

Relación entre variables cuantitativas:

Regresión

La recta de regresión es otra manera de expresar la relación lineal entre dos variables continuas y describe cómo como es la variación de una variable en función de los valores de la otra. La recta de regresión está muy relacionada con el coeficiente de correlación.

La recta de regresión como una recta de medias:

El coeficiente de correlación es una medida resumen de la asociación lineal entre dos variables continuas. Una manera alternativa de resumir la relación es la recta de regresión, determinada por una ecuación que resume el diagrama de dispersión. Sin duda, tal como se vio, es sencillo el trazado del diagrama de dispersión cuando se tienen datos no agrupados.

Cuando se tienen datos agrupados en intervalos tenemos el problema de que no podemos saber el valor concreto de cada dato. La solución es tener en cuenta que la media constituye la medida descriptiva principal y más simple de una variable y que las principales relaciones estadísticas son relaciones que nos dicen como varía *en promedio* una variable al variar otra. Por lo tanto, para expresar la relación entre dos variables podemos calcular cómo varía la media de la distribución de la variable dependiente condicionada a los valores de la variable independiente. Es eso lo que se representará en el diagrama de dispersión correspondiente a datos agrupados en intervalos y lo que será resumido por la correspondiente recta de regresión.

Debe hallarse así la media de la variable dependiente para cada intervalo de la variable independiente. Es decir que los \bar{x}_i del cálculo de la media, serían los datos agrupados en cada casilla (la frecuencia absoluta de cada celda en que ese intervalo intersecciona con los intervalos de la variable dependiente) y el N sería el valor que toma la distribución marginal de la variable x en ese intervalo. Al graficar, usaremos como valor de la variable independiente la marca de clase de cada uno de sus intervalos.

La recta de regresión tiene la utilidad de que por medio de ella podemos realizar predicciones de la variable dependiente al conocer el valor de la variable independiente. Naturalmente, se desea que la predicción sea la mejor o, en otros términos, que el error que pueda cometerse al predecir sea lo más pequeño posible. Se trata de minimizar la diferencia entre el valor observado y el valor previsto por la predicción.

Esta es la idea que se sigue en la recta de regresión. Si se desea prever el valor de una variable de un elemento de una población y se conoce el valor que otra variable, relacionada con la primera toma en dicho elemento, se elegirá como predictor la media de la distribución de la primera variable condicionada al valor observado de la segunda. La ecuación de regresión nos proporciona automáticamente la mejor predicción de la variable dependiente cuando conocemos los valores de la variable independiente.

Cálculo de la recta de regresión:

Como antes, llamaremos y a la variable dependiente, o endógena, y x a la variable independiente, o exógena. La recta de regresión indica el valor medio de y para cada valor observado de x . Llamaremos $\hat{y}(x)$ a las estimaciones de la variable y que vamos a obtener con la recta, para indicar que dependen del valor particular de x .

Representaremos la ecuación de la recta por

$$\hat{y} = a + bx$$

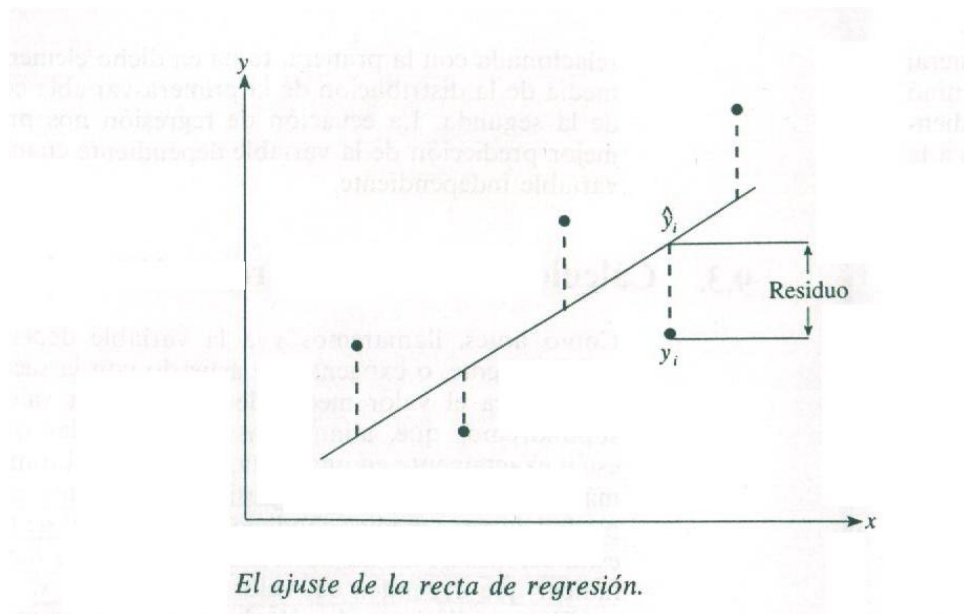
Esta ecuación depende de dos constantes o parámetros, a y b , que deben calcularse a partir de los datos observados. El parámetro b_1 es el más importante. Se denomina pendiente de la recta y nos dice cuánto aumenta la media de la variable dependiente cuando la independiente aumenta en una unidad.

