Educaguía.com

ESTADÍSTICA

REGRESIÓN

REGRESIÓN

Consideramos la recta de regresión aquella que más se aproxima a la nube de puntos generada por la distribución bidimensional.
Esta recta es la línea que haga que la suma de las desviaciones de los puntos de la nube respecto de los correspondientes de la línea sea lo menor posible. En decir es la línea que menos se separa de la nube de puntos.
Cuando ajustamos una línea de regresión a una nube de puntos podemos hacerlo mediante una recta, una parábola, una cúbica, una exponencial, etc...
Nosotros vamos a hacer solo la regresión lineal.

La regresión lineal

Supongamos que una vez estudiada la correlación existente entre las dos variables que componen la variable bidimensional, se observa que están fuertemente correladas y que el diagrama de puntos se puede ajustar mediante una recta.
La ecuación de la recta buscada, cuando consideramos a X como variable independiente y a Y como variable dependiente será de la forma:

$$y - \bar{y} = m(x - \bar{x})$$

Donde m recibe el nombre de coeficiente de regresión y se demuestra que es igual al cociente de la covarianza y la desviación típica al cuadrado.
De donde la ecuación de la recta que estábamos buscando es:

$$y - \bar{y} = \frac{s_{xy}}{s_x^2}(x - \bar{x})$$

A esta recta de regresión se le llama de y sobre x, ya que hemos considerado a y como variable dependiente de x.
Análogamente se puede obtener la recta de regresión de x sobre y:

$$x - \bar{x} = \frac{s_{xy}}{s_y^2}(y - \bar{y})$$

La fiabilidad de la aproximación de esta recta será tanto mayor cuanto mayor sea el coeficiente de correlación lineal en valor absoluto. Así pues:

- Si r es muy pequeño, realmente no hay ninguna posibilidad de hacer estimaciones.
- Si r es próximo a –1 o a 1, probablemente los valores reales serán próximos a nuestras estimaciones.
- Si r = –1 o r = 1, las estimaciones realizadas coincidirán con los valores reales.

Pero incluso para estos valores próximos a 1, las estimaciones pueden resultar poco fiables, por ejemplo, cuando se pretende extrapolar más allá del recorrido de los datos observados.

Por ejemplo:
Supongamos que el número de casos de SIDA detectados desde su aparición hasta el momento tiene una fuerte correlación con el número de individuos fallecidos. En el supuesto de que dentro de 15 años se presenten 5000 casos, ¿se puede predecir cuántos individuos fallecerán de esos 5000?
No es porque es probable que en este tiempo se consiga un tratamiento que haga disminuir la mortalidad.