REGRESIÓN LINEAL POR MINIMOS CUADRADOS

PROYECTO FINAL

INTEGRANTES:

Priscila Vaca Diez Soliz- 221090096

Rene Eduardo Chungara Martinez-221044191

Adriana Rodríguez Sotelo-217177778

SANTA CRUZ DE LA SIERRA

Bolivia

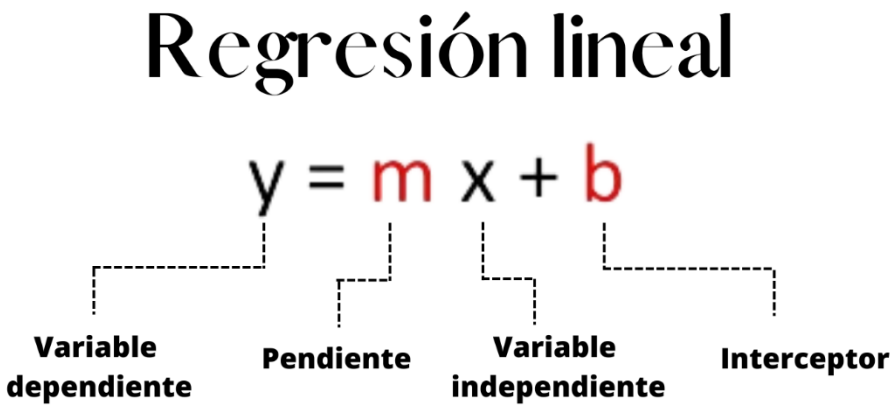
ÍNDICE

Tabla de contenido

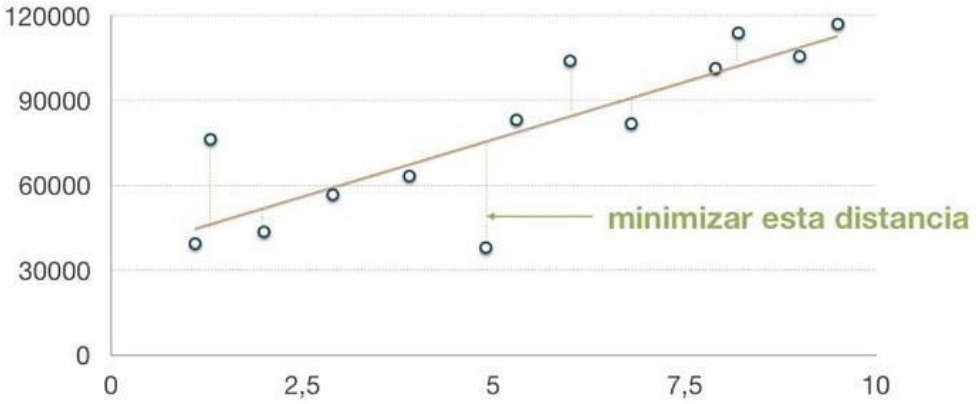
RESUMEN

La Regresión Lineal es una técnica paramétrica utilizada para predecir variables continuas, dependientes, dado un conjunto de variables independientes. Es de naturaleza paramétrica porque hace ciertas suposiciones basadas en el conjunto de datos. Si el conjunto de datos sigue esas suposiciones, la regresión arroja resultados increíbles, de lo contrario, tiene dificultades para proporcionar una precisión convincente.

Matemáticamente, la regresión usa una función lineal para aproximar o predecir la variable dependiente dada como:



Esta es la ecuación de Regresión Lineal Simple. Se llama simple porque solo hay una variable independiente involucrada, que vendría siendo “x”.



El objetivo con Regresión Lineal Simple es minimizar la distancia vertical entre todos los datos y nuestra línea, por lo tanto, para determinar la mejor línea, debemos minimizar la distancia entre todos los puntos y la distancia de nuestra línea. Existen muchos métodos para cumplir con este objetivo, pero todos estos métodos tienen un solo objetivo que es el de minimizar la distancia.

Una forma en que el modelo de regresión encuentre la mejor línea de ajustes es utilizando el criterio de mínimos cuadrados para reducir el error.