El parámetro a se denomina ordenada en el origen y representa el valor de la variable dependiente cuando la independiente toma el valor cero. En muchos problemas es absurdo que x sea considerada de valor cero (¿cuál es la estatura cuando el valor es cero?). Entonces a debe tomarse como un valor de referencia necesario para describir la recta, pero sin pretender atribuirle una interpretación lógica en el problema.

Para hallar a y b a partir de los datos se define el error de predicción o residuo de la recta, que es el error al predecir cada uno de los valores observados de la variable dependiente.

Residuo= error de predicción = valor observado – valor de predicho por la recta

La recta se calcula imponiendo la condición de que el error promedio, definido como la raíz cuadrada de la suma de los cuadrados de los errores al prever cada punto de la recta, sea mínimo. Observemos que el error que minimizamos es el de prever la variable dependiente, por eso minimizamos las distancias verticales a la recta. Este criterio se denomina de mínimos cuadrados.



Como resultado de aplicar este criterio se obtiene que la pendiente b viene dada por la ecuación

$$b = \frac{\text{COV}(x, y)}{s_x^2}$$

es decir, la pendiente de la recta resulta de estandarizar la covarianza de manera que tenga las unidades apropiadas de aumento en y por unidad de aumento en x . Observemos que si la covarianza es cero, también lo será la pendiente de la recta.

Es importante resaltar que la pendiente de la recta es la medida básica de la asociación, ya que si $b = 0$, no hay relación entre las variables.

El segundo parámetro a , ordenada en el origen de la recta, indica su posición concreta y se calcula

$$a = \bar{y} - b \bar{x}$$

Por lo tanto, teniendo el valor del parámetro a y un valor estimado de y para un x cualquiera, podemos trazar la recta de regresión, ya que una recta queda definida por dos puntos.

Datos agrupados

Para los datos agrupados, si bien la idea es la misma, la fórmula es levemente diferente, según ya vimos que ocurría con el cálculo del coeficiente de correlación. Así que ahora el cálculo de b se efectuaría de la siguiente manera

$$b = \frac{N \sum_{i=1}^k \sum_{j=1}^m n_{ij} \cdot i' \cdot j' - \left(\sum_{i=1}^k n_i \cdot i' \right) \left(\sum_{j=1}^m n_j \cdot j' \right)}{N \sum_{i=1}^k n_i \cdot i'^2 - \left(\sum_{i=1}^k n_i \cdot i' \right)^2} \cdot \frac{j'}{i'}$$

Ahora a puede calcularse de manera sencilla aplicando la fórmula de cálculo que ya vimos para hallar esa constante.

La desviación estándar residual:

La recta de regresión puede, entonces, ser considerada una recta de medias. Por ello conviene complementarla con una medida de variabilidad de los puntos con relación a la media. La desviación entre cada observación de la variable dependiente y la recta es lo que hemos llamado el error de predicción al prever ese punto con la recta, o residuo.

La desviación estándar residual es una medida del valor promedio de los residuos. Se obtiene de la siguiente forma: Se elevan al cuadrado cada uno de los desvíos, se suman, se divide por el número de residuos (de casos) y se extrae la raíz cuadrada, de manera análoga a lo que hacíamos con la desviación estándar.

$$s_r = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

La desviación estándar residual tiene una interpretación similar a la desviación estándar: representa la variabilidad promedio de los datos observados con relación a la recta de regresión, nos da el error promedio al estimar mediante la recta de regresión

Regresión y correlación:

El coeficiente de correlación es un resumen del gráfico de dispersión entre dos variables. La recta de regresión es otra manera de resumir esta información, y su parámetro fundamental, la pendiente, está relacionado con el coeficiente de correlación.

La diferencia entre regresión y correlación es que en el cálculo de la correlación ambas variables se tratan simétricamente, mientras que en la regresión no. En la regresión se trata de prever la variable dependiente en función de los valores de la independiente. En consecuencia, si cambiamos el papel de las variables cambiará también la ecuación de regresión, porque la recta se adapta a las unidades de la variable que se desea predecir.

El cuadrado del coeficiente de correlación está relacionado con la reducción de variabilidad que se consigue al predecir con la recta de regresión en lugar de hacerlo con la media de la variable. Si pretendemos prever el valor de la variable dependiente en un elemento de los datos y no conocemos el valor de la variable independiente para ese elemento, tendremos que utilizar la media de la variable dependiente y cometeremos un error promedio de predicción S_y . Sin embargo si conocemos el valor de la variable independiente para ese fenómeno podemos dar como predicción la media de la distribución condicionada de y dado x , que es el valor de la recta de regresión para el valor de x en ese elemento. Entonces, el error promedio de la predicción será la desviación estándar residual, S_r . El cociente entre ambas medidas del error de predicción es

$$\frac{S_r^2}{S_y^2} = 1 - r^2$$

Por ejemplo si $r = 0,7$, tendremos que $1 - 0,49 = 0,51$ y $\sqrt{0,51}$ es 0,71. Esta cantidad se interpreta diciendo que en la predicción de y con la recta de regresión el error promedio es el 74% del que tendríamos si no conociésemos x .

El cuadrado del coeficiente de correlación, r^2 , puede interpretarse como la proporción de variación total en una de las variables explicadas por la otra.