PROCESO DE GESTIÓN DE FORMACIÓN PROFESIONAL INTEGRAL
ANEXO COMPONENTE FORMATIVO

Distribuciones bidimensionales y rectas de regresión

Distribuciones bidimensionales

En una distribución bidimensional se consideran dos variables estadísticas sobre una misma población. Las representaremos, en general, por (X, Y). Estas distribuciones suelen presentarse mediante una tabla de tres columnas, apareciendo en las dos primeras los valores de las variables, y en la tercera, la frecuencia del par correspondiente, es decir, en la forma:

X	Y	n_{ij}
x_1	y_1	n_{11}
x_1	y_2	n_{12}
$\vdots$	$\vdots$	$\vdots$
x_i	y_j	n_{ij}
$\vdots$	$\vdots$	$\vdots$
x_h	y_k	n_{hk}

En ocasiones es preferible hacerlo mediante una tabla de doble entrada, con disposición rectangular, en la forma:

$Y \setminus X$	x_1	x_2	x_3	$\cdots$	x_i	$\cdots$	x_h
y_1	n_{11}	n_{21}	n_{31}	$\cdots$	n_{i1}	$\cdots$	n_{h1}
y_2	n_{12}	n_{22}	n_{32}	$\cdots$	n_{i2}	$\cdots$	n_{h2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_j	n_{1j}	n_{2j}	n_{3j}	$\cdots$	n_{ij}	$\cdots$	n_{hj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_k	n_{1k}	n_{2k}	n_{3k}	$\cdots$	n_{ik}	$\cdots$	n_{hk}

Ejemplo: supongamos que contamos el número de pizzerías (X) y el número de hamburgueserías (Y) en 80 localidades de una región. Obtenemos la siguiente tabla de frecuencias:

X	Y	n_{ij}
0	1	4
1	1	3
1	3	4
2	0	2
2	2	9
2	3	3
3	1	6
3	2	12
3	3	5
3	4	2
4	0	2
4	1	7
4	2	15
4	4	1
5	2	5

La interpretación es que hay 4 localidades que tienen 0 pizzerías y 1 hamburguesería, hay 12 localidades con 3 pizzerías y 2 hamburgueserías, etc. La tabla de doble entrada sería la siguiente:

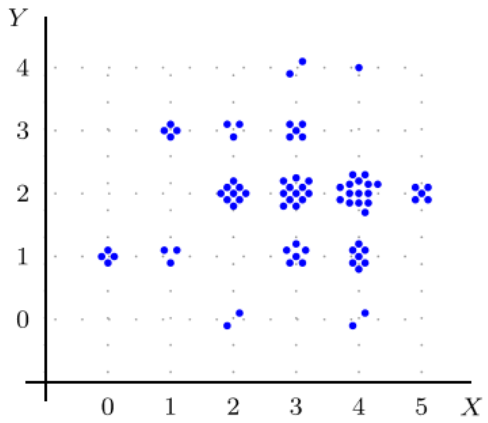
Y \ X	**0**	**1**	**2**	**3**	**4**	**5**
0	0	0	2	0	2	0
1	4	3	0	6	7	0
2	0	0	9	12	15	5
3	0	4	3	5	0	0
4	0	0	0	2	1	0

La forma más usual de representar gráficamente las distribuciones bidimensionales es el diagrama de dispersión o nube de puntos, que se obtiene al considerar dos ejes coordenados, situando en el eje horizontal los valores de la variable **X** y en el vertical los de la variable Y; en las proximidades del par ($\mathbf{x_i}$, $\mathbf{y_j}$) se colocan tantos puntos como indica su frecuencia conjunta $\mathbf{n_{ij}}$.

Ejemplo: la distribución bidimensional que consideramos anteriormente:

Y \ X	**0**	**1**	**2**	**3**	**4**	**5**
0	0	0	2	0	2	0
1	4	3	0	6	7	0
2	0	0	9	12	15	5
3	0	4	3	5	0	0
4	0	0	0	2	1	0

Tendría la siguiente representación:



También se puede optar por utilizar puntos de distinto tamaño según la frecuencia, o simplemente representar un punto por cada valor (x_i, y_i), en los casos en que las variables varían continuamente y no hay repeticiones.

