen-Mondragon, E., Revelo-Romo, D., Hidalgo, A., García, H., & Galeano, L. A. (2020).

Mathematical modelling of MS2 virus inactivation by Al/Fe-PILC-activated catalytic wet peroxide oxidation (CWPO). Environmental science and pollution research international, 27(16), 19836–19844. https://doi.org/10.1007/s11356-020-08365-4