El error es una parte inevitable del proceso de predicción, no importa cuán poderoso sea el algoritmo que elijamos, siempre habrá un error irreductible. Sabemos que no podemos eliminar por completo el error, pero aún podemos intentar reducirlo al nivel más bajo. Justamente es en este momento en que se usa la técnica conocida como mínimos cuadrados.

La técnica de mínimos cuadrado intenta reducir la suma de los errores al cuadrado, buscando el mejor valor posible de los coeficientes de regresión.

INTRODUCCIÓN

En la retícula de cada una de estas carreras se encuentra ubicada la materia de Estadística, en dicha materia se presenta el tema de regresión lineal el cual se desarrolla por medio del método de mínimos cuadrados. Se propone realizar una investigación sobre las metodologías de Mínimos Cuadrados y el Método de Gauss-Jordan para la resolución de regresión lineal con el objetivo de ver cual método es más sencillo en su desarrollo, además de que el alumno pueda tener una mejor comprensión del tema y una resolución del problema, obteniendo los parámetros sin errores.

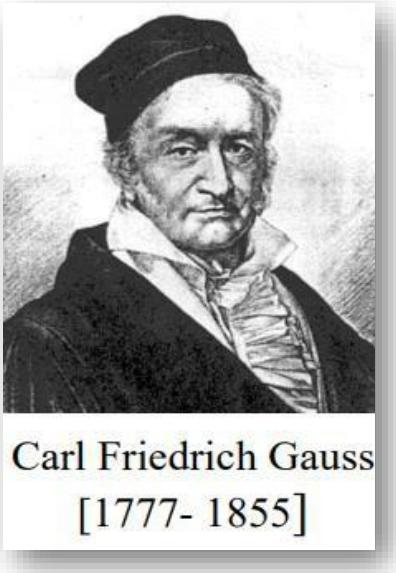
La metodología a seguir es: Analizar los dos métodos con el mismo problema de aplicación y observar que cada uno de ellos llega al mismo resultado, primero se mostrará el desarrollo de los dos métodos para la solución de dos sistemas y después concluir con un caso práctico de regresión lineal por los dos métodos con esta aplicación se puede hacer un análisis del comportamiento de dichos métodos. Los resultados que se obtienen tomando en consideración la elaboración y resultados de los dos métodos es que dichos métodos llegan al mismo valor para cada variable de los coeficientes de las ecuaciones.

Pero al emplear el método de mínimos cuadrados se presenta la desventaja de usar ocho cifras después del punto para lograr un resultado más exacto, además se tiene que llevar a cabo muy minuciosamente todas las operaciones del álgebra lineal elemental, o en su defecto usar una paquetería para poder evitar errores. Cabe mencionar que en el método de la eliminación gaussiana se tiene la ventaja de denotar operaciones más fáciles, pues solo se requiere de sumar, multiplicar y dividir los elementos de la matriz, teniendo mayor exactitud el resolver sistemas de ecuaciones, se puede hacer de forma manual evitando el uso de paquetería.

Atendiendo los resultados obtenidos se concluye que tanto el método de Gauss-Jordan como el método de Mínimos Cuadrados tienen un nivel de precisión y exactitud adecuados en cuanto a solución de sistemas de cualquier orden, la diferencia radica en la labor que implica usar un método u otro haciendo notorio que el método de Gauss-Jordan es menos laborioso que el método de mínimos cuadrados.

Palabras clave: algebra lineal, mínimos cuadrados, Gauss-Jordan, regresión múltiple, sistema de ecuaciones

MARCO TEÓRICO


Carl Friedrich Gauss
[1777- 1855]

1.1. Historia

El método de mínimos cuadrados tiene una larga historia que se remonta a los principios del siglo XIX. En junio de 1801, Zach, un astrónomo que Gauss había conocido dos años antes, publicaba las posiciones orbitales del cuerpo celeste Ceres, un nuevo "pequeño planeta" descubierto por el astrónomo italiano G. Piazzi en ese mismo año.

