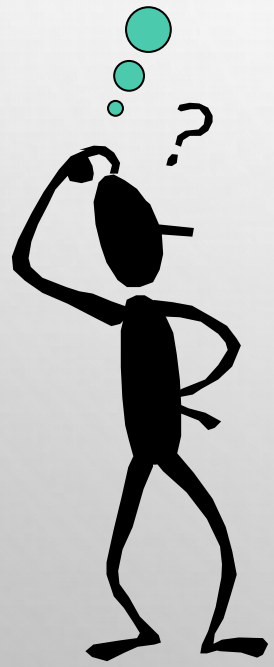
REGRESIÓN

UNIDAD N° 1

ANÁLISIS DE REGRESIÓN

Qué es el análisis de regresión lineal ?



El análisis de regresión tiene como objetivo determinar la relación funcional que vincula una *variable dependiente* con una *variable independiente*

En este caso debe existir una relación de *causalidad* entre las variables –condición que no era necesaria para el análisis de correlación-

La regresión lineal es un tema que analiza la relación entre dos o más variables para determinar una predicción mediante una recta

ANÁLISIS DE REGRESIÓN

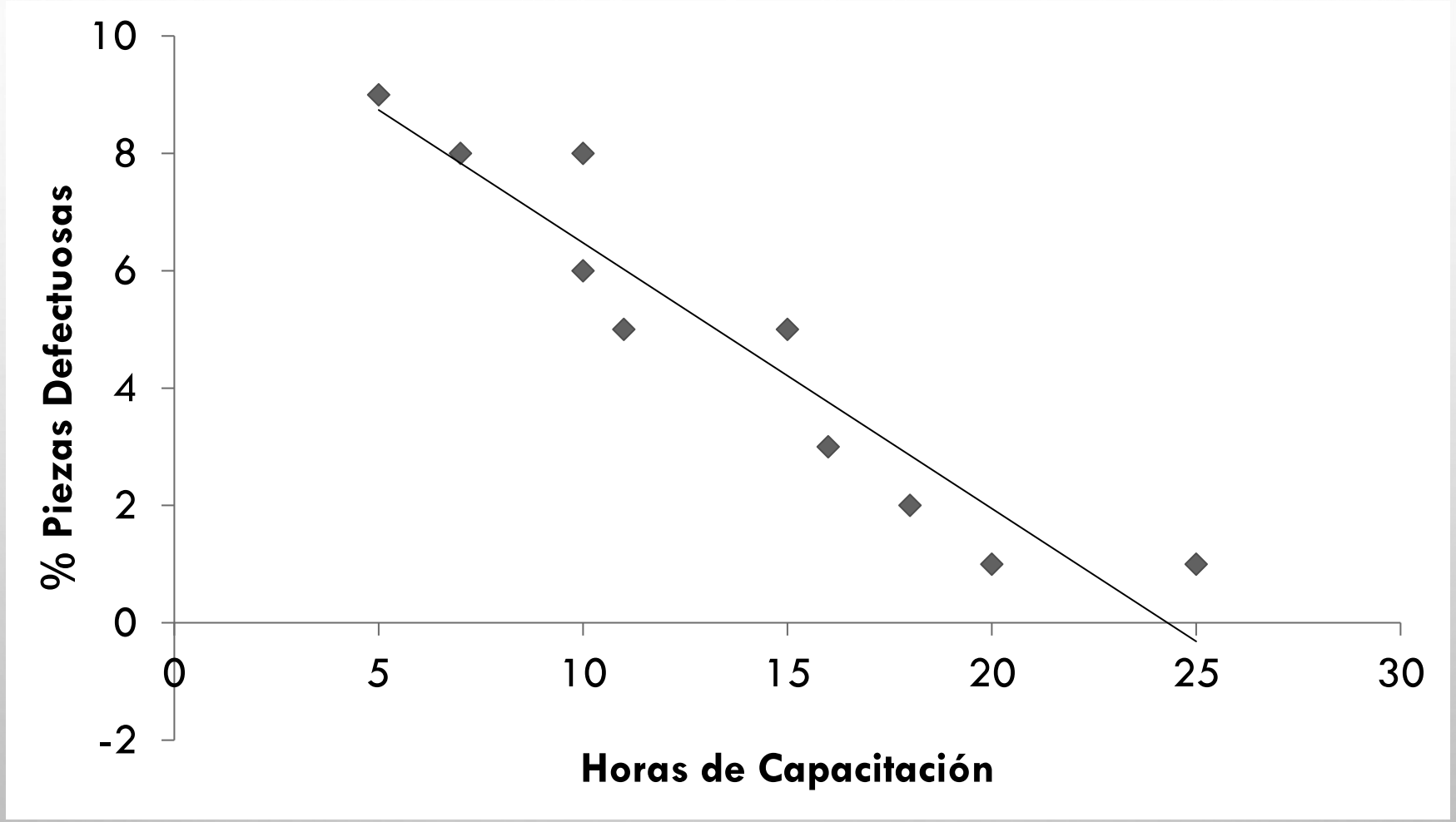
Si analizamos un ejemplo sobre la relación entre las *horas de capacitación* recibidas por un operario y el *porcentaje de piezas defectuosas* que produce

