Mínimos Cuadrados – Modelos de regresión Lineal y Cuadrática

Regresión Lineal

Aproximación por rectas que pasan por el origen

A continuación, efectuaremos el cálculo de la pendiente de la recta que pasa por el origen que mejor se aproxima a un conjunto de valores (x_1, y_1), (x_2, y_2), (x_3, y_3) … (x_i, y_i), experimentales.

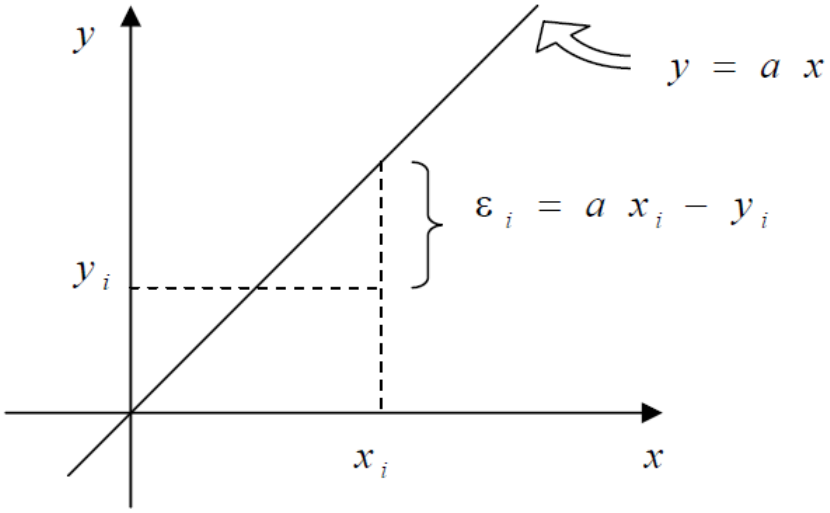
Este procedimiento es de gran importancia debido a que en las experiencias muchas veces las magnitudes físicas dependen linealmente, como por ejemplo, la intensidad de corriente eléctrica es directamente proporcional a la diferencia de potencial en los elementos óhmicos. En reiteradas ocasiones, nos será útil encontrar la pendiente de la recta que mejor aproxime los datos experimentales, debido a que tendrá un importante significado físico. En el ejemplo anterior, la pendiente del gráfico diferencia de potencial (V) en función de la intensidad (I) es la resistencia eléctrica (R) del elemento a estudio. En los casos en los que la relación entre las variables no es lineal, muchas veces se puede linealizar las relaciones para llevarlas a este caso.

Podemos expresar la relación lineal entre ambas magnitudes de la siguiente forma:

$$y = a \cdot x$$

En donde a es la pendiente de la recta, o sea, el valor que deseamos hallar. En el ejemplo anterior, y corresponde a la diferencia de potencial (V), x a la intensidad (I), y a es una constante de proporcionalidad, la cual es igual a la resistencia (R) del elemento. Cuando tratemos datos provenientes de una experiencia, debido a los errores experimentales, generalmente los datos experimentales no satisfacerán exactamente dicha ecuación, sino que estarán próximos a la recta, pero no perfectamente alineados.

Es decir la distancia de cada punto del gráfico a la recta, calculado como $\epsilon = a \cdot x_i - y_i$ no será exactamente cero:



La pendiente de la recta que minimiza la suma de las distancias al cuadrado de los valores

experimentales a la recta (la recta que en cierto modo más se aproxima a los valores experimentales y por ende comete el menor error) tiene como pendiente:

$$a=\frac{\sum_{1}^{N} x_i \cdot y_i}{\sum_{1}^{N} x_i^2}$$

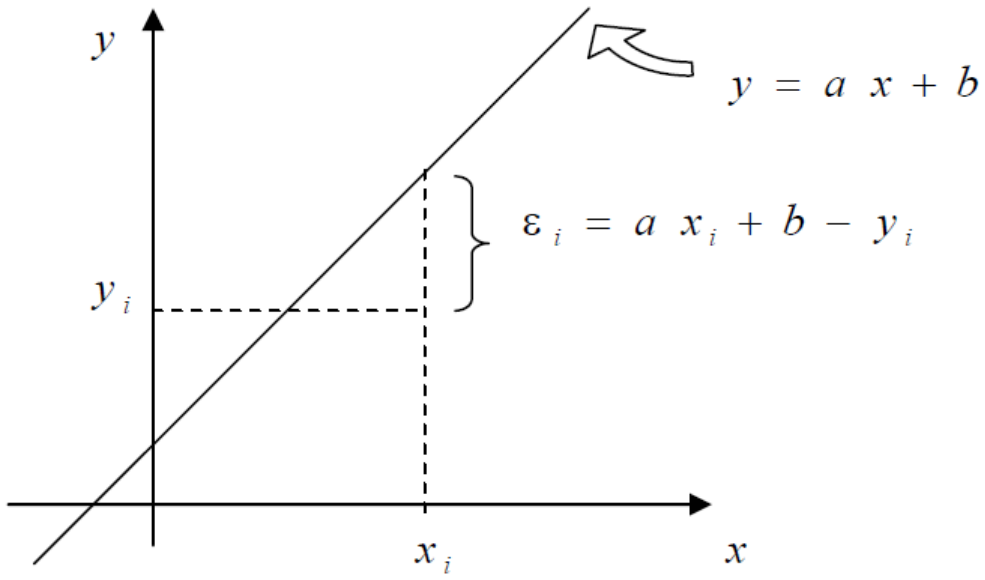
Aproximación por rectas que no necesariamente pasan por el origen

Calculemos ahora la mejor aproximación de un conjunto de valores experimentales (x_1, y_1), (x_2, y_2), (x_3, y_3) … (x_i, y_i) por una recta general, que no necesariamente pase por el origen. Podemos expresar la relación entre ambas magnitudes de la siguiente forma:

$$y=a\cdot x+b$$

en donde *a* es la pendiente de la recta y *b* es el punto de corte de la recta con el eje *y*, o sea, los valores que deseamos hallar.

Procedamos de la misma manera que en el caso anterior.

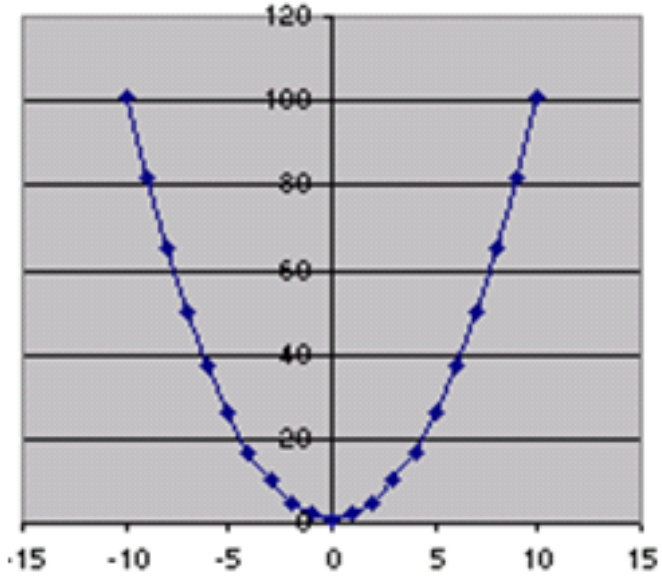


Es decir la distancia de cada punto del gráfico a la recta, calculado como $\epsilon=a\cdot x_i+b-y_i$ no será exactamente cero:

Sin demostrar los valores de pendiente y el punto de corte con el eje y de la recta que minimiza la suma de las distancias al cuadrado de los valores experimentales a la recta (la recta que en cierto modo más se aproxima a los valores experimentales) tienen como expresiones:

$$a=\frac{N\cdot\sum_{i=1}^{N} x_i\cdot y_i-\sum_{i=1}^{N} y_i\cdot\sum_{i=1}^{N} x_i}{N\cdot\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2} \qquad b=\frac{\sum_{i=1}^{N} y_i\cdot\sum_{i=1}^{N} x_i^2-\sum_{i=1}^{N} x_i\cdot y_i\cdot\sum_{i=1}^{N} x_i}{N\cdot\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2}$$

Regresión Cuadrática



La regresión cuadrática es el proceso por el cuál encontramos los parámetros de una parábola que mejor se ajusten a una serie de datos que poseemos, ya sean mediciones hechas o de otro tipo. Bueno, pero por que habríamos de querer ajustar nuestros datos precisamente a una parábola y no a otra función.

Una función cuadrática o de segundo grado se puede representar de manera genérica como :

$$y = a + b \cdot x + c \cdot x^2$$

Entonces lo que nos interesa es encontrar los valores de a, b y c que hacen que el valor de y calculado sea lo mas cercano posible al medido.

Dichos valores se obtienen de resolver el siguiente sistema de ecuaciones:

$$\left\langle \begin{array}{cccc} N & \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} y_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i \cdot y_i \\ \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i^3 & \sum_{i=1}^{N} x_i^4 & \sum_{i=1}^{N} x_i^2 \cdot y_i \end{array} \right|$$

Una vez se haya reemplazado el valor de N, y de las sumatorias, sólo habrá que solucionar el sistema de ecuaciones por su método preferido. Después de que ha solucionado el sistema de ecuaciones entonces tendrá el valor de los parámetros: a,b,c.

Ejemplo:

En determinado proceso se realizaron una serie de 24 mediciones, que luego al graficarse se determinó que es de naturaleza cuadrática. Se desea encontrar los parámetros del polinomio de segundo grado, que mejor se ajusta a esa serie de datos, y cuál es el valor de la variable dependiente, cuando el valor de la variable independiente es de 20.

La tabla con los datos medidos es la siguiente:

X	Y
0	10,08

0,5	12,03
1	11,38

1,5	18,81
2	20,53

2,5	28,5
3	31,38

3,5	38,4
4	48,39
4,5	60,6
5	66,66
5,5	82,61

6	91,37
6,5	105,44
7	122,53
7,5	137,77
8	152,74

8,5	172,65
9	188,84
9,5	207,77
10	230,94
10,5	251,35

11	274,07
11,5	295,95

Ahora, teniendo en cuenta la matriz que dedujimos anteriormente, sabemos que tenemos que encontrar los valores de la suma de x, la suma de x2, de x3, x4, de Yi, xYi, x2*Yi y N=24.

X	Y	X^2	X^3	X^4	Xyi	X^2Yi
0	10,08	0	0	0	0	0
0,5	12,03	0,25	0,13	0,06	6,01	3,01
1	11,38	1	1	1	11,38	11,38
1,5	18,81	2,25	3,38	5,06	28,21	42,31
2	20,53	4	8	16	41,06	82,13
2,5	28,5	6,25	15,63	39,06	71,24	178,11
3	31,38	9	27	81	94,14	282,41
3,5	38,4	12,25	42,88	150,06	134,39	470,36
4	48,39	16	64	256	193,56	774,26
4,5	60,6	20,25	91,13	410,06	272,68	1227,08
5	66,66	25	125	625	333,31	1666,55
5,5	82,61	30,25	166,38	915,06	454,37	2499,02
6	91,37	36	216	1296	548,23	3289,38
6,5	105,44	42,25	274,63	1785,06	685,39	4455,05
7	122,53	49	343	2401	857,74	6004,2
7,5	137,77	56,25	421,88	3164,06	1033,24	7749,32
8	152,74	64	512	4096	1221,9	9775,23
8,5	172,65	72,25	614,13	5220,06	1467,54	12474,08