Covarianza

Para una distribución estadística bidimensional (X, Y), se llama covarianza a la media aritmética de los productos de las desviaciones de cada variable respecto a su media aritmética; se indicará por S_{XY}, y está dada por la fórmula:

$$S_{XY} = \frac{\sum_{i=1}^{h} \sum_{j=1}^{k} (x_i - \overline{X})(y_j - \overline{Y}) n_{ij}}{N}$$

Se puede calcular más fácilmente en la forma:

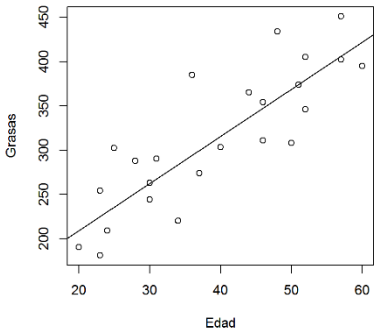
$$S_{XY} = \frac{\sum_i \sum_j x_i y_j n_{ij}}{N} - \left(\frac{\sum_i \sum_j x_i n_{ij}}{N} \right) \left(\frac{\sum_i \sum_j y_j n_{ij}}{N} \right),$$

es decir, la covarianza es igual a la media de los productos menos el producto de las medias.

Regresión

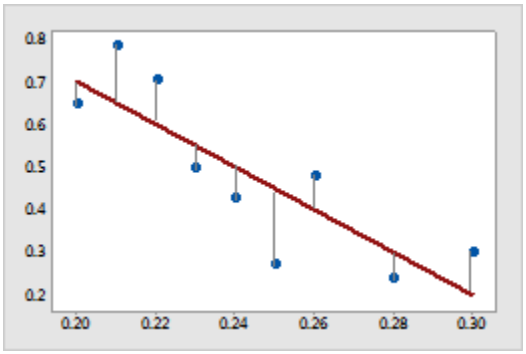
La regresión es un método estadístico utilizado para el análisis de los datos en las finanzas, inversiones y otras disciplinas que intenta determinar la fuerza y el carácter de la relación entre una variable dependiente (generalmente denotada por Y) y una serie de otras variables (conocidas como variables independientes).

Una aplicación muy común de la regresión es ayudar a los administradores financieros y de inversiones a valorar los activos y comprender las relaciones entre las variables, como los precios de los productos básicos y las acciones de las empresas que comercian con esos productos básicos.



Regresión

Rectas de regresión: las rectas de regresión son líneas que se utiliza para describir el comportamiento de un conjunto de datos. En otras palabras, da la mejor tendencia de los datos proporcionados. Las rectas de regresión son útiles en los procedimientos de pronóstico. Su propósito es describir la interrelación de la variable dependiente (variable y) con una o muchas variables independientes (variable x).



Recta de regresión simple

El uso de la ecuación obtenida de la recta de regresión actúa como un analista que puede pronosticar comportamientos futuros de las variables dependientes ingresando diferentes valores para las independientes.

Los dos tipos básicos de regresión son **la recta de regresión** normal o simple y **la recta de regresión múltiple**, aunque existen métodos de regresión no lineal para datos y análisis más complicados. **la recta de regresión simple** usa una variable independiente para explicar o predecir el resultado de la variable dependiente Y, mientras que **la recta de regresión múltiple** usa dos o más variables independientes para predecir el resultado. La forma general de cada tipo de regresión es:

Fórmula de la recta de regresión simple: $Y = a + bX + u$

Fórmula de la recta de regresión múltiple: $Y = a + b_1X_1 + b_2 X_2 + b_3 X_3 + ... + b_t X_t + u$

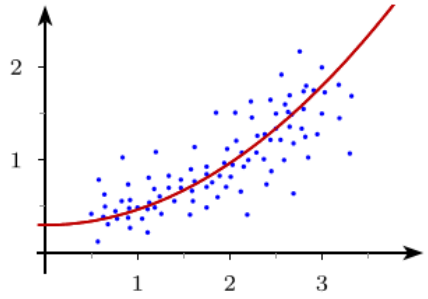
Donde:

Y = la variable que está intentando predecir (variable dependiente).
X = la variable que está utilizando para predecir Y (variable independiente).
a = la intersección.
b = la pendiente.
u = el residual de regresión.

Las rectas de regresión se utilizan en el sector financiero y empresarial entre otros. Varios analistas financieros emplean regresiones lineales para pronosticar los precios de las acciones, los precios de las materias primas y para realizar valoraciones de muchos valores diferentes. Varias empresas emplean regresiones lineales con el propósito de pronosticar ventas, inventarios y muchas otras variables.