Es razonable suponer que el porcentaje de rechazo producido por un operario *depende* de las horas de entrenamiento recibidas por el mismo

El diagrama de dispersión muestra que existe una fuerte relación entre ambas variables

Es posible modelar la relación entre las variables, y el modelo adecuado puede ser una recta

ANÁLISIS DE REGRESIÓN



MÉTODO DE MÍNIMOS CUADRADOS

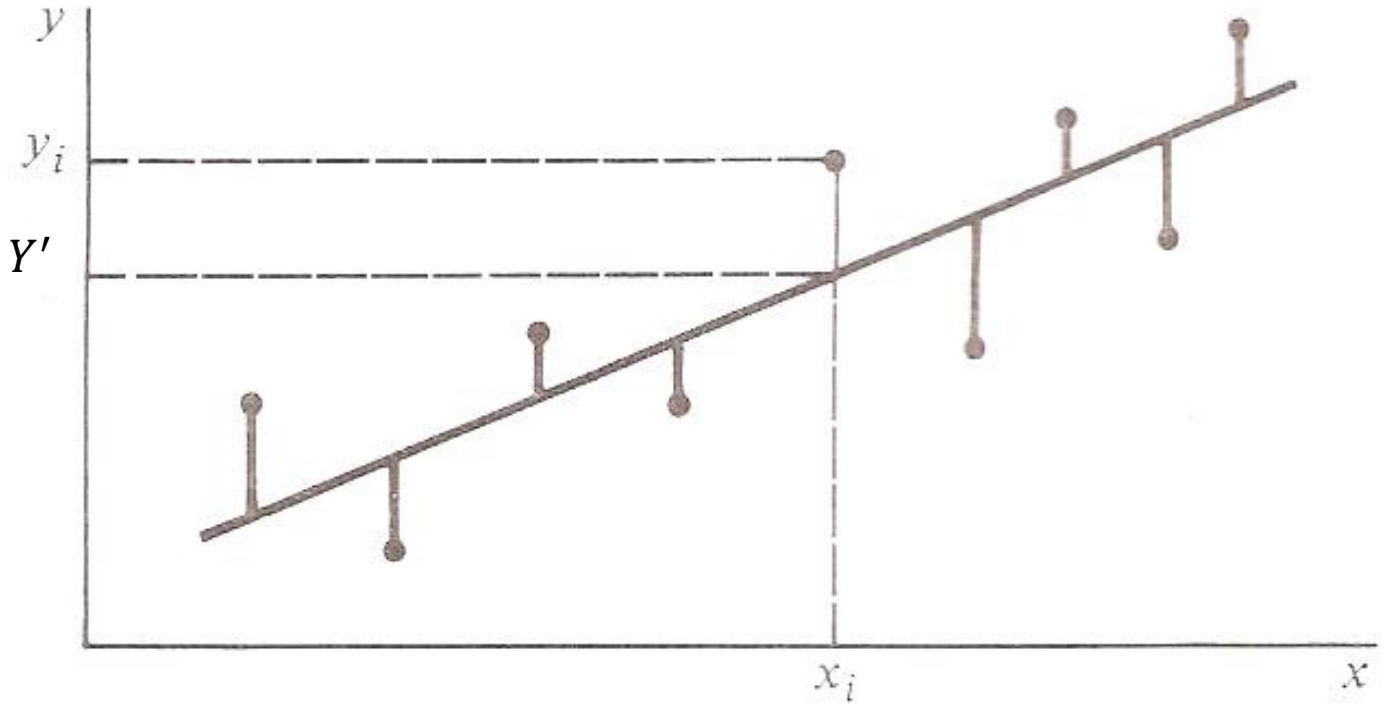
Un procedimiento para estimar los parámetros de cualquier modelo lineal es el método de los mínimos cuadrados, que se puede ilustrar sencillamente aplicándolo para ajustar una línea recta a través de un conjunto de puntos que representan los datos