Desafortunadamente, Piazzi sólo había podido observar 9 grados de su órbita antes de que este cuerpo desapareciese tras del sol. Zach publicó varias predicciones de suposición incluyendo una de Gauss que difería notablemente de las demás. Cuando Ceres fue redescubierto por Zach en Diciembre de1801 estaba casi exactamente en donde Gauss había predicho. Aunque todavía no había revelado su método, Gauss había descubierto el método de mínimos cuadrados.

En un trabajo brillante logró calcular la
órbita de Ceres a partir de un número reducido de observaciones, de hecho, el método de Gauss requiere sólo un mínimo de 3 observaciones y todavía es, en esencia, el utilizado en la actualidad para calcular las órbitas.

1.1. ¿Qué son los Mínimos Cuadrados?

Es un procedimiento de análisis numérico en la que, dados un conjunto de datos (pares ordenados y familia de funciones), se intenta determinar la función continua que mejor se aproxime a los datos (línea de regresión o la línea de mejor ajuste), proporcionando una

demostración visual de la relación entre los puntos de los mismos.

En su forma más simple, busca minimizar la suma de cuadrados de las diferencias ordenadas (llamadas residuos) entre los puntos generados por la función y los correspondientes datos.

1.2. ¿Para qué sirve este método?

Este método se utiliza comúnmente para analizar una serie de datos que se obtengan de algún estudio, con el fin de expresar su comportamiento de manera lineal y así minimizar los errores de la data tomada.

1.3. Definición

Su expresión general se basa en la *ecuación de una recta* **y = mx + b**. Donde m es la pendiente y b el punto de corte, y vienen expresadas de la siguiente manera:

$$m = \frac{n \cdot \Sigma(x \cdot y) - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

$$b = \frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma(x \cdot y)}{n \cdot \Sigma x^2 - |\Sigma x|^2}$$

Donde:

Σ es el símbolo sumatorio de todos los términos,

(x, y) son los datos en estudio, y

n la cantidad de datos que existen.

El método de mínimos cuadrados calcula a partir de los N pares de datos experimentales (x, y), los valores m y b que mejor ajustan los datos a una recta. Se entiende por el mejor ajuste aquella recta que hace mínimas las distancias de los puntos medidos a la recta. Teniendo una serie de datos (x, y), mostrados en un gráfico o gráfica, si al conectar punto a punto no se describe una recta, debemos aplicar el método de mínimos cuadrados, basándonos en su expresión general:

$$y = \left(\frac{n \cdot \Sigma(x \cdot y) - \Sigma x \cdot \Sigma y}{n \cdot \Sigma x^2 - |\Sigma x|^2}\right) x + \left(\frac{\Sigma y \cdot \Sigma x^2 - \Sigma x \cdot \Sigma(x \cdot y)}{n \cdot \Sigma x^2 - |\Sigma x|^2}\right)$$

Ejemplo 1.

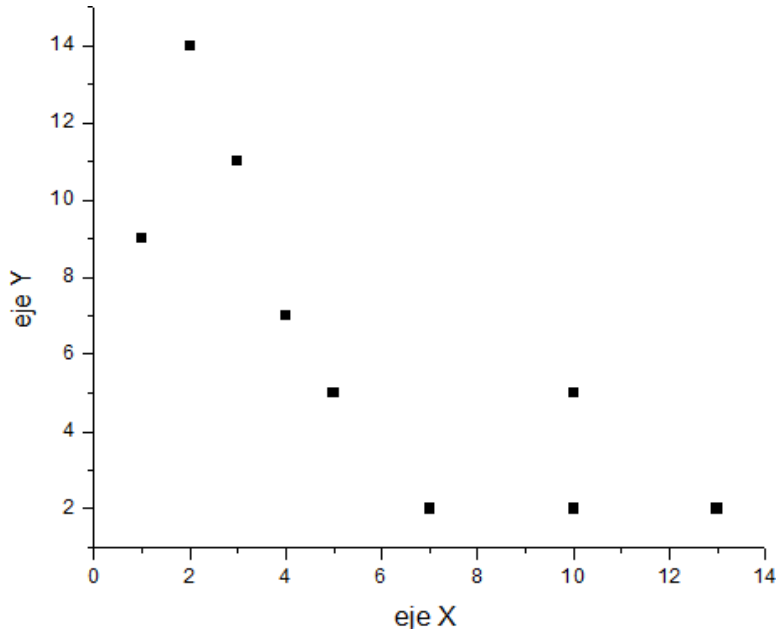
Encontrar la recta que mejor se ajusta a los siguientes datos:

Tabla 1.

n	x	y
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

Siendo su gráfico:

Gráfica 1.



