
ESTADÍSTICA APLICADA A LA INVESTIGACIÓN CIENTÍFICA

Unidad 5: Relación entre variables

REGRESIÓN LINEAL SIMPLE

Decir que x e y están relacionadas de manera **determinista** significa que una vez que se sabe el valor de x , el valor de y queda especificado por completo.

Un **modelo determinista** es un formulismo matemático donde las mismas entradas o condiciones iniciales producirán invariablemente las mismas salidas o resultados, no contemplándose la existencia del azar, o incertidumbre en el proceso que se ha modelado.

Ejemplos de modelos deterministas:

- 1) Supóngase que se decide rentar un auto por un día y la renta es de \$ 2.500, más \$ 30 por km recorrido. Si x es el número de km recorridos e y el cargo de la renta, entonces $y = 2.500 + 30x$. Si se recorren 100 km ($x=100$), entonces

$$y = 2.500 + 30(100) = \$ 5.500$$

- 2) Si la velocidad inicial de una partícula es v_0 y experimenta una aceleración constante a , entonces la distancia recorrida es

$$y = v_0 x + \frac{1}{2} a x^2$$

donde x = tiempo

El **análisis de regresión** es la parte de la estadística que se ocupa de investigar la relación entre dos o más variables relacionadas en una forma **no determinística**.

Si dos variables no están relacionadas de manera determinista, entonces con un valor fijo de x , la segunda variable tendrá un valor aleatorio.

Ejemplos de relación no determinista:

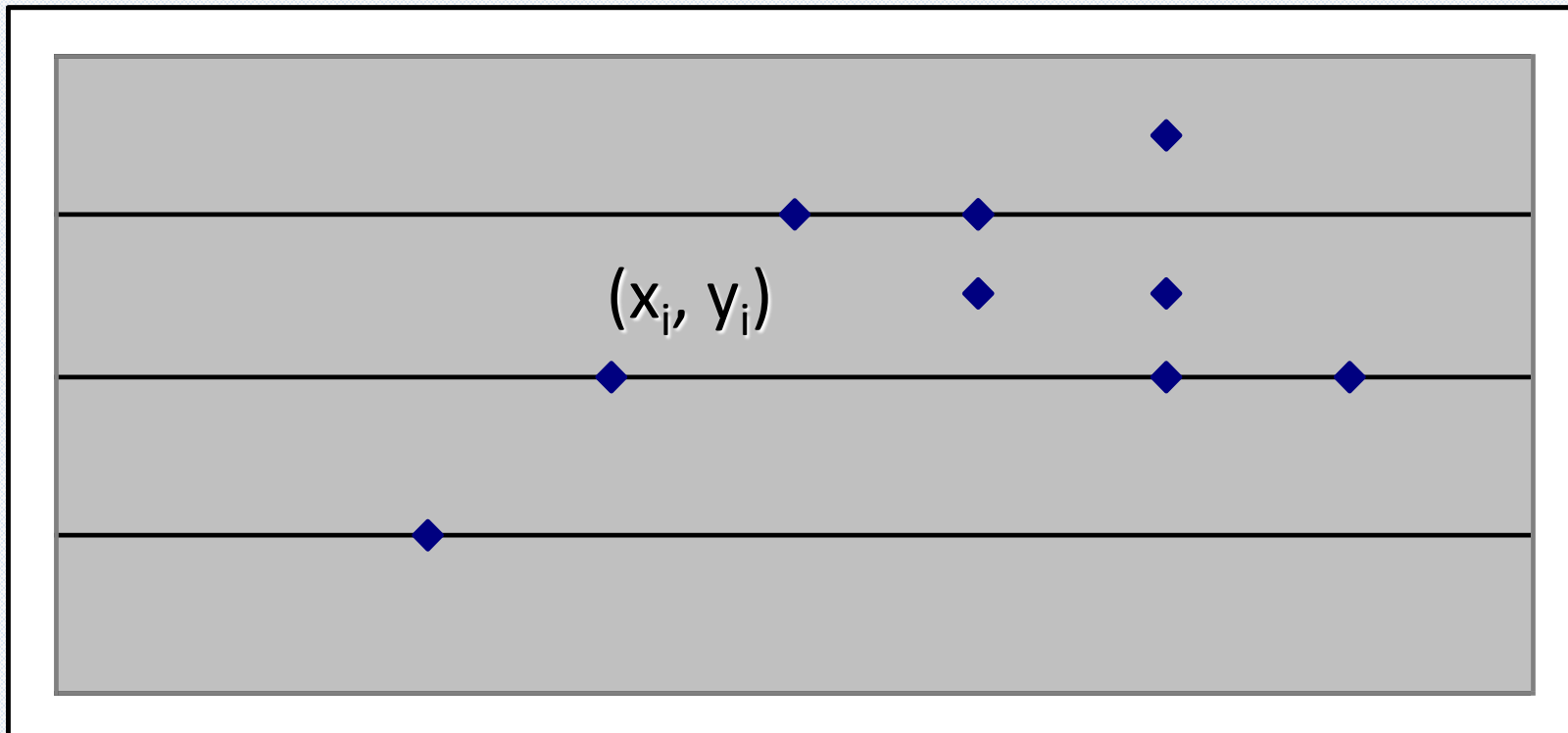
- 1) La estatura (E) y el peso (P) de los estudiantes de cierta universidad.
- 2) El rendimiento (Y) de un proceso químico que depende de la temperatura (T).

La variable cuyo valor fija el experimentador será denotada por x , que se la conoce como **variable independiente, predictora o explicativa**. La segunda variable será aleatoria, es denotada por y , conocida como **variable dependiente, de respuesta o criterio**.

Normalmente se realizarán observaciones para varios escenarios de la variable independiente. Sean $x_1, x_2, \dots, x_n$ los valores de la variable independiente para la que se realizan las observaciones, y sean Y_i y y_i , respectivamente, la variable aleatoria y el valor observado asociado con los x_i . Los **datos bivariantes** disponibles se componen entonces de los n pares $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.

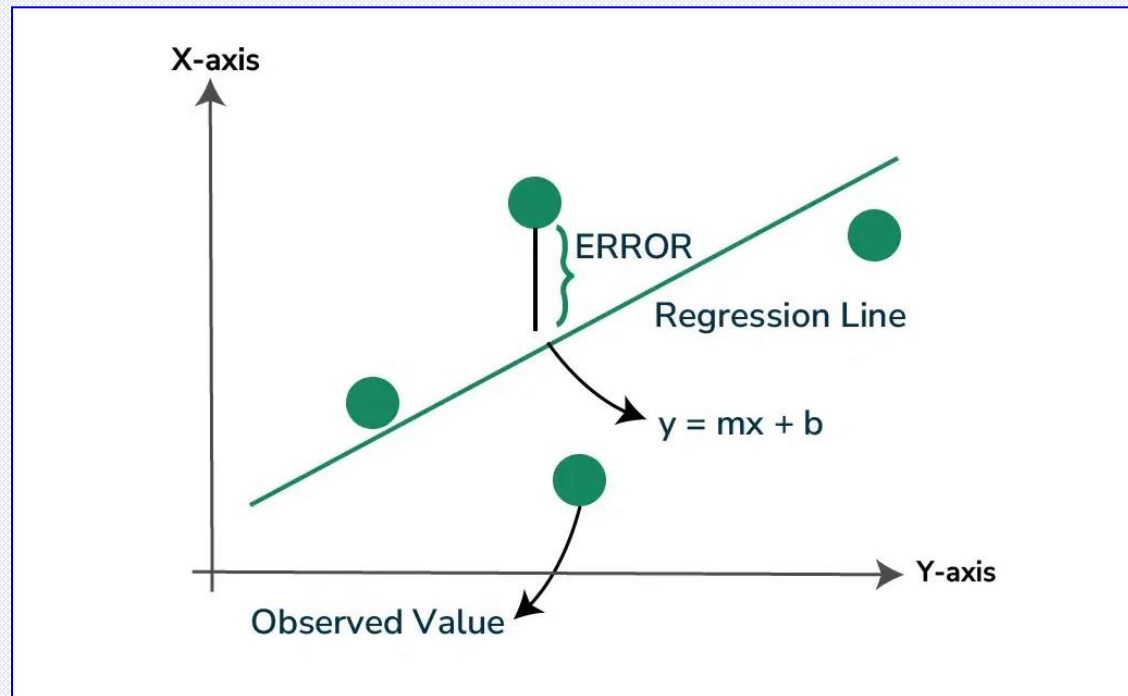
Diagrama de dispersión o dispersograma

Un primer paso en el análisis de regresión que implica dos variables es construir una **gráfica de puntos** de los datos observados. En esta gráfica, cada (x_i, y_i) está representado como un punto colocado en un sistema de coordenadas bidimensional.



MÉTODO DE MÍNIMOS CUADRADOS

El método de mínimos cuadrados es un enfoque matemático popular utilizado en el ajuste de datos, el análisis de regresión y el modelado predictivo. Ayuda a encontrar la línea o curva de mejor ajuste que minimice la suma de las diferencias al cuadrado entre los puntos de datos observados y los valores predichos. Esta técnica se utiliza ampliamente en estadística, aprendizaje automático e ingeniería.



RECTA DE MÍNIMOS CUADRADOS

Cuando **dos variables** tienen una **relación lineal**, el diagrama de dispersión tiende a estar agrupado alrededor de la **recta de mínimos cuadrados**.

Ejemplo: experimento hipotético con un resorte

Un resorte se cuelga verticalmente con un extremo fijo, y los pesos se cuelgan uno tras otro del otro extremo. Después de colgar cada peso se mide la longitud del resorte. Sean $x_1, \dots, x_n$ los pesos, y sea l la longitud del resorte bajo la carga x_i .

La ley de Hooke, en el contexto de un resorte, puede ser establecida como:

$$l = \alpha + \beta x \quad (1)$$

donde α representa la longitud del resorte cuando no tiene carga y β es la constante del resorte.

Sea Y la longitud **medida** del resorte bajo carga x_i . Debido al error de medición, Y será diferente de la longitud verdadera I . Se escribe

$$Y = I + \varepsilon \quad (2)$$

donde ε_i constituye el error en la i -ésima medición.

A partir de las ecuaciones (1) y (2)

$$Y = \alpha + \beta X + \varepsilon \quad (3)$$

La **ecuación (3)** se denomina **modelo lineal**; Y se llama variable dependiente, X se conoce como variable independiente, α y β son los **coeficientes de regresión**, y ε se denomina **error aleatorio**.

Si no hubiera error en la medición, los puntos que relacionan los diferentes x_i e y_i se encontrarían en una línea recta con pendiente β e intercepto α . Debido al error de medición, α y β no se pueden determinar exactamente, pero se pueden **estimar** a partir de la recta de mínimos cuadrados.

MODELO DE REGRESIÓN LINEAL SIMPLE

DEFINICIÓN:

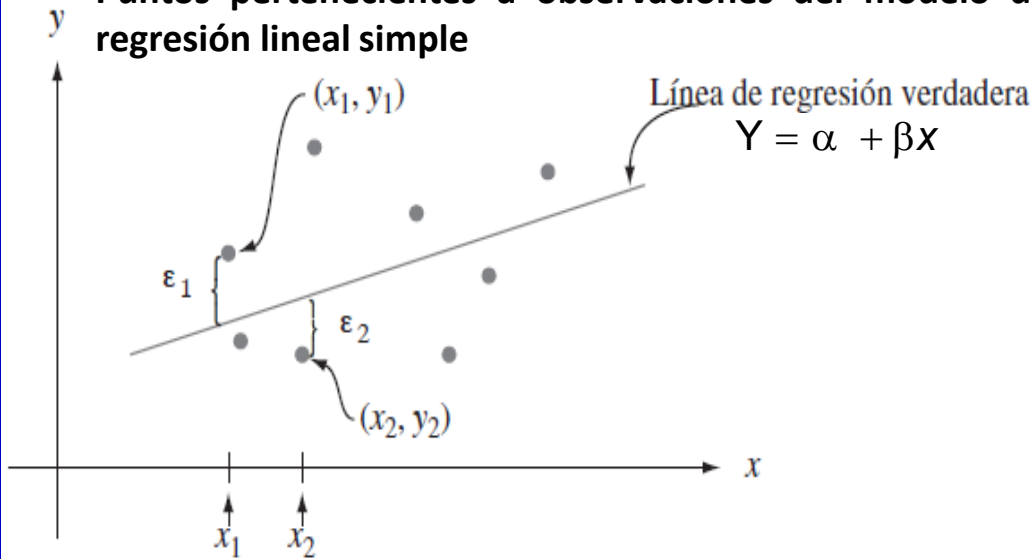
Existen parámetros α , β y σ^2 de tal manera que con cualquier valor fijo de la *variable independiente* X , la *variable dependiente* Y está relacionada con X por conducto de la ecuación de modelo

$$Y = \alpha + \beta X + \varepsilon$$

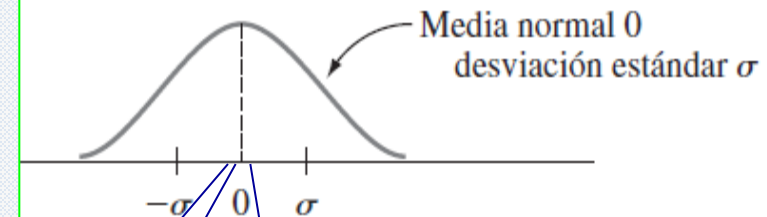
la cantidad ε en la ecuación de modelo es una variable aleatoria, que se supone **normalmente distribuida** con $E(\varepsilon) = 0$ y $V(\varepsilon) = \sigma^2$

La variable ε se conoce como **término de error aleatorio** o **desvío aleatorio** en el modelo. Sin ε , cualquier par observado (x, y) correspondería a un punto que queda exactamente sobre la línea $Y = \alpha + \beta x$, llamada **línea de regresión** (o de **población**) **verdadera**.

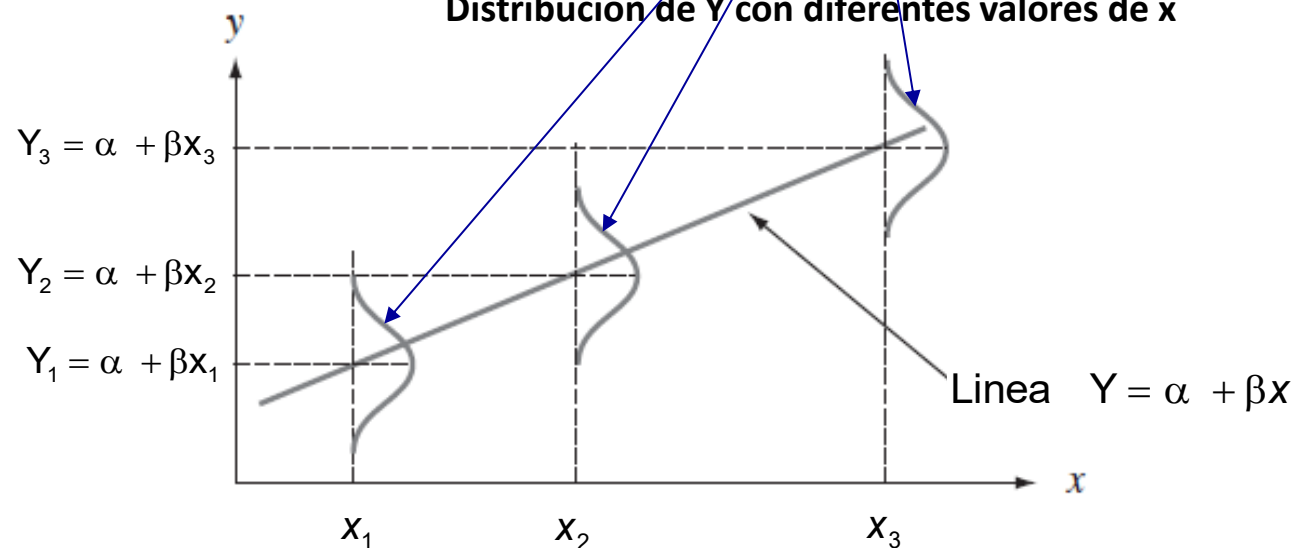
Puntos pertenecientes a observaciones del modelo de regresión lineal simple



Distribución de ϵ



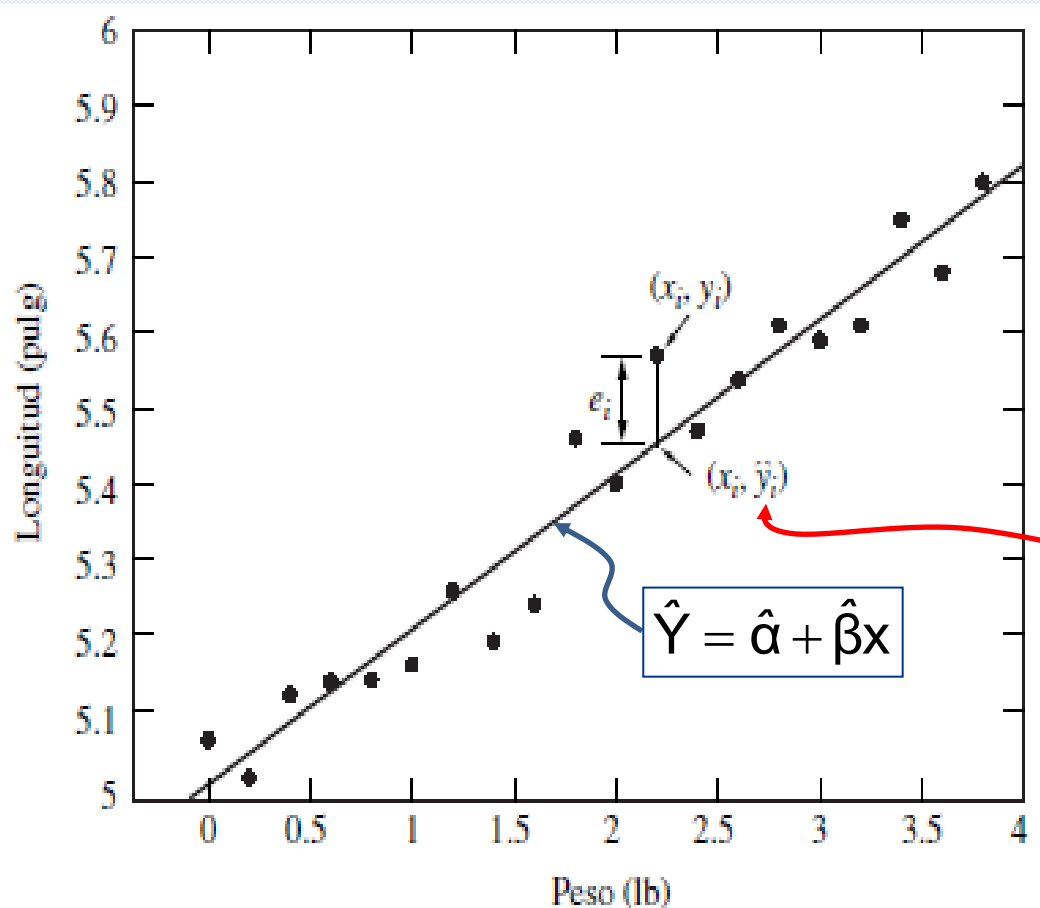
Distribución de Y con diferentes valores de x



Continuando con el ejemplo

Longitudes *medidas* de un resorte bajo diferentes cargas

Peso (lb)	Longitud medida (pulg)	Peso (lb)	Longitud medida (pulg)
x	y	x	y
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80



Gráfica de longitudes medidas de un resorte contra carga

Estimadores de coeficientes de mínimos cuadrados $\left\{ \begin{array}{l} \hat{\alpha} = a \\ \hat{\beta} = b \end{array} \right.$

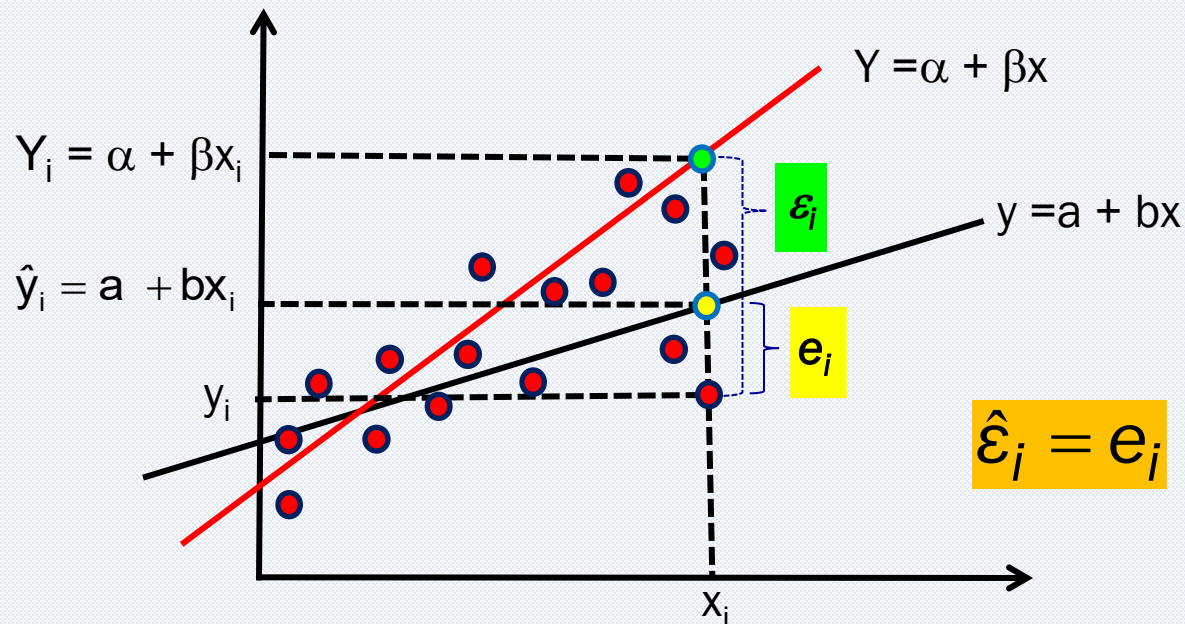
Para cada punto de datos (x_i, y_i) la distancia vertical al punto $(x_i, \hat{y}_i)$ en la recta de mínimos cuadrados es $e_i = y_i - \hat{y}_i$.

La cantidad $\hat{y}_i = a + bx_i$ se llama **valor ajustado**, y la cantidad e_i se denomina **residuo** asociado con el punto (x_i, y_i) .

El residuo e_i es la diferencia entre el valor observado y_i en los datos y el valor ajustado $\hat{y}_i$ pronosticado por la recta de mínimos cuadrados.

La recta de mínimos cuadrados se define como aquella para la que la **suma de cuadrados de residuos** $\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ **se minimiza**.

En el ejemplo de ley de Hooke hay solamente una variable independiente (peso), ya que es razonable suponer que la única variable que afecta la longitud del resorte es el peso colgado en él. Este tipo de modelos lineales, con sólo una variable independiente, se conocen como modelos de **regresión lineal simple**.



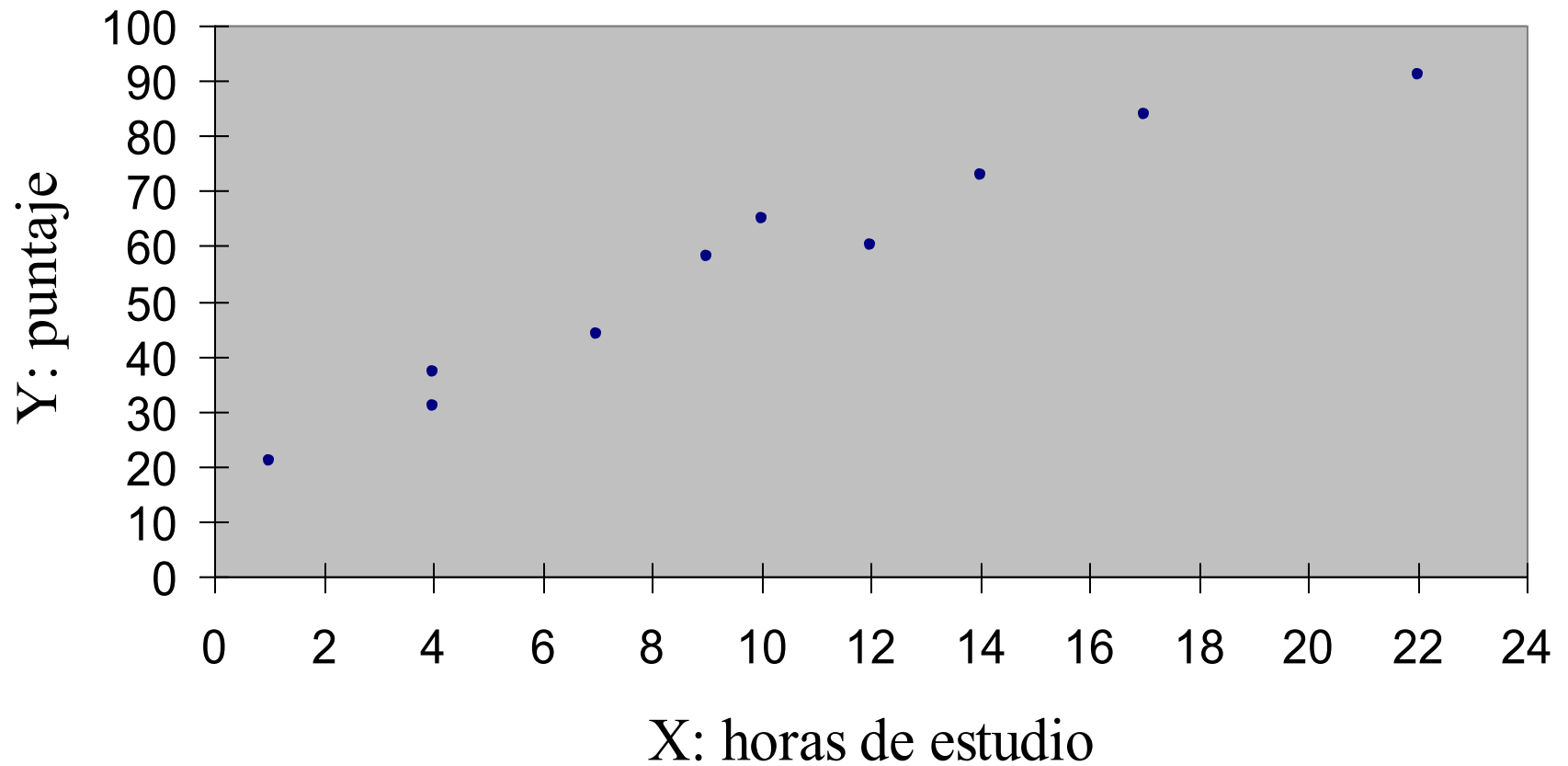
$$y = a + bx \quad (\text{Recta de regresión lineal})$$

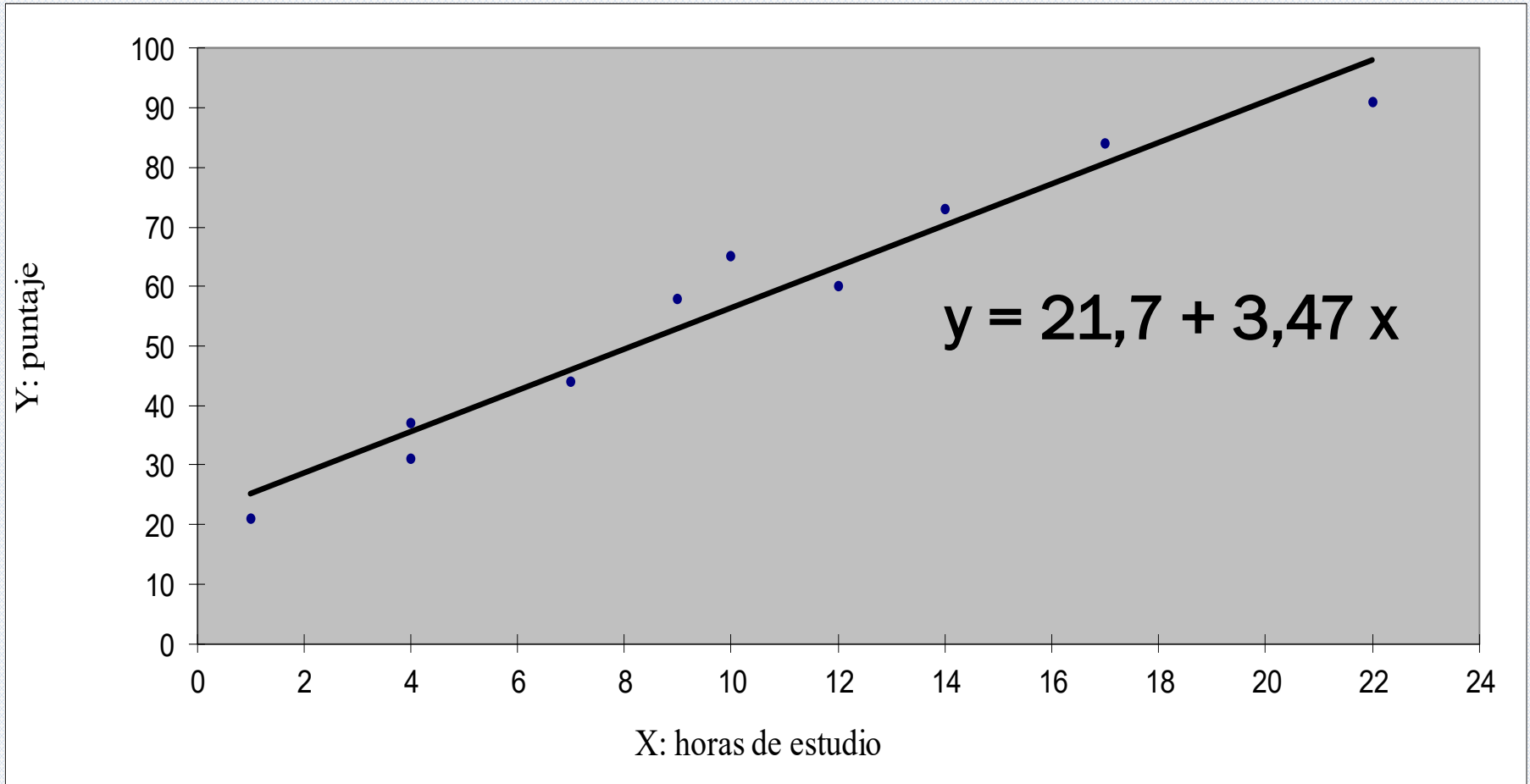
Ejemplo: La siguiente tabla muestra el número de horas que 10 personas ocuparon para estudiar un parcial y su correspondiente puntaje. Hallar la regresión lineal

X : hs de estudio	4	9	10	14	4	7	12	22	1	17
Y : puntaje	31	58	65	73	37	44	60	91	21	84

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = 3,47$$

$$a = \bar{y} - b \bar{x} = 21,7$$



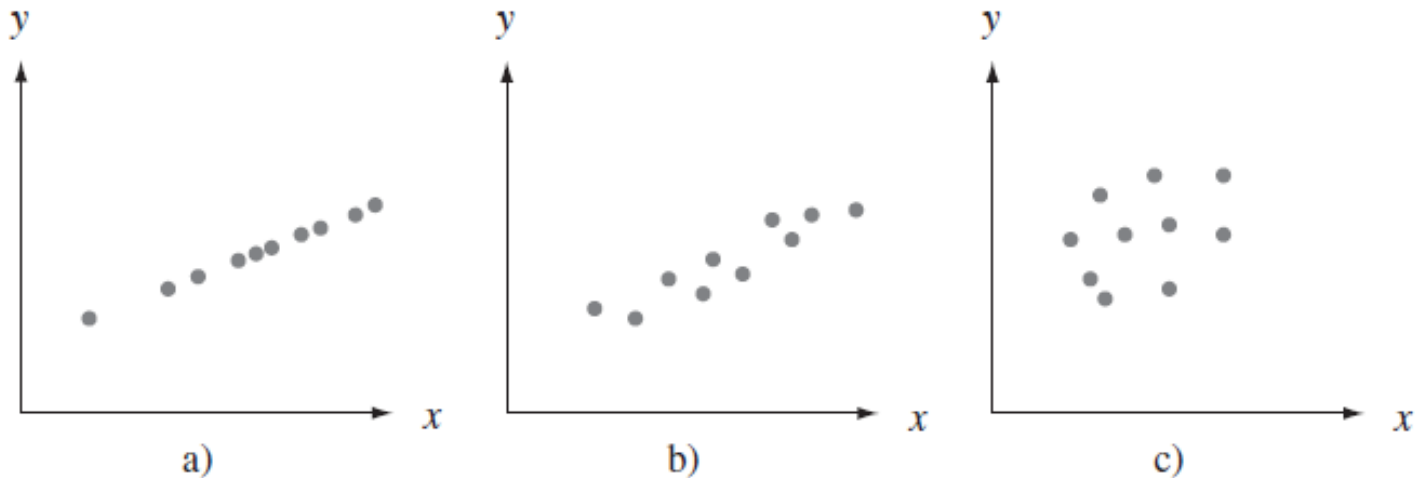


SUPUESTOS

1. La variable X es una variable matemática, es decir no es aleatoria, no está sujeta a errores. Los únicos valores que existen para X son los de la muestra, o sea son fijos los $x_1, x_2, x_3, \dots, x_n$.
2. La variable Y es una variable aleatoria, es la que medimos en el experimento y está sujeta a errores.
3. Las variables $Y_1, Y_2, \dots, Y_n$ se consideran variables aleatorias independientes.
4. Para cada x_i , la distribución de Y_i es normal con $E[Y_i] = \alpha + \beta x_i$ y $V[Y_i] = \sigma^2$, es decir $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$
5. La varianza σ^2 es constante.

Coeficiente de determinación (Medición de la bondad del ajuste)

Un **estadístico de la bondad del ajuste** representa una cantidad que mide qué **tan bien un modelo explica un conjunto específico de datos**. Un **modelo lineal** ajusta bien si hay una fuerte **relación lineal** entre x e y .



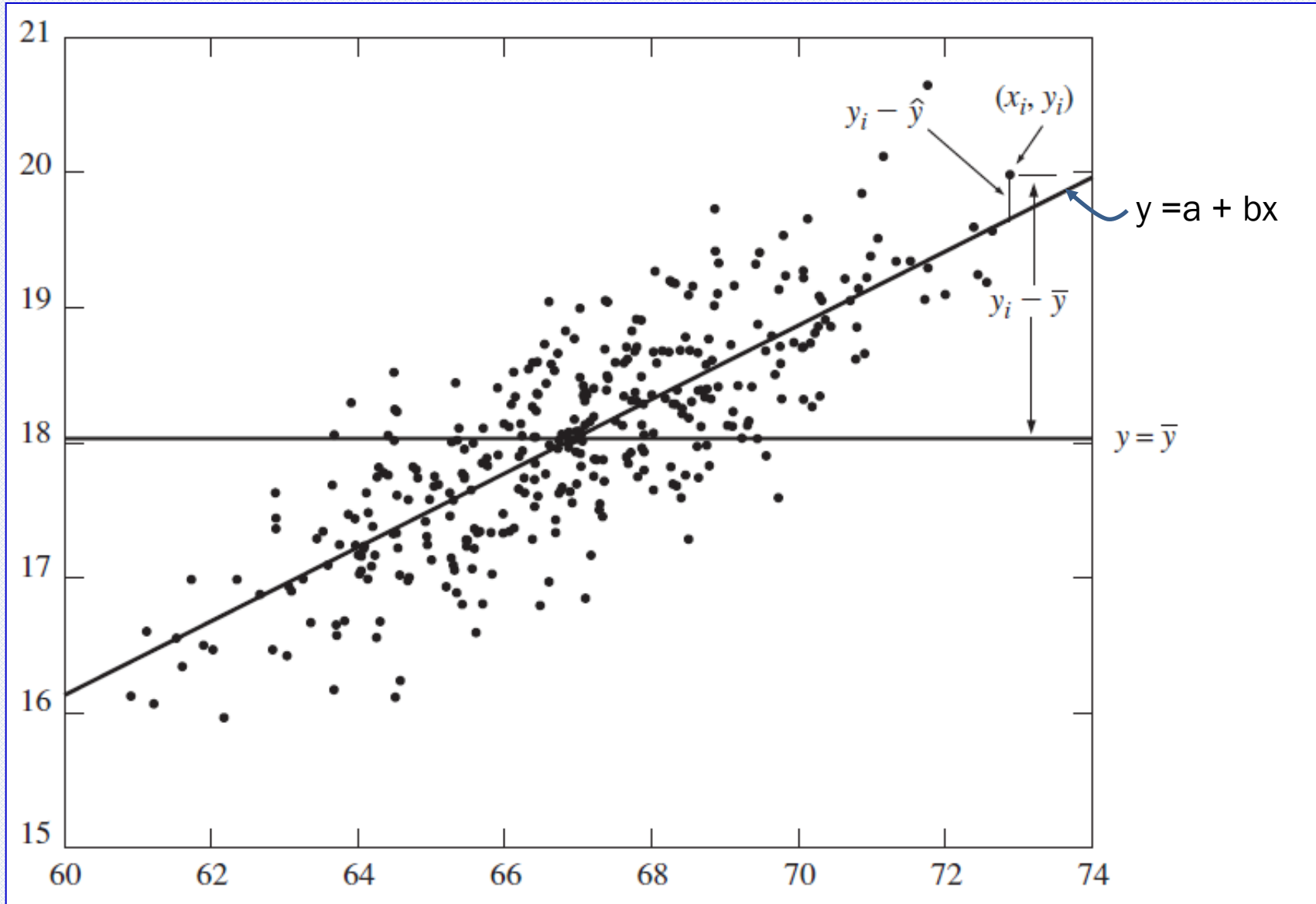
Utilización del modelo para explicar la variación de y : a) datos con los cuales toda la variación es explicada; b) datos con los cuales la mayor parte de la variación es explicada; c) datos con los cuales poca variación es explicada.

La **suma de cuadrados de residuos (SCR)** $\sum e_i^2$ es la suma de los desvíos al cuadrado, con respecto a la recta de mínimos cuadrados $y = a + b x$.
Puede ser interpretada como una medida de cuánta variación de **y** permanece **sin ser explicada por el modelo**, es decir, cuánta no puede ser atribuida a una **relación lineal**.

$$\text{SCR} = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

La **suma total de los cuadrados (STC)** S_{yy} es la suma de los desvíos al cuadrado de los valores **y observados** con respecto a la media muestral $\bar{y}$.
Da una medida cuantitativa de la cantidad de variación total en los valores de **y observados**.

$$\text{STC} = S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$$



$SCR < STC$ (a menos que la recta horizontal sea la recta de mínimos cuadrados)

La razón SCR/STC es la **proporción de variación total que no puede ser explicada por el modelo de regresión lineal simple**, y $1 - SCR/STC$ (un número entre 0 y 1) es la **proporción de variación de "y observada" que es explicada por el modelo**.

DEFINICIÓN:

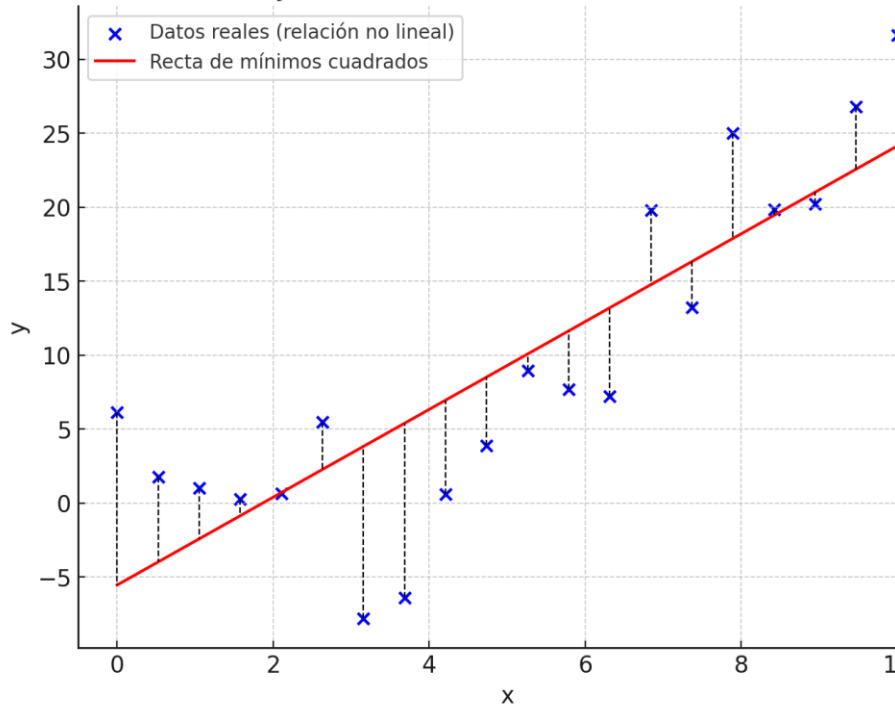
El coeficiente de determinación, denotado por R^2 , está dado por

$$R^2 = 1 - \frac{SCR}{STC}$$

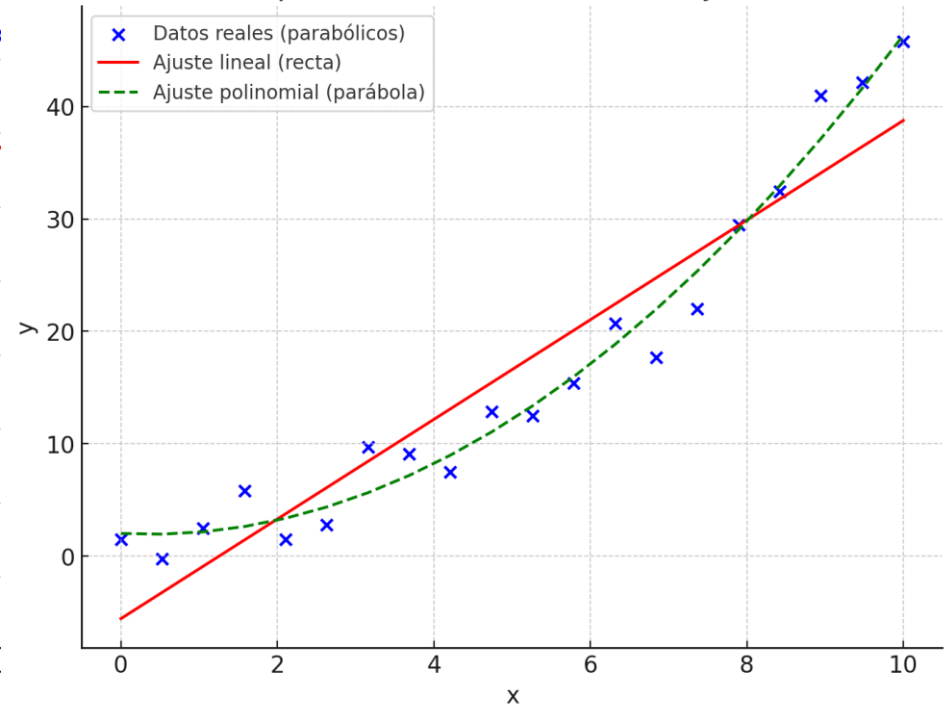
Se interpreta como la *proporción de variación y observada que puede ser explicada por el modelo de regresión lineal simple* (atribuida a una relación lineal aproximada entre y y x).

Valor de R^2 recomendado $\geq 0,8$

Errores al Ajustar una Línea a una Relación No Lineal



Comparación: Recta vs. Curva Ajustada



Recta roja (ajuste lineal): intenta forzar una línea recta sobre los datos parabólicos, lo que genera grandes errores en los extremos.

Curva verde (ajuste polinomial): sigue mucho mejor la forma real de los datos, ya que refleja su naturaleza cuadrática.

Esto muestra que **el método de mínimos cuadrados se puede extender a modelos no lineales**, y al elegir el modelo correcto (en este caso cuadrático), se logra un ajuste mucho más preciso.

Comparación de los **coeficientes de determinación (R^2)**

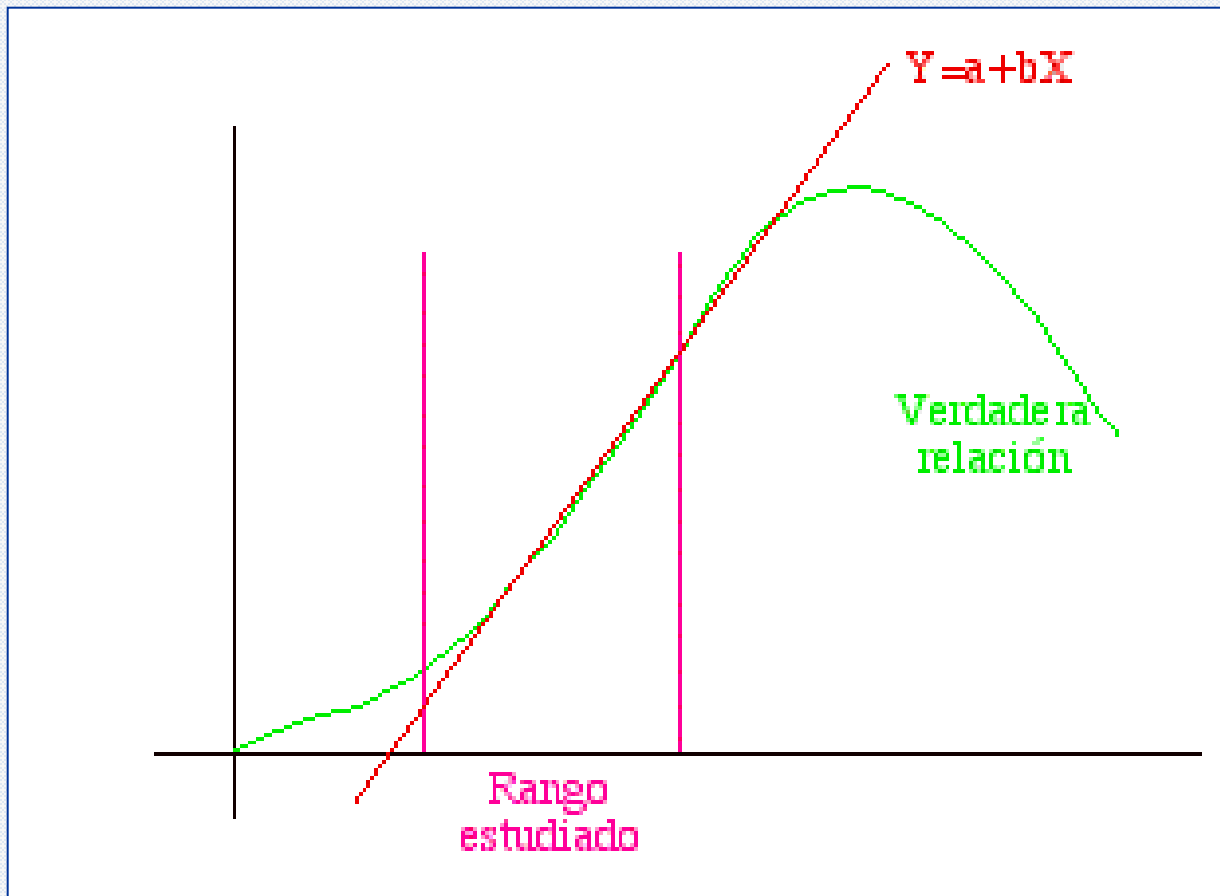
El **R^2** mide qué proporción de la variabilidad de los datos puede explicar el modelo (cuanto más cerca de 1, mejor):

- Para el **ajuste lineal** en datos parabólicos, el R^2 suele ser relativamente bajo (ej. 0.6–0.7), porque la recta no capta bien la curvatura.
- Para el **ajuste polinomial de grado 2**, el R^2 se acerca mucho a 1 (ej. 0.95–0.99), ya que el modelo cuadrático representa adecuadamente la relación verdadera.

En conclusión: el **modelo polinomial ajusta mejor** porque explica mucha más variabilidad en los datos que el ajuste lineal.

EXTRAPOLACIÓN

Es importante resaltar que, cuando se hacen predicciones, no deben extrapolarse los resultados más allá del rango de la variable X utilizado para ajustar el modelo, ya que más allá de ese rango no sabemos qué puede estar ocurriendo.



CORRELACIÓN

Existen muchas situaciones en las que el objetivo al estudiar el comportamiento conjunto de dos variables, y considerar que tanto X como Y son variables aleatorias, y que las mediciones (x_i, y_i) son las observaciones de una población que tiene la función de densidad conjunta $f(x, y)$.

DEFINICIÓN:

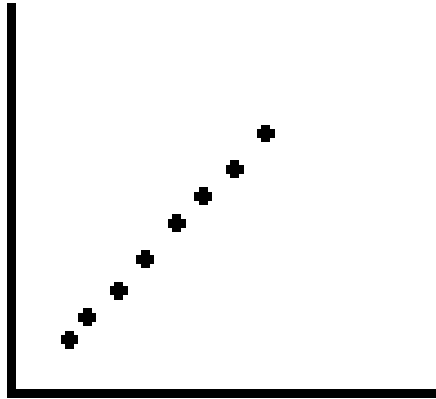
El **coeficiente de correlación muestral** para los n pares $(x_i, y_i), \dots, (x_n, y_n)$ es

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}} \sqrt{S_{yy}}}$$

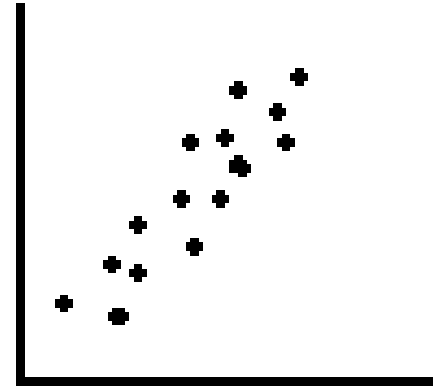
$$\mathbf{-1 \leq r \leq 1}$$

Propiedades de r

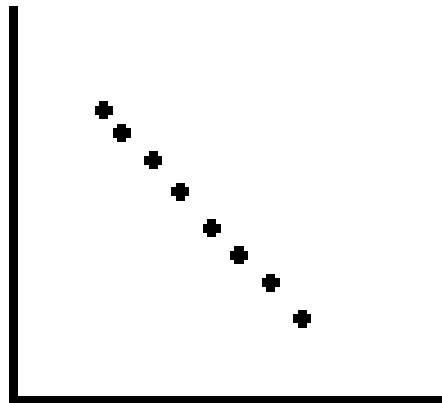
1. El valor de r no depende de cuál de las dos variables estudiadas es x y cual es y .
2. El valor de r es independiente de las unidades en las cuales x y y estén medidas.
3. Es valido $-1 \leq r \leq 1$
4. Se cumple que $r = 1$ si y sólo si todos los pares (x_i, y_i) quedan en una línea recta con pendiente positiva y $r = -1$ si y sólo si los pares (x_i, y_i) quedan en una línea recta con pendiente negativa.
5. El cuadrado del coeficiente de correlación muestral da el valor del coeficiente de determinación que resultaría de ajustar el **modelo de regresión lineal simple**, en símbolos $(r)^2 = R^2$.



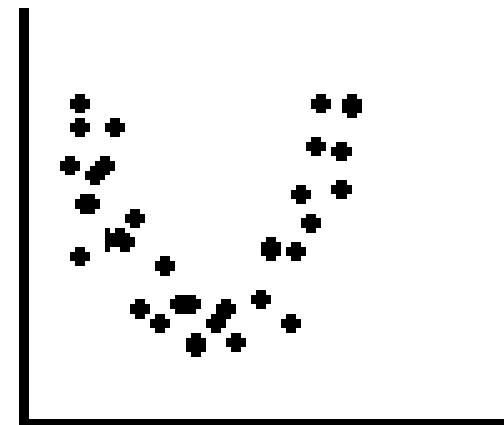
$$r = 1$$



$$r = 0,8$$



$$r = -1$$



$$r = 0$$

Las correlaciones no necesariamente implican causalidad

Posibles explicaciones para una aparente correlación:

- 1. La correlación podría ser una *coincidencia*.**
- 2. Ambas variables correlacionadas podrían ser influidas directamente por alguna *causa subyacente común*.**
- 3. Una de las variables correlacionadas en realidad podría ser una causa de la otra (observe que, incluso en este caso, podría ser sólo una de varias causas).**

Diferencias entre el coeficiente de determinación y el coeficiente de correlación

Característica	Coeficiente de correlación r	Coeficiente de determinación R^2
Mide	Fuerza y dirección de la relación	Proporción de variabilidad explicada
Tipo de relación	Lineal	Lineal (modelo de regresión)
Rango de valores	-1 a 1	0 a 1
Signo	Importa (+ o -)	No importa (es una proporción)
Interpretación	Asociación lineal	Calidad del ajuste del modelo
En regresión lineal simple	$R^2 = r^2$	Cuadrado del coeficiente de correlación