Esta recta denominada *línea de regresión por mínimos cuadrados* es la que minimiza los errores de predicción

La distancia vertical entre cada punto y la recta representa el error de la predicción

MÉTODO DE MÍNIMOS CUADRADOS

$$Y' = b_y X + a_y$$



CONSTRUCCIÓN DE LA RECTA DE REGRESIÓN POR MÍNIMOS CUADRADOS

La regresión siempre se realiza para predecir una variable, dada otra variable

Se pueden realizar predicciones de Y dado X, o predecir X dado Y

Según la regresión que se va a realizar será el modelo a utilizar

CONSTANTES DE REGRESIÓN
(REGRESIÓN DE Y SOBRE X)

Las ecuaciones que se muestran a continuación utilizan los datos en bruto

Constante de regresión b_y para predecir Y dado X

$$b_y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

Constante de regresión a_y para predecir Y dado X

$$a_y = \bar{y} - b_y \bar{x}$$

REGRESIÓN DE X SOBRE Y

Para predecir X dado Y, debemos obtener una nueva línea de regresión

Se deben obtener nuevas constantes de regresión

Ahora lo que se quiere es minimizar los errores de la variable X. estos errores se representan mediante rectas horizontales paralelas al eje x

El proceso de minimizar los errores y' y el de minimizar los errores x' no conducen a las mismas líneas de regresión. La excepción ocurre cuando la relación es perfecta

La línea de regresión para la predicción de X a partir de Y se conoce como la línea de regresión de X sobre Y

La ecuación de regresión para predecir X dado Y es:

$$X' = b_x y + a_x$$

X': predicción del valor de x

b_x: pendiente que minimiza los errores X'

a_x: ordenada al origen que minimiza los errores X'

CONSTANTES DE REGRESIÓN
(REGRESIÓN DE X SOBRE Y)

Las ecuaciones que se muestran a continuación utilizan los datos en bruto

Constante de regresión b_x para predecir X dado Y

$$b_x = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}}$$

Constante de regresión a_x para predecir X dado Y

$$a_x = \bar{x} - b_x \bar{y}$$

EJEMPLO

- SE TOMA UNA MUESTRA DE OPERARIOS Y SE DESEA RELACIONAR LAS *HORAS DE CAPACITACIÓN* RECIBIDAS POR CADA OPERARIO CON EL *PORCENTAJE DE PIEZAS DEFECTUOSAS* QUE PRODUCE CADA UNO.

	X	Y	X^2	Y^2	XY
1	5	9	25	81	45
2	7	8	49	64	56
3	10	6	100	36	60
4	10	8	100	64	80
5	11	5	121	25	55
6	15	5	225	25	75
7	16	3	256	9	48
8	18	2	324	4	36
9	20	1	400	1	20
10	25	1	625	1	25
Total	***137***	***48***	***2225***	***310***	***500***