Necesitamos encontrar una recta y = mx + b aplicando el método de mínimos cuadrados.

Debemos encontrar (x*y), (x^2), y la sumatoria de cada columna:

Tabla 2.

n	**x**	**y**	**x*y**	**x^2**
1	7	2	14	49
2	1	9	9	1
3	10	2	20	100
4	5	5	25	25
5	4	7	28	16
6	3	11	33	9
7	13	2	26	16
8	10	5	50	100
9	2	14	28	4
∑	**55**	**57**	**233**	**473**

Sustituimos en cada una de las expresiones:

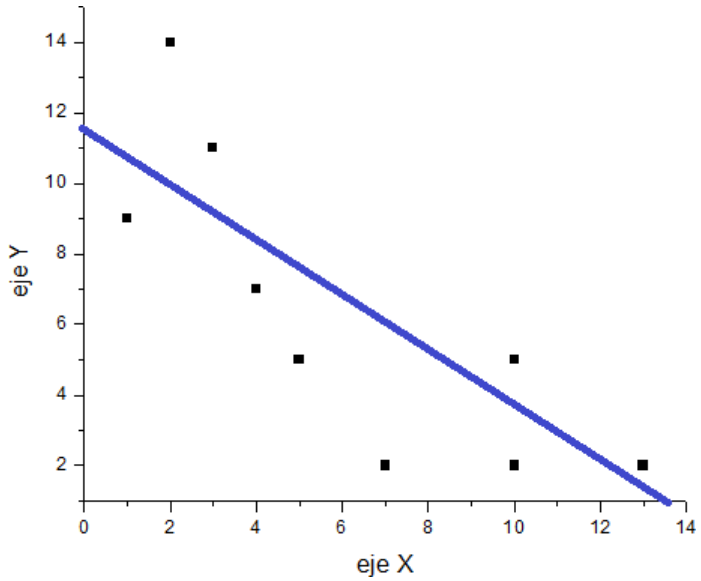
$$m = \frac{(9 \cdot 233 - 55 \cdot 57)}{9 \cdot 473 - |55|^2} = -\frac{1038}{1232} = -0,84$$

$$b = \frac{(57 \cdot 473 - 55 \cdot 233)}{9 \cdot 473 - |55|^2} = \frac{14146}{1232} = 11,48$$

La recta obtenida con el método de los mínimos cuadrados es la siguiente:

$$y = (-0,84) \cdot x + 11,48$$

Gráfico 2.



Vemos que la recta corta al eje y en 11,48 y en el eje x en 13,57. Por lo tanto, si queremos saber dónde corta en el eje x igualamos la ecuación y = 0:

$$0 = (-0,84) \cdot x + 11,48$$

Despejamos x:

$$x = -\frac{11,48}{-0,84} = 13,57$$

Ejemplo 2.

Una empresa que se dedica a la venta de pizzas a domicilio, desea saber la relación que existe entre la publicidad y las ventas de pizzas.

