Aula360
Aprendizaje para el siglo XXI

Apuntes Unidad 2

Diferencia entre valor observado y estimado

FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS UNIVERSIDAD DE CHILE

CMM Centro de Modelamiento Matemático

Curso: Probabilidad y estadística descriptiva
Unidad 2 : Media muestral, dispersión y correlación
Tema: Modelo de regresión lineal
Contenido: Diferencia entre valor observado y estimado

DIFERENCIA ENTRE EL VALOR OBSERVADO Y EL ESTIMADO POR EL MODELO

Supongamos que estamos considerando una recta $y = mx + b$ para modelar la relación entre las variables y queremos ver qué tan bien se ajusta a los datos.

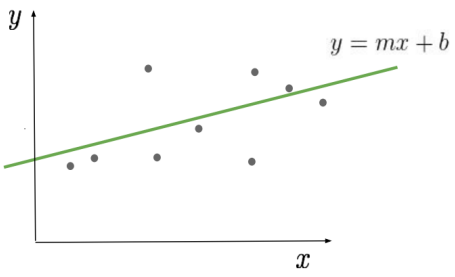

Figura 1: Gráfico de la situación.

Para cada punto del gráfico (x_i, y_i), podemos encontrar el valor que estima el modelo, es decir, cuando $x = x_i$. Este valor se denominará $\hat{y}_i$.

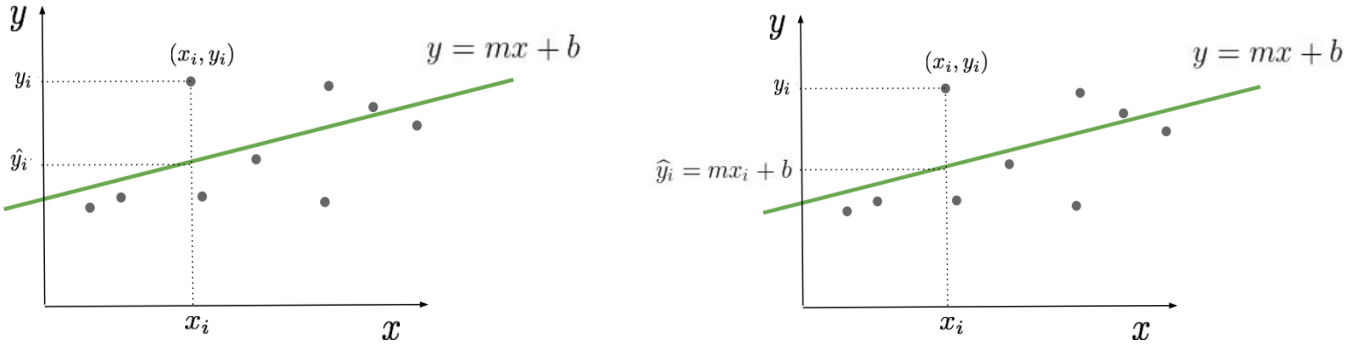

Figura 2: Representación gráfica de y_i y $\hat{y}_i$.

A la diferencia entre el valor observado y_i y el valor estimado por el modelo $\hat{y}_i$ le llamaremos error, cuya definición matemática es:

$$e_i = y_i - \hat{y}_i = y_i - (mx_i + b)$$

Por último, observemos que el error puede ser positivo o negativo, dependiendo si el punto está por sobre o bajo la recta, respectivamente.

Curso: Probabilidad y estadística descriptiva
Unidad 2 : Media muestral, dispersión y correlación
Tema: Modelo de regresión lineal
Contenido: Diferencia entre valor observado y estimado

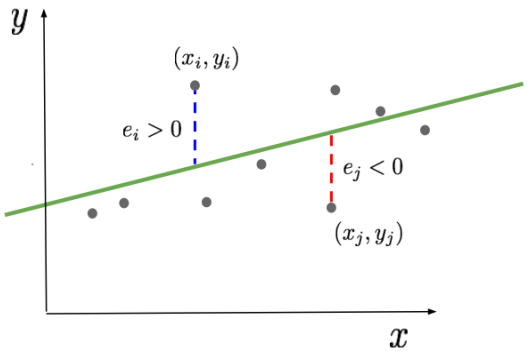

Figura 3: Representación gráfica del error junto a un caso donde es positivo (color azul) y otro donde es negativo (rojo).