EJEMPLO

ECUACIÓN PARA PREDECIR Y DADO X

$$b_y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}} = \frac{500 - \frac{137 \cdot 48}{10}}{2225 - \frac{137^2}{10}} = -0{,}45$$

$$a_y = \bar{y} - b_y\bar{x} = \frac{48}{10} + 0{,}45 \cdot \frac{137}{10} = 10{,}96$$

$$Y' = -0{,}45x + 10{,}96$$

ECUACIÓN PARA PREDECIR X DADO Y

$$b_x = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum Y^2 - \frac{(\sum Y)^2}{N}} = \frac{500 - \frac{137 \cdot 48}{10}}{310 - \frac{48^2}{10}} = -1{,}98$$

$$a_x = \bar{x} - b_x\bar{y} = \frac{137}{10} + 1{,}98 \cdot \frac{48}{10} = 23{,}20$$

$$X' = -1{,}98y + 23{,}20$$

ERROR ESTÁNDAR

A menos que la relación entre X e Y sea perfecta, la mayor parte de los valores de Y no estarán sobre la línea de regresión

Cuando la relación es imperfecta, habrá errores en la predicción y es útil conocer estas magnitudes

La medición de los errores de predicción implica el cálculo del ***Error Estándar de la Estimación*** (el error es similar a la desviación estándar)

El error estándar de la estimación nos da una medida de la desviación promedio de los errores de predicción en torno a la línea de regresión

Mientras mayor sea su valor, menor confianza tendremos en la predicción

La ecuación para el error estándar de la estimación para predecir Y dado X:

$$S_{y/x} = \sqrt{\frac{\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right] - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]}}{N-2}}$$

Dividimos entre N-2 debido a que el cálculo del error estándar de la estimación implica el ajuste de los datos a una línea recta

EJEMPLO

- CÁLCULO DEL ERROR ESTÁNDAR PARA EL EJEMPLO CON EL QUE VENIMOS TRABAJANDO:

$$S_{\frac{Y}{X}} = \sqrt{\frac{\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right] - \frac{\left[\sum XY - \frac{(\sum X)(\sum Y)}{N}\right]^2}{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]}}{N-2}} = \sqrt{\frac{\left[310 - \frac{48^2}{10}\right] - \frac{\left[500 - \frac{137 \cdot 48}{10}\right]^2}{2225 - \frac{137^2}{10}}}{10-2}} = 1{,}02$$

- COMO PODEMOS OBSERVAR EL ERROR ESTÁNDAR ES PEQUEÑO, DEBIDO A LA FUERTE RELACIÓN ENTRE LAS VARIABLES

CONSIDERACIONES FINALES

Para utilizar la regresión lineal debemos tener las siguientes consideraciones:

La relación entre X e Y debe ser lineal

Se determina una línea de regresión para utilizarla con sujetos en los que una de las variables es desconocida

Ejemplo: Regresión

	Temp.	Costo	x^2	y^2	x*y
1	35	250	1225	62500	8750
2	29	360	841	129600	10440
3	36	165	1296	27225	5940
4	60	43	3600	1849	2580
5	65	92	4225	8464	5980
6	30	200	900	40000	6000
7	10	355	100	126025	3550
8	7	290	49	84100	2030
9	21	230	441	52900	4830
10	55	120	3025	14400	6600
11	54	73	2916	5329	3942
12	48	205	2304	42025	9840
13	20	400	400	160000	8000
14	39	320	1521	102400	12480
15	60	72	3600	5184	4320
16	20	272	400	73984	5440
17	58	94	3364	8836	5452
18	40	190	1600	36100	7600
19	27	235	729	55225	6345
20	30	139	900	19321	4170
	744	4105	33436	1055467	124289

$$b_y = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{N}}{\sum X^2 - \frac{(\sum X)^2}{N}}$$

$$= \frac{124289 - \frac{744 \cdot 4105}{20}}{33436 - \frac{744^2}{20}} = -4{,}93$$

$$a_y = \bar{y} - b_y\bar{x} = \frac{744}{20} + 4{,}93 \cdot \frac{4105}{20} = 388{,}80$$

$$Y' = -4{,}93x + 388{,}80$$

EJEMPLO: REGRESIÓN

- RECTA DE REGRESIÓN ESTIMADA:

$$Y' = -4{,}93x + 388{,}80$$

- NOTE QUE LA PENDIENTE -4.9342 TIENE SIGNO NEGATIVO, LO CUAL REFLEJA QUE LA RELACIÓN ES INVERSA, ANÁLOGO AL SIGNO DEL COEFICIENTE DE CORRELACIÓN (-0.812).
- EL VALOR DE LA PENDIENTE SIGNIFICA QUE POR CADA GRADO QUE DESCIENDA LA TEMPERATURA EXTERIOR HABRÁ UN AUMENTO PROMEDIO DE 5 DÓLARES EN EL COSTO DE LA CALEFACCIÓN.