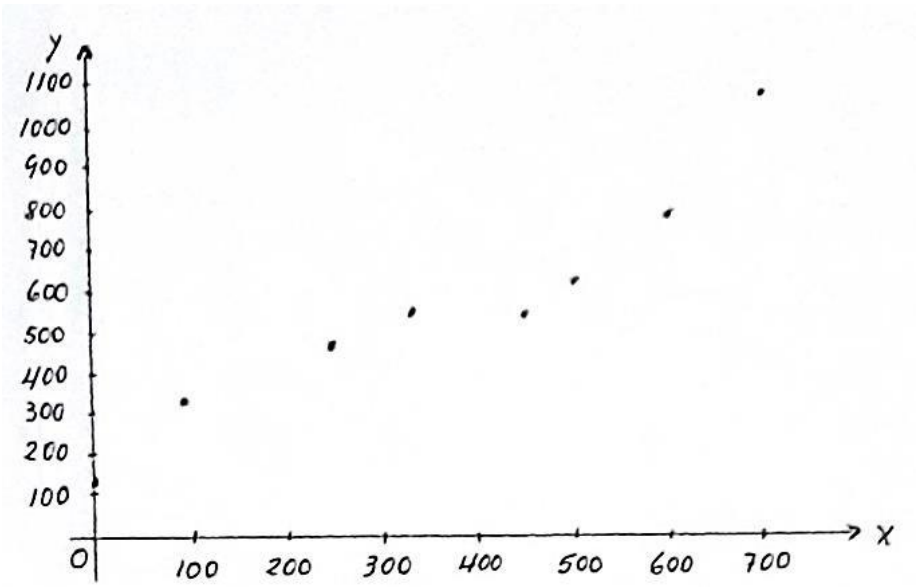
La tabla adjunta muestra los resultados:

Tabla 3.

Gastos en Publicidad (Bs)	**Ventas Mensuales (N)**
0	120
100	350
250	500
350	550
450	550
500	650
600	800
700	1100

El primer paso es analizar, si existe o no relación entre las dos variables:

Publicidad **(X)** y número de pizzas vendidas **(Y)**



Cada punto representa las coordenadas **(x, y)** y podremos observar que a medida que aumenta los gastos de publicidad, aumenta las ventas de pizza, indicando que hay una relación directa.

Como siguiente paso debemos encontrar los valores para m y b

Tabla 4.

n	**x**	**y**	**x*y**	**x^2**
1	0	120	0	0
2	100	350	35000	10000
3	250	500	125000	62500
4	350	550	192500	122500
5	450	550	247500	202500
6	500	650	325000	250000
7	600	800	480000	360000
8	700	1100	770000	490000
∑	**2950**	**4620**	**2775000**	**1497500**

$$m = \frac{8*(2775000)-(2950)*(4620)}{8(1497500)-(2950)^2}$$

$m = 2.615$

$$b = \frac{(4620)(1497500)-(2950)(2775000)}{8(1497500)-(2950)^2}$$

$b = -386.8$

La recta obtenida con el método de los mínimos cuadrados es la siguiente:

$$\boldsymbol{y = 2.615x - 386.8}$$

1.3.1. Ejemplo 3

Ajuste a una línea recta los valores *x* y *y*

Tabla 5.

n	x	y
1	1	0.5
2	2	2.5
3	3	2.0
4	4	4.0
5	5	3.5
6	6	6.0
7	7	5.5

Se calculan las siguientes cantidades:

Tabla 6.

n	**x**	**y**	**x*y**	**x^2**
1	1	0.5	0.5	1
2	2	2.5	5	4
3	3	2.0	6	9
4	4	4.0	16	16
5	5	3.5	17.5	25
6	6	6.0	36	36
7	7	5.5	38.5	49
∑	**28**	**24**	**119.5**	**140**

$\bar{y} = \frac{24}{7} = 3.428571$; $\bar{x} = \frac{28}{7} = 4$

Se reemplazan los valores en las ecuaciones para obtener los resultados

$a_1 = \frac{7(119.5)(28)(24)}{7(140)-(28)^2} = 0.8392857$

$a_o = 3.428571 - 0.8392857(4) = 0.07142857$

Por lo tanto, el ajuste por mínimos cuadrados es ***y = 0.07142857 + 0.839285***

Cuantificación del error en la regresión lineal

Cualquier otra línea diferente a la calculada en el ejemplo anterior, dará como resultado una suma mayor de los cuadrados de los residuos. Así, la línea es única y, en términos de nuestro criterio elegido, es la "mejor" línea a través de los puntos. Varias propiedades de este ajuste se observan al examinar más de cerca la forma en que se calcularon los residuos. Recuerde que la suma de los cuadrados se define como:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i)^2$$

La analogía se puede extender aún más en casos donde 1. la dispersión de los puntos alrededor de la línea es de magnitud similar en todo el rango de los datos, y 2. la distribución de estos puntos cerca de la línea es normal. Es posible demostrar que, si estos criterios se cumplen, la regresión por mínimos cuadrados proporcionará la mejor (es decir, la más adecuada) estimación de a_0 y a_1 (Draper y Smith, 1981).

