REGRESIÓN LINEAL

Docente: Johanna Trochez

Regresión lineal

Permite establecer asociaciones entre variables de interés, entre las cuáles la relación usual no es necesariamente de causa – efecto.

El objetivo es obtener estimaciones razonables de Y para distintos valores de X a partir de una muestra de n pares de valores (x1, y1), . . . ,(xn, yn).

El modelo más simple de regresión corresponde a:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

y_i variable respuesta o dependiente para la i-ésima observación

β_0 Intercepto

β_1 Pendiente

x_i variable predictora independiente para la i-ésima observación

ε_i error aleatorio para la i-ésima observación

$$\varepsilon_i \sim N(0, \sigma^2)$$

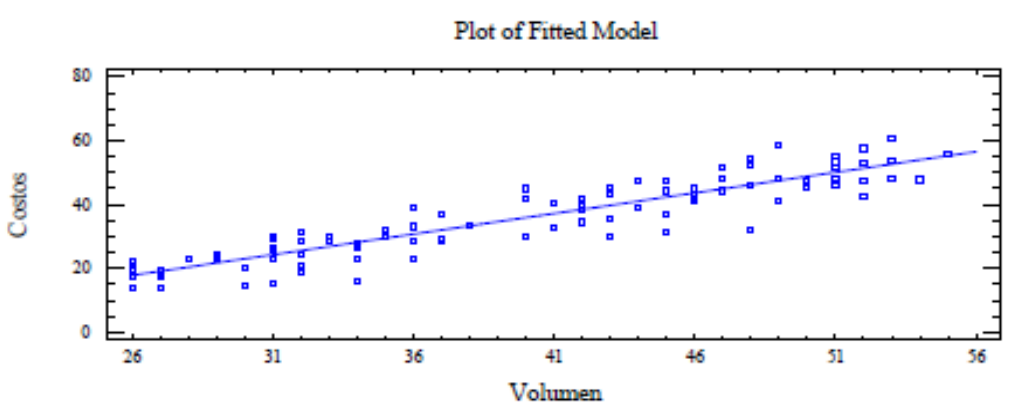
El objetivo es obtener estimaciones $\beta_0,\ \beta_1, \sigma$ para calcular la recta de regresión:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

que se ajuste lo mejor posible a los datos.

Ejemplo: Supongamos que la recta de regresión

Costo = 15,65 + 1,29 Volumen



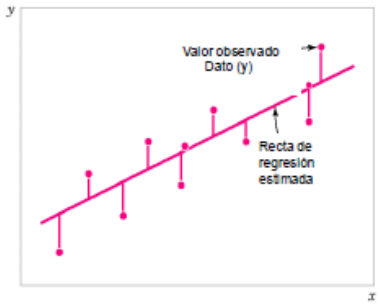
Se estima que una empresa que produce 25 mil unidades tendrá un costo:

costo = 15,65 + 1,29 × 25 = 47,9mil

Errores o residuales

La diferencia entre cada valor y_i de la variable respuesta y su estimación $\widehat{y_i}$ se llama residuo:

$$\varepsilon_i = y_i - \widehat{y_i}$$



Ejemplo (cont.): una empresa determinada que haya producido exactamente 25 mil unidades no va a tener un gasto de exactamente 47,9 mil euros. La diferencia entre el costo estimado y el real es el residuo. Si por ejemplo el costo real de la empresa es de 55 mil , el residuo es:

$$\varepsilon_i = 55 - 47.9 = 7.1$$

Objetivos de la regresión lineal simple