EJEMPLO: REGRESIÓN

- **CÁLCULO DEL ERROR ESTÁNDAR DE LA ESTIMACIÓN:**
  - MIDE LA VARIABILIDAD O DISPERSIÓN DE LOS VALORES OBSERVADOS ALREDEDOR DE LA LÍNEA DE REGRESIÓN
  - MIENTRAS MÁS GRANDE SEA EL ERROR ESTÁNDAR DE LA ESTIMACIÓN, MAYOR SERÁ LA DISPERSIÓN DE LOS PUNTOS ALREDEDOR DE LA LÍNEA DE REGRESIÓN
- EN NUESTRO EJEMPLO EL ERROR ESTÁNDAR DE ESTIMACIÓN QUE SE COMETE AL USAR LA RECTA PARA ESTIMAR EL COSTO ES DE $63,553

EJEMPLO: REGRESIÓN

- **CÁLCULO DEL COEFICIENTE DE DETERMINACIÓN:**
  - MIDE EL **PODER EXPLICATIVO DEL MODELO DE REGRESIÓN**, ES DECIR, LA PARTE DE LA VARIACIÓN DE Y EXPLICADA POR LA VARIACIÓN DE X
  - EL VALOR DE R^2 HA DE ESTAR ENTRE 0 Y 1, SI $R^2 = 0{,}70$ SIGNIFICA QUE EL 70% DE LA VARIACIÓN DE Y ESTÁ EXPLICADA POR LAS VARIACIONES DE X. ES EVIDENTE QUE CUANTO MAYOR SEA R^2, MAYOR PODER EXPLICATIVO TENDRÁ NUESTRO MODELO
- **EN NUESTRO EJEMPLO:**
  - **SI ANALIZAMOS EL VALOR DEL COEFICIENTE DE DETERMINACIÓN $R^2 = 0.659$, APRECIAMOS QUE APROXIMADAMENTE EL 66% DE LA VARIABILIDAD DEL COSTO ESTA DETERMINADO POR LA VARIABILIDAD EN LA TEMPERATURA EXTERIOR**

EJEMPLO: REGRESIÓN

- **PREDICCIONES: CÁLCULO DE UNA PREDICCIÓN PUNTUAL**
  - SUPONGA QUE SE DESEA UN ESTIMADOR PUNTUAL DEL COSTO DE UN APARTAMENTO, SI LA TEMPERATURA EXTERIOR ES DE 35 GRADOS
  - SOLUCIÓN: SUSTITUCIÓN DEL VALOR DE X = 35, EN LA ECUACIÓN DE LA RECTA PARA OBTENER UN VALOR

$$Y' = -4{,}93x + 388{,}80$$

$$Y' = -4{,}93 \cdot 35 + 388{,}80$$

$$Y' = 216{,}25$$

REGRESIÓN MÚLTIPLE

ANÁLISIS DE REGRESIÓN MÚLTIPLE

El Análisis de Regresión Lineal Múltiple nos permite establecer la relación que se produce entre una variable dependiente Y, y un conjunto de variables independientes (X_1, X_2, ... X_K)

El análisis de regresión lineal múltiple, a diferencia del simple, se aproxima más a situaciones de análisis real puesto que los fenómenos, hechos y procesos sociales, por definición, son complejos y, en consecuencia, deben ser explicados en la medida de lo posible por la serie de variables que, directa e indirectamente, participan en su concreción

La anotación matemática del modelo o ecuación de regresión lineal múltiple es la que sigue:

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k$$

en donde:

$\hat{y}$ es la variable a predecir

$b_0, b_1, \ldots, b_k$, son parámetros desconocidos a estimar

ANÁLISIS DE REGRESIÓN MÚLTIPLE

Hay modelos que se describen de una forma más apropiada con el ***modelo de regresión polinomial***, y la respuesta estimada se obtiene de la ecuación de regresión polinomial

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + \cdots + b_r x^r$$

En ocasiones surge confusión cuando hablamos de un modelo polinomial como un modelo lineal

Sin embargo, los estadísticos se refieren a un modelo lineal como uno en el cual los parámetros ocurren linealmente, sin importar cómo entran las variables independientes al modelo

Un ejemplo de un modelo no lineal es la ***relación exponencial***, que se estima con la ecuación de regresión

$$\hat{y} = ab^x$$

ESTIMACIÓN DE LOS COEFICIENTES

Si la regresión múltiple es lineal, podemos generar un conjunto de ecuaciones linealmente independientes para poder calcular los coeficientes mediante cualquier método apropiado para resolver sistemas de ecuaciones lineales y de esta manera obtener $b_0, b_1, b_2, \ldots, b_k$

$$nb_0 + b_1\sum_{i=1}^{n} x_{1i} + b_2\sum_{i=1}^{n} x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{ki} = \sum_{i=1}^{n} y_i$$

$$b_0\sum_{i=1}^{n} x_{1i} + b_1\sum_{i=1}^{n} x^2{}_{1i} + b_2\sum_{i=1}^{n} x_{1i}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x_{1i}x_{ki} = \sum_{i=1}^{n} x_{1i}y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$b_0\sum_{i=1}^{n} x_{ki} + b_1\sum_{i=1}^{n} x_{ki}x_{1i} + b_2\sum_{i=1}^{n} x_{ki}x_{2i} + \cdots + b_k\sum_{i=1}^{n} x^2{}_{ki} = \sum_{i=1}^{n} x_{ki}y_i$$

EJEMPLO: REGRESIÓN MÚLTIPLE LINEAL

- SE REALIZÓ UN ESTUDIO SOBRE UN CAMIÓN DE REPARTO LIGERO A DIESEL PARA VER SI LA HUMEDAD, TEMPERATURA DEL AIRE Y PRESIÓN BAROMÉTRICA INFLUYEN EN LA EMISIÓN DE ÓXIDO NITROSO (EN PPM). LAS MEDICIONES DE LAS EMISIONES SE TOMARON EN DIFERENTES MOMENTOS, CON CONDICIONES EXPERIMENTALES VARIANTES.
- LOS DATOS SE MUESTRAN EN LA TABLA:

	Óxido nitroso y	*Humedad x1*	*Temperatura x2*	*Presión x3*
1	0,90	72,4	76,3	29,18
2	0,91	41,6	70,3	29,35
3	0,96	34,3	77,1	29,24
4	0,89	35,1	68,0	29,27
5	1,00	10,7	79,0	29,78
6	1,10	12,9	67,4	29,39
7	1,15	8,3	66,8	29,69
8	1,03	20,1	76,9	29,48
9	0,77	72,2	77,7	29,09
10	1,07	24,0	67,7	29,60
11	1,07	23,2	76,8	29,38
12	0,94	47,4	86,6	29,35
13	1,10	31,5	76,9	29,63
14	1,10	10,6	86,3	29,56
15	1,10	11,2	86,0	29,48
16	0,91	73,3	76,3	29,40
17	0,87	75,4	77,9	29,28
18	0,78	96,6	78,7	29,29
19	0,82	107,4	86,8	29,03
20	0,95	54,9	70,9	29,37

EJEMPLO: REGRESIÓN MÚLTIPLE LINEAL

- E1: $20b_0 + 863{,}1b_1 + 1530{,}4b_2 + 587{,}84b_3 = 19{,}42$
- E2: $863{,}1b_0 + 54876{,}89b_1 + 67000{,}09b_2 + 25283{,}40b_3 = 779{,}48$
- E3: $1530{,}40b_0 + 67000{,}09b_1 + 117912{,}32b_2 + 44976{,}87b_3 = 1483{,}44$
- E4: $587{,}84b_0 + 25283{,}40b_1 + 44976{,}87b_2 + 17278{,}51b_3 = 571{,}12$
- ***B_0=-3.30705, B_1=-0.00265768, B_2=0.000801712, B_3=0.147366***
- ECUACIÓN DE REGRESIÓN

$$\hat{y} = -3{,}31 - 0{,}0027x_1 + 0{,}000802x_2 + 0{,}147x_3$$

- PARA 50% DE HUMEDAD, UNA TEMPERATURA DE 76°F Y UNA PRESIÓN BAROMÉTRICA DE 29,30, LA CANTIDAD ESTIMADA DE ÓXIDO NITROSO ES:

$$\hat{y} = -3{,}31 - 0{,}0027(50) + 0{,}000802(76) + 0{,}147(29{,}3) = 0{,}9388$$

REGRESIÓN POLINOMIAL

Suponga ahora que deseamos ajustar la ecuación polinomial

$$y_i = b_0 + b_1 x_i + b_2 {x_i}^2 + \cdots + b_r {x_i}^r + e_i$$

Donde *r* es el grado del polinomio y e_i es el error residual asociado con la respuesta y_i

El número de pares *n* debe ser al menos tan grande como *r+1*, el número de parámetros a estimar

Las ecuaciones toman la forma:

$$nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} {x_i}^2 + \cdots + b_r \sum_{i=1}^{n} {x_i}^r = \sum_{i=1}^{n} y_i$$

$$b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} {x_i}^2 + b_2 \sum_{i=1}^{n} {x_i}^3 + \cdots + b_r \sum_{i=1}^{n} {x_i}^{r+1} = \sum_{i=1}^{n} x_i y_i$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$b_0 \sum_{i=1}^{n} {x_i}^r + b_1 \sum_{i=1}^{n} {x_i}^{r+1} + b_2 \sum_{i=1}^{n} {x_i}^{r+2} + \cdots + b_r \sum_{i=1}^{n} {x_i}^{2r} = \sum_{i=1}^{n} {x_i}^r y_i$$

EJEMPLO: REGRESIÓN POLINOMIAL

- DADOS LOS DATOS EXPRESADOS EN LA TABLA, HALLAR LA ECUACIÓN DE REGRESIÓN POLINOMIAL

$$\hat{y} = b_0 + b_1 x + b_2 x^2$$

	x	*y*
1	0	9,1
2	1	7,3
3	2	3,2
4	3	4,6
5	4	4,8
6	5	2,9
7	6	5,7
8	7	7,1
9	8	8,8
10	9	10,2

EJEMPLO: REGRESIÓN POLINOMIAL

- E1: $10b_0 + 45b_1 + 285b_2 = 63{,}7$
- E2: $45b_0 + 285b_1 + 2025b_2 = 307{,}3$
- E3: $285b_0 + 2025b_1 + 15333b_2 = 2153{,}3$
- B_0=8.69818, B_1= -2.34061, B_2=0.287879
- ECUACIÓN DE REGRESIÓN

$$\hat{y} = 8,70 - 2,34x + 0,29x^2$$

- CUANDO X=2 NUESTRA ESTIMACIÓN ES:

$$\hat{y} = 5{,}17$$

ERROR ESTÁNDAR DE ESTIMACIÓN

El error estándar de estimación se puede calcular en términos de los coeficientes de correlación r_{12}, r_{13}, r_{23} (siendo r_{12} la correlación entre las variables X_1 y X_2; r_{13} la correlación entre las variables X_1 y X_3; r_{23} la correlación entre las variables X_2 y X_3

Se puede considerar que $X_1 = y$, el resto de las variables (X_2, X_3...) corresponden a variables independientes

Para el cálculo se emplea la siguiente fórmula:

$$s_{1.23} = s_1 \sqrt{\frac{1 - {r_{12}}^2 - {r_{13}}^2 - {r_{23}}^2 + 2r_{12}r_{13}r_{23}}{1 - {r_{23}}^2}}$$

Donde s_1 es el desvío estándar de X_1