Ahora, si no consideramos los signos de los errores, estos se pueden interpretar en el gráfico de dispersión como las distancias verticales de los puntos a la recta. Una forma de encontrar la recta que mejor se ajusta a un conjunto de datos, es encontrar la recta para la que se cumple que las distancias verticales de cada uno de los puntos a la recta son lo más pequeñas posibles.

A continuación, es posible observar la explicación previa de manera gráfica.

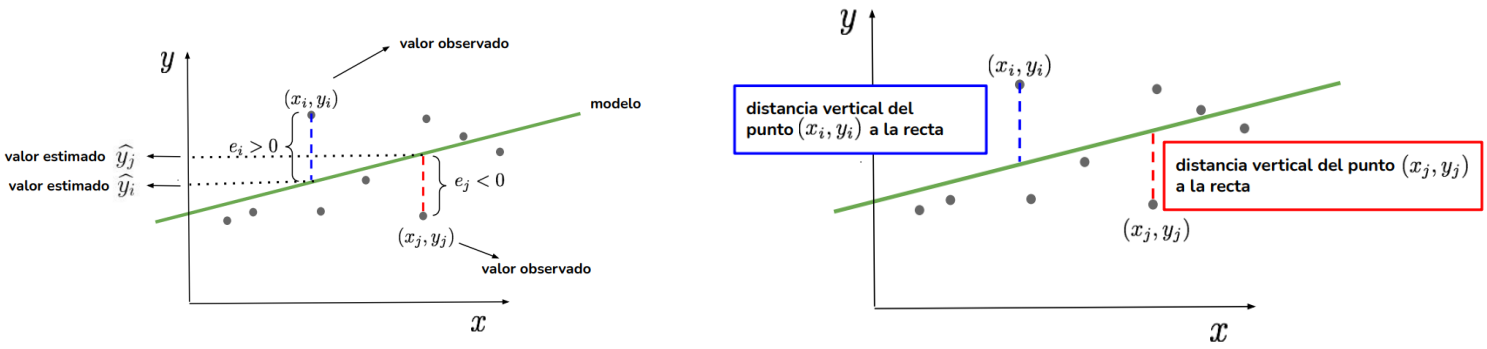

Figura 4: Representación gráfica de los errores, donde en el caso azul existe un error positivo y en el caso rojo uno negativo.

Esta idea intuitiva la podemos formalizar definiendo una medida para el error de todos los datos en conjunto. Como convención, se utiliza la suma de los cuadrados de los errores, es decir, la suma de los cuadrados de las diferencias entre el valor observado y estimado por el modelo:

$$e_1^2 + e_2^2 + \cdots + e_n^2 = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + \cdots + (y_n - \hat{y}_n)^2$$

SÍNTESIS

- Para encontrar un procedimiento que considere a todos los puntos para obtener la recta que mejor se ajuste a estos, partimos definiendo el error e_i entre el valor observado y_i y el valor estimado por el modelo $\hat{y}_i$ como la diferencia entre ellos:

$$e_i = y_i - \hat{y}_i = y_i - (mx_i + b)$$

- El error puede ser positivo o negativo, dependiendo si el punto está por sobre o bajo la recta, respectivamente.

- Si no consideramos los signos de los errores, estos se pueden interpretar en el gráfico de dispersión como las distancias verticales de los puntos a la recta.

- Una forma de encontrar la recta que mejor se ajusta a un conjunto de datos, es encontrar la recta para la que se cumple que las distancias verticales de cada uno de los puntos a la recta son lo más pequeñas posibles.

- Esta idea intuitiva la podemos formalizar definiendo una medida para el error de todos los datos en conjunto. Como convención, se utiliza la suma de los cuadrados de los errores, es decir, la suma de los cuadrados de las diferencias entre el valor observado y estimado por el modelo.