Esto se conoce en estadística como el principio de máxima verosimilitud. Además, si estos criterios se satisfacen, una "desviación estándar" para la línea de regresión se determina como: $S_{y/x} = \sqrt{\frac{S_r}{n-2}}$, donde a $S_{y/x}$ se le llama error estándar del estimado. El subíndice "y/x" designa que el error es para un valor predicho de y correspondiente a un valor particular de x. También, observe que ahora dividimos entre $n - 2$ debido a que se usaron dos datos estimados (a_0 y a_1), para calcular Sr; así, se han perdido dos grados de libertad.

Otra justificación para dividir entre *n – 2* es que no existe algo como "datos dispersos" alrededor de una línea recta que une dos puntos. De esta manera, en el caso donde n = 2, da un resultado sin sentido, infinito. Así como en el caso de la desviación estándar, el error estándar del estimado cuantifica la dispersión de los datos. Aunque, $S_{y/x}$ cuantifica la dispersión alrededor de la línea de regresión, a diferencia de la desviación estándar original S_y que cuantifica la dispersión alrededor de la media.

Los conceptos anteriores se utilizan para cuantificar la "bondad" de nuestro ajuste. Esto es en particular útil para comparar diferentes regresiones. Para hacerlo, regresamos a los datos originales y determinamos la suma total de los cuadrados alrededor de la media para la variable dependiente (en nuestro caso, y). Esta cantidad se designa por S_t. Ésta es la magnitud del error residual asociado con la variable dependiente antes de la regresión. Después de realizar la regresión, calculamos S_r, es decir, la suma de los cuadrados de los residuos alrededor de la línea de regresión. Esto caracteriza el error residual que queda después de la regresión. Es por lo que, algunas veces, se le llama la suma inexplicable de los cuadrados.

La diferencia entre estas dos cantidades, $S_t - S_r$, cuantifica la mejora o reducción del error por describir los datos en términos de una línea recta en vez de un valor promedio.

Como la magnitud de esta cantidad depende de la escala, la diferencia se normaliza a S_t para obtener: $r^2 = \frac{S_t - S_r}{S_t}$, donde r^2 se conoce como el coeficiente de determinación y r es el coeficiente de correlación ($= \sqrt{r^2}$). En un ajuste perfecto, $\mathbf{S_r = 0}$ y $\mathbf{r = r^2 = 1}$, significa que la línea explica el 100% de la variabilidad de los datos. Si $\mathbf{r = r^2 = 0}$ y $\mathbf{S_r = S_t}$, el ajuste no representa alguna mejora. Una representación alternativa para r que es más conveniente para implementarse en una computadora es:

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

Aunque el coeficiente de correlación ofrece una manera fácil de medir la bondad del ajuste, se deberá tener cuidado de no darle más significado del que ya tiene. El solo hecho de que r sea “cercana” a 1 no necesariamente significa que el ajuste sea “bueno”. Por ejemplo, es posible obtener un valor relativamente alto de r cuando la relación entre y y x no es lineal. Draper y Smith (1981) proporcionan guías y material adicional respecto a la evaluación de resultados en la regresión lineal. Además, como mínimo, usted deberá inspeccionar siempre una gráfica de los datos junto con su curva de regresión.

CONCLUSIÓN

En esta práctica pudimos concluir que nos ayudó a obtener la ecuación de la recta a partir de pares ordenados es decir de los datos, utilizando el método de los mínimos cuadrados y a ser más dóciles en el manejo de las fórmulas, también pudimos determinar que con el error estándar podemos interpretar el resultado para futuros resultados.

LINK AL VIDEO DE LA EXPOSICION:

- https://www.youtube.com/watch?v=mN7NlQTUeKQ