

15 Regresión Lineal Simple

15.1 Introducción

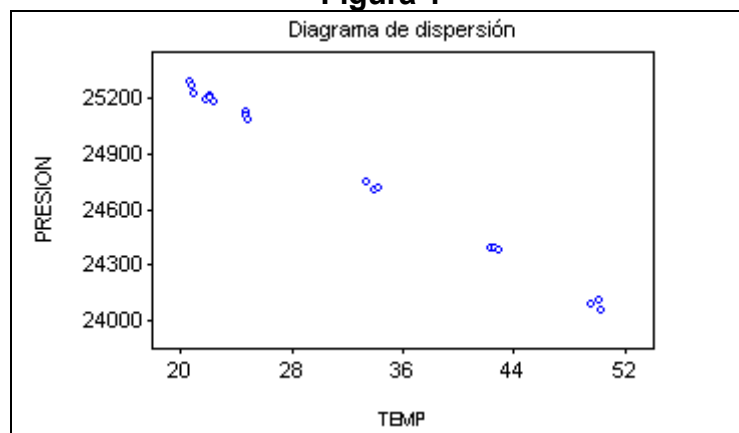
En muchos problemas científicos interesa hallar la relación entre una variable (Y), llamada variable de respuesta, ó **variable de salida**, ó variable dependiente y un conjunto de variables ($X_1, X_2, \dots$) llamadas variables explicativas, ó variables independientes ó **variables de entrada**.

Consideraremos el caso más simple que consiste en describir la relación entre dos variables continuas mediante una recta. Aún cuando el análisis incluya eventualmente más de una variable explicativa, es habitual mirar inicialmente una variable explicativa por vez.

Ejemplo. Interesa estudiar la relación entre la presión (bar) de transición de Bismuto I-II con la temperatura ($^{\circ}\text{C}$)

TEMP	PRESION	TEMP	PRESION	TEMP	PRESION	TEMP	PRESION
20.8	25276	22.1	25187	25	25080	42.7	24394
20.9	25256	22.4	25177	33.5	24750	42.9	24384
21	25216	22.5	25177	34	24701	49.7	24077
21.9	25187	24.8	25112	34.2	24716	50.1	24106
22.1	25217	24.8	25093	42.5	24374	50.3	24057

Figura 1



Vemos que la presión de transición de Bismuto I-II, decrece a medida que aumenta la temperatura, observamos una tendencia lineal decreciente aunque los puntos del diagrama de dispersión no están “perfectamente alineados”.

15.2 Puntos sobre una recta

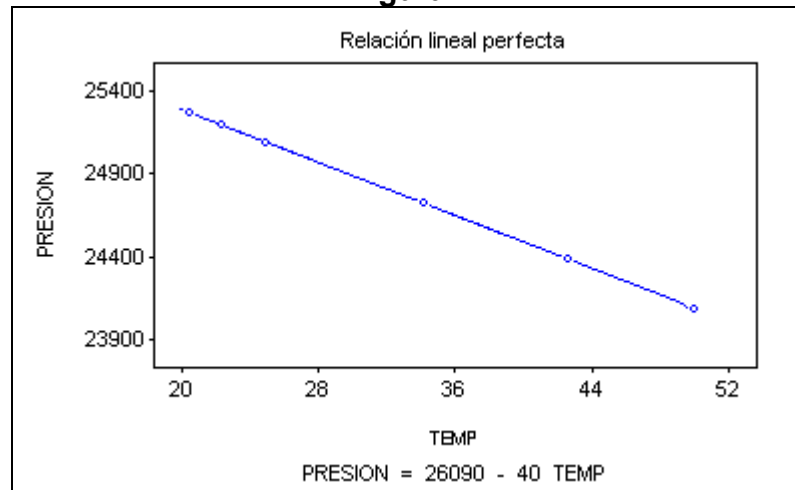
Diremos que la relación entre dos variables X e Y es “perfectamente lineal”, si todos los pares de valores observados (x_i, y_i) de dichas variables satisfacen la ecuación de una recta:

$$y_i = \alpha + \beta x_i \quad (1)$$

En esta expresión α y β son constantes: α es la ordenada al origen y β la pendiente.