9	188,84	81	729	6561	1699,55	15295,92
9,5	207,77	90,25	857,38	8145,06	1973,8	18751,13
10	230,94	100	1000	10000	2309,4	23093,97
10,5	251,35	110,25	1157,63	12155,06	2639,18	27711,38
11	274,07	121	1331	14641	3014,81	33162,86
11,5	295,95	132,25	1520,88	17490,06	3403,37	39138,76
Total 138	266,078,166	1081	9522	89452,75	22494,51	208137,88

Reemplacemos los valores en la matriz...

$$\left(\begin{array}{ccc|c} 24 & 138 & 1081 & 2660,8 \\ 138 & 1081 & 9522 & 22495 \\ 1081 & 9522 & 89453 & 208138 \end{array}\right)$$

Resolviendo:

Por lo tanto: a=9,6 b=1,76 c=2,02

la parábola de mejor ajuste es entonces:

$$y = 9,6 + 1,76 \cdot x + 2,02 \cdot x^2$$

Problema 1

La población en los Estados Unidos de América durante el siglo XX ha seguido la evolución indicada en la tabla adjunta, se pide hallar la recta de regresión y pronosticar el número de habitante en al año 2010.

Año	1900	1910	1920	1930	1940	1950	1960	1970	1980	1990	2000
Hab.	75995	91972	105711	123203	131669	150697	179323	203212	226505	249633	281422

Problema 2

Para conocer la relación entre la velocidad de caída de un paracaidista y la fuerza de fricción hacia arriba, se han efectuado las siguientes mediciones

v	1	2	3	4	5
f	5	15,3	29,3	46,4	66,3

Donde v se mide en centímetros por segundo y rozamiento f en 10^6 dinas. Dibuje los puntos de la tabla y aproxime mediante mínimos cuadrados, de forma lineal y cuadrática.

UTN TSEVMA – Mínimos Cuadrados – Prof. Ing. Martin, Milton

AÑOS	1987	1988	1989	1990	1991	1992	1993
GASTOS	21	22	25	26	27	29	30
VENTAS	…	19	20	22	23	24	26

En el informe final de su análisis, deberá responder a las siguientes preguntas:

a) ¿Se incrementarán las ventas del período siguiente al aumentar los gastos en publicidad del período actual?

b) ¿Es adecuado suponer que el ajuste entre estas variables es efectivamente lineal teniendo en cuenta los valores de las variables? Ajuste el modelo lineal e interprete los coeficientes del mismo.

c) ¿Cuál será la predicción de las ventas para 1994?

d) Si para el año 1994 se piensa incrementar los gastos de publicidad en un 10%, ¿qué incremento relativo cabría esperar para las ventas de 1995 con respecto a las de 1994, según el modelo ajustado?

Un estudiante de la UTN Facultad Regional Parana, para poder pagarse sus estudios, debe trabajar como camarero en un bar de su localidad. A este establecimiento, suelen acudir todos los jóvenes de la zona. Este año, con los conocimientos aprendidos, decide por fin estudiar la relación existente entre la cantidad de sal de las galletas saladas y el consumo de bebidas, ya que es costumbre dar al cliente este aperitivo cuando pide una consumición. Se sabe que las galletas no pueden tener una concentración de sal superior a 3'5 gramos por cada 1000 galletas y, por ello, decide ir variando a partir de 1 gramo la concentración de 0'5 en 0'5 gramos cada semana e ir anotando el incremento en caja semanalmente, obteniendo la siguiente tabla:

Gramos de sal por 1000 galletas	1	1,5	2	2,5	3
Ingresos ($)	1403	1500	1650	1750	2000

A partir de tales cifras, se quiere conocer:

a) ¿Considera justificado el planteamiento de un modelo lineal para expresar la relación entre las variables?

b) Si el propietario desea unos ingresos de 1600 pesos, ¿qué cantidad de sal debería aportar por cada 1000 galletas? Si aporta el máximo permitido de sal, ¿cuál sería el ingreso en caja?

c) ¿Cuál sería la variación porcentual de los ingresos cuando la cantidad de sal aumenta en un 1%