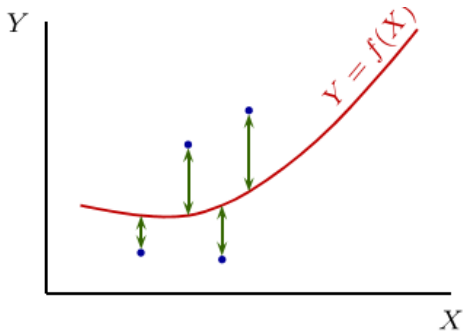
Dada una distribución bidimensional, cabe preguntarse si las dos variables son independientes o si están relacionadas entre sí. Y si están relacionadas, cuál es esa relación. Por ejemplo, si consideramos sobre un grupo de personas la estatura (X) y el sueldo mensual (Y), lo lógico es pensar que se trata de dos variables completamente independientes. Sin embargo, si consideramos la estatura (X) y el peso (Z) sí que va a haber una relación importante, ya que las personas más altas suelen por lo general tener mayor peso.

La relación entre dos variables puede observarse al representar gráficamente la nube de puntos. Cuando dos variables están relacionadas, la nube de puntos tiende a concentrarse en torno a la gráfica de una determinada función. El problema de la regresión consiste en encontrar esa función.



El método más habitual es la regresión por mínimos cuadrados. Primero hemos de decidir qué tipo de función creemos que es la más apropiada para nuestro caso. Por ejemplo, podemos decidir aproximar por un polinomio de segundo grado $f(x) = ax^2 + bx + c$. El siguiente paso consistiría en encontrar los valores de a, b y c que hacen que la diferencia entre la gráfica de la función y la distribución sea lo más pequeña posible. Esta diferencia se cuantifica calculando, para cada valor de la distribución ($\mathbf{x_i}$, $\mathbf{y_i}$) la diferencia $\mathbf{y_i - f(x_i)}$, y después sumando los cuadrados de todas esas diferencias:

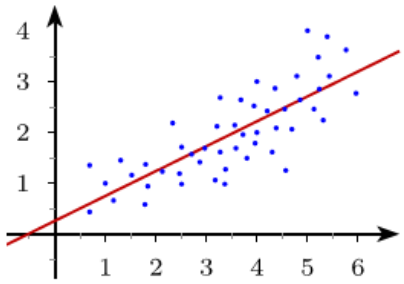
$$\sum_{i=1}^{N} (y_i - f(x_i))^2$$



En la figura anterior los puntos azules son los de la gráfica de la distribución. El proceso de regresión por mínimos cuadrados consiste en encontrar la función f cuya gráfica tiene la propiedad de que la suma de los cuadrados de las longitudes indicadas en verde sea lo menor posible.

El caso más sencillo es la regresión lineal, cuando el tipo de función por el que aproximamos es una función lineal f(x) = ax + b. Es decir, se trata de encontrar una recta y = ax + b que sea la que mejor aproxime a nuestra distribución bidimensional (X, Y) por mínimos cuadrados. Esta recta se halla mediante la siguiente fórmula:

$$y - \overline{Y} = \frac{S_{XY}}{S_X^2}(x - \overline{X})$$



donde recordamos que y son las medias aritméticas de X e Y respectivamente, SX es la varianza de X, SXY es la covarianza de X e Y. La ecuación anterior nos indica que la recta hallada pasa por el punto (,), llamado centro de gravedad de la distribución bidimensional, y tiene por pendiente:

$$a = \frac{S_{XY}}{S_X^2}$$

llamado el coeficiente de regresión. Esta recta se llama recta de regresión de Y sobre X. Es importante especificar el orden de las variables, puesto que la recta de regresión de Y sobre X no coincide con la de X sobre Y. Ambas rectas aspiran a ser las que mejor se aproximan a la distribución, pero en un caso hemos tratado de minimizar distancias medidas en vertical y en otro caso en horizontal, por lo que el resultado no será exactamente el mismo.

Se llama coeficiente de correlación lineal de Pearson al valor:

$$r = \frac{S_{XY}}{S_X \cdot S_Y}$$

Este coeficiente sirve para medir hasta qué punto la recta de regresión es una buena aproximación de la distribución: mejor cuanto más próximo esté el valor de |r| a 1 y peor cuanto más se acerque a 0. Si r = ±1, la correlación lineal es perfecta, directa o inversa, es decir, la nube de puntos está situada, toda ella, sobre la recta de regresión, con pendiente positiva para r = 1 y negativa para r = −1. Si r = 0, no existe dependencia lineal entre las variables, pudiendo darse una dependencia no lineal, o bien puede ocurrir que las variables sean independientes.