Decimos que X es una **variable predictora** de Y , ecuación (1). El valor i del subíndice indexa las observaciones: $i = 1, 2, 3, \dots, n$. Para el ejemplo y_1 representa el valor de la presión obtenido para la temperatura x_1 .

Figura 2



Gráficamente, (1) define una línea recta, donde:

- α (la *ordenada al origen*) es el punto donde la recta corta al eje vertical y
- β (la *pendiente*), indica cuantas unidades cambia y cuando x aumenta 1 unidad.
- Si β positivo la recta sube β unidades por cada aumento de x en 1 unidad. Si β es negativo la recta cae cuando x aumenta. Si $\beta = 0$ la recta es horizontal.

Figura 3 a

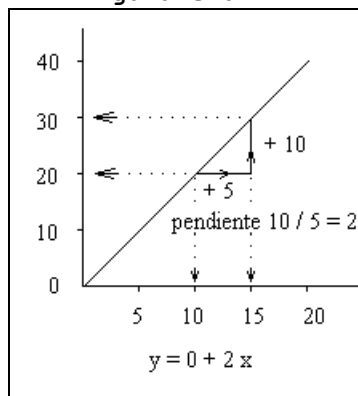
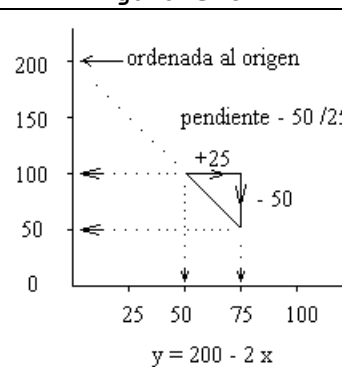


Figura 3 b



La figura 3 muestra dos ejemplos hipotéticos.

Si la relación entre X e Y es perfectamente lineal y conocemos los valores α y β , la ecuación (1) permite predecir qué valor de Y corresponde a cualquier valor de X. Más aún, dos pares de datos son suficientes para determinar los parámetros α y β , de la misma manera que **dos puntos y una regla alcanzan para dibujar una línea recta**. La relación entre datos reales es rara vez tan simple.

15.3 Modelo de Regresión Lineal Simple

En forma más realista podríamos plantear que el **valor esperado** (la media poblacional) de Y, más que los valores individuales, **cambia linealmente con X**:

$$E[Y_i / X_i = x_i] = \alpha + \beta x_i, \quad (2)$$

donde α es igual a la media poblacional de Y cuando $X = 0$. Con un aumento de una unidad en X se obtiene un aumento de la media poblacional de Y en β unidades. Este tipo de modelos tiene muchas aplicaciones prácticas.

En el caso de la presión y la temperatura el modelo dice que la **media poblacional** de las mediciones de la presión para una temperatura fija está dada por

$$\alpha + \beta \text{ TEMP}$$

Otras cosas, además de X , causan que los valores observados de Y_i varíen alrededor de la media de todos los valores de Y cuando X toma el valor x , $E[Y/X=x]$. Esas otras cosas son lo que determinan el error (de medición en nuestro ejemplo) ε_i .

$$\varepsilon_i = Y_i - (\alpha + \beta x_i) = Y_i - E[Y_i / X_i = x_i] \quad (3)$$

El valor de Y_i es igual a la media más un error:

$$\begin{aligned} Y_i &= E[Y_i / X_i = x_i] + \varepsilon_i \\ &= \alpha + \beta x_i + \varepsilon_i. \end{aligned}$$

Por lo tanto, otra forma de expresar el modelo lineal dado en (2) es: los valores de la variable respuesta se encuentran relacionados linealmente con la variable explicativa más un error. Tenemos así el siguiente

Modelo de regresión lineal simple

$$Y_i = \alpha + \beta x_i + \varepsilon_i. \quad (4)$$

Si nos interesa predecir PRESION a partir de TEMP (tabla1), llamaremos a la primera variable **respuesta** y a la segunda variable **explicativa** o **predictora**. La variable respuesta siempre se grafica en el **eje vertical**, o **eje Y**, y la variable predictora en el **eje horizontal**, o **eje X**, como muestra el **diagrama de dispersión** de la figura 1.

El problema consiste en ajustar una recta que represente al conjunto de datos de la mejor manera, para obtener la predicción de Y para cualquier valor de X . Hay muchas maneras de evaluar si una recta representa bien al conjunto de datos. El enfoque tradicional consiste en hallar la recta que en promedio tenga la menor **distancia vertical**, **residuo**, al cuadrado a cada uno de los puntos. Este procedimiento se llama método de **Cuadrados Mínimos (CM)** y lo describiremos en la Sección 14.5.

UNWEIGHTED LEAST SQUARES LINEAR REGRESSION OF PRESION				
PREDICTOR VARIABLES	COEFFICIENT	STD ERROR	STUDENT'S T	P
CONSTANT	26079.9	11.9034	2190.97	0.0000
TEMP	-39.8935	0.35843	-111.30	0.0000

La **recta de regresión** ($\hat{y} = a + b x$) obtenida por el método de cuadrados mínimos para los datos de la tabla 1 es:

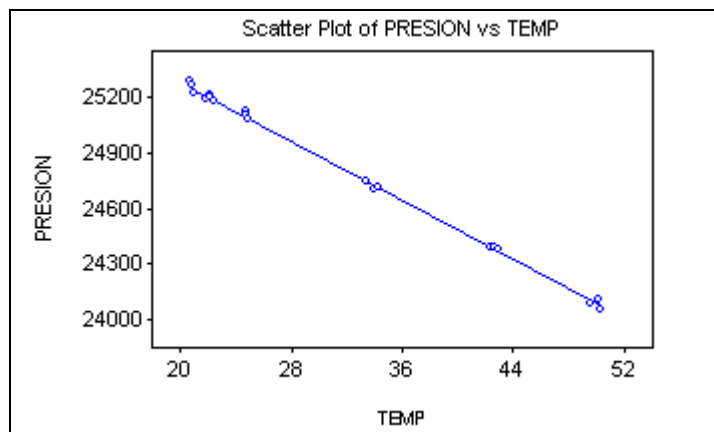
$$\text{PRESION ajustada } (\hat{y}) = 26079.9 - 39.8935 \text{ TEMP} \quad (5)$$

El valor *ajustado* ($\hat{y}$) puede utilizarse de dos maneras distintas:

- como *estimador* de la media poblacional de Y para cada x fijo, en este caso como estimador de la media de la presión de transición para una temperatura fija.
- como *predictor* de un valor futuro de Y para un valor fijo de x.

La diferencia entre a) y b) se encuentra únicamente en la varianza de $\hat{y}$.

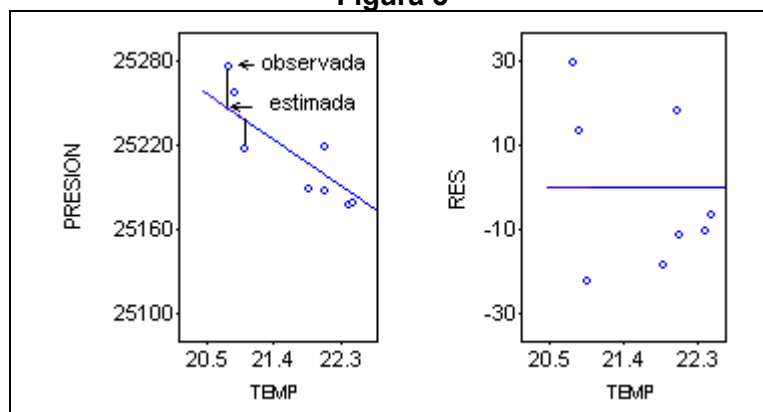
Figura 4. Diagrama de dispersión junto con la recta ajustada.



15. 4 Residuos

El **residuo** de un punto a una recta en un diagrama de dispersión es la **distancia vertical** del punto a dicha recta. La figura 5 muestra el diagrama de dispersión de los datos junto con la recta ajustada y el diagrama de dispersión de los residuos vs. la temperatura para los primeros 8 datos.

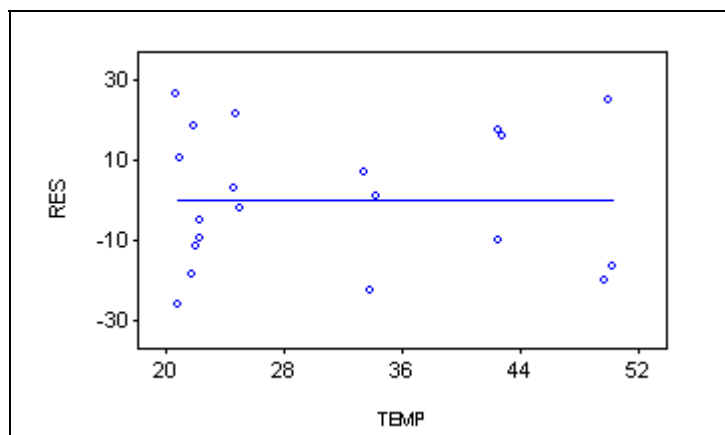
Figura 5



Algunos residuos son positivos, la presión observada está por encima de la recta, y otros son negativos, la presión observada está por debajo de la recta. La suma de todos los residuos es cero.

La figura 6 muestra el diagrama de dispersión de los residuos vs. la temperatura del conjunto de datos completo.

Figura 6



Para la primera medición TEMP = 20.8, PRESION = 25276. El residuo = 25.8799 es positivo y se obtiene como

$$\begin{aligned}\text{residuo} &= \text{valor de } Y \text{ observado} - \text{valor de } Y \text{ estimado por la recta} \\ &= y - \hat{y} = \text{PRESION} - \{26079.9 - 39.8935 \text{ TEMP}\} \\ &= 25276 - \{26079.9 - 39.8935 * 20.8\} \\ &= 25.8799\end{aligned}$$

Para la tercera medición TEMP= 21, PRESION = 25216. El residuo= -26.1413 es negativo.

15.5 El Método de Cuadrados Mínimos

La **suma de los cuadrados de los residuos (RSS)** da una medida de la "bondad de ajuste" de la recta. Cuanto más pequeño es ese número tanto mejor es el ajuste.

Hemos observado valores de dos variables, X (TEMP) e Y (PRESION), y hemos realizado una "regresión de Y sobre X ", obteniendo una recta que da un valor "ajustado" estimado de Y ($\hat{y}$, y "sombrero") para cada valor de la variable X .

Un estudio comienza por un modelo lineal (4) porque existe una teoría que lo sugiere o porque se desea comenzar de manera simple. En cualquiera de los dos casos, nos interesa obtener los mejores estimadores de los parámetros α y β . Si llamamos a y b a nuestros estimadores, la ecuación de la recta estimada es:

$$\hat{y}_i = a + b x_i ,$$

donde $\hat{y}_i$ (y_i "sombrero") indica el valor ajustado (o predicho) de la variable Y para el caso i (es el valor de la ordenada para x_i sobre la recta ajustada) (ver figura 5).

Los residuos e_i , la contraparte muestral de los errores (ε_i), son las diferencias entre el valor observado y el valor predicho:

$$\begin{aligned}e_i &= y_i - \hat{y}_i \\ &= y_i - (a + b x_i).\end{aligned}$$

Los residuos miden el error de predicción. Como hemos visto, si el valor observado es mayor que el valor predicho ($y_i > \hat{y}_i$) el residuo es positivo; en caso contrario es negativo.

Con una predicción perfecta ($y_i = \hat{y}_i$) resulta un residuo nulo. La suma de los cuadrados de los residuos (RSS) refleja la precisión y exactitud global de nuestras predicciones:

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - a - bx_i)^2 \quad (6)$$

Cuanto más cerca estén los valores observados de los predichos tanto menor será RSS.

El **método de Cuadrados Mínimos** (CM) consiste en elegir a y b de manera que la suma de cuadrados de los residuos (RSS) sea lo más pequeña posible.

¿Cómo hallamos a y b ?

$$\begin{aligned} \frac{\partial \left(\sum_{i=1}^n (y_i - a - bx_i)^2 \right)}{\partial a} &= -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \\ \frac{\partial \left(\sum_{i=1}^n (y_i - a - bx_i)^2 \right)}{\partial b} &= -2 \sum_{i=1}^n x_i (y_i - a - bx_i) = 0 \end{aligned} \quad (7)$$

Las ecuaciones anteriores definen los estimadores de los parámetros α y β resultan de derivar (6) con respecto a a y a b . Se trata de dos ecuaciones lineales con dos incógnitas cuyas soluciones son

$$\begin{aligned} b &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ a &= \bar{y} - b\bar{x} \end{aligned} \quad (8)$$

Observaciones

- De la primera ecuación de (7) tenemos que la suma de los residuos es 0.
- La segunda ecuación de (8), nos dice que la recta de cuadrados mínimos pasa por $(\bar{x}, \bar{y})$, ya que $\bar{y} = a + b\bar{x}$.

Podemos pensar al método de cuadrados mínimos como fijando un punto, dado por el promedio de los valores de la variable explicativa (x 's) y el promedio de los valores de la variable respuesta (y 's) y luego girando la recta que pasa por ese punto elegimos la que deja en promedio, en forma pareja, tantos valores observados por arriba como por abajo.

Ninguna otra recta tendrá, para el mismo conjunto de datos, una RSS tan baja como la obtenida por CM. En este sentido, el método de mínimos cuadrados brinda la solución que mejor ajusta a ese conjunto de datos.

Advertencia: en general no pueden realizarse predicciones fuera del rango de valores observados de la variable independiente.

¿Que nos dice la ecuación de la recta (PRESION *ajustada* ($\hat{y}$) = 26079.9 - 39.8935 TEMP) ajustada?:

- La estimación de la variable PRESION, obtenida a partir de la ecuación de regresión ajustada, es el **valor predicho** de PRESION.

- Para cualquier valor de la variable TEMP un aumento en un grado de la temperatura produce una reducción de 39.8935(bar) en la **presión media** (“verdadera”) de transición de Bismuto I-II.

El método de CM permite estimar una recta a partir de un conjunto de datos. Si estos datos son una muestra “adecuada” de una población, la recta nos permite extender resultados a dicha población. Ciertas características de los datos podrían invalidar los resultados del método.

15.6 Supuestos

Antes de utilizar el análisis de regresión y considerar medidas de incerteza o dispersión, es necesario conocer los supuestos en los que se basa el método. Veremos primero cuáles son esos supuestos y luego qué procedimientos pueden utilizarse para validarlos.

15.6.1 Descripción de los supuestos

Supuesto a: Normalidad de los errores.

Para cada valor x , de la variable predictora X , la variable respuesta Y debe tener distribución Normal

Por ejemplo, si se cumple este supuesto, la presión de transición (Y) es una variable aleatoria Normal con media μ_x que depende de x (temperatura).

Supuesto b: Linealidad

La media de la variable Y varía linealmente con X .

Si pasar de 21 a 22 °C no fuera lo mismo que pasar 41 a 42 °C respecto del cambio de la presión de transición, este supuesto no se cumpliría.

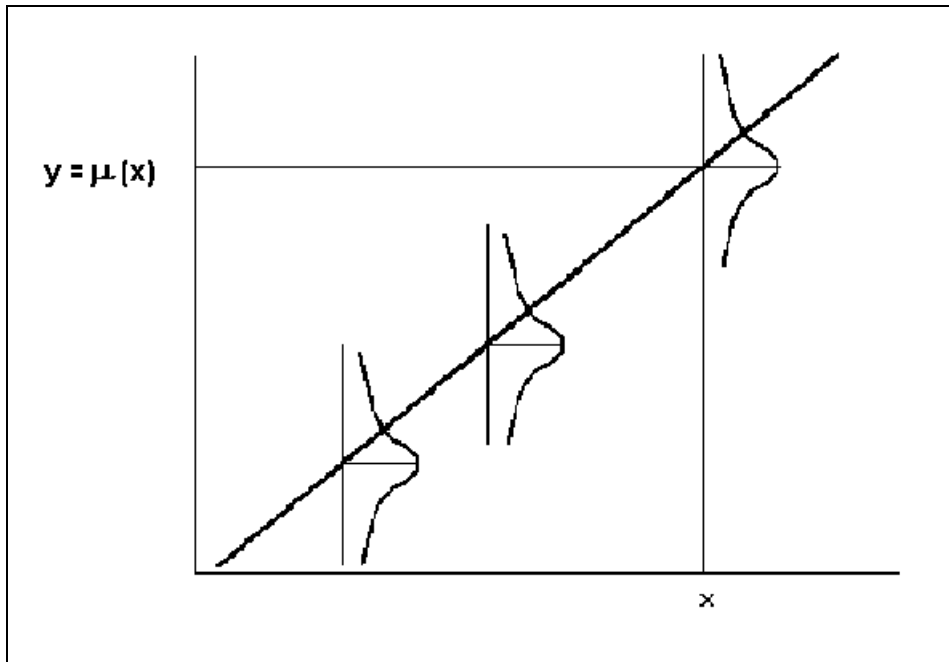
Supuesto c: Homoscedasticidad

La variabilidad de Y , que es medida por su varianza (σ^2), o por su desvío estándar (σ), debe ser la misma para cada valor x de la variable X .

Este supuesto no se cumpliría por ejemplo si a medida que aumenta (o disminuye) la temperatura los valores de la presión de transición de Bismuto I-II estuvieran comprendidos en un rango más amplio.

No sabemos si los supuestos se satisfacen, ni conocemos los verdaderos valores de los parámetros α y β .

Figura 7. Supuestos de Normalidad, linealidad y homoscedasticidad



La figura 7 representa dos variables para las cuales se satisfacen los supuestos de linealidad ($\mu(x) = \alpha + \beta x$, la media de la variable Y crece linealmente con x), normalidad y homocedasticidad de los errores.

Supuesto d: Independencia de los errores

Hemos visto que cuando dos variables son independientes su correlación es cero, en general la recíproca no es cierta pero bajo el supuesto de normalidad el supuesto de independencia de los errores se reduce a que no estén correlacionados ($\text{corr}(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$).

Por ejemplo, si las presiones de transición fueron obtenidas en un orden secuencial con la temperatura, podría ocurrir que los errores fueran mayores en temperaturas más bajas que en temperaturas más altas invalidando el supuesto de independencia de los errores.

15.6.2 Validación de los Supuestos

La validación de los supuestos se realiza en base a los datos y a los residuos de los mismos respecto de la recta ajustada. El diagrama de dispersión de los datos permite obtener una impresión sobre el supuesto de linealidad y homocedasticidad. El análisis posterior de residuos permitirá confirmar la impresión inicial y validar los supuestos de Normalidad e independencia.

Veamos algunas estructuras que suelen verse en los diagramas de dispersión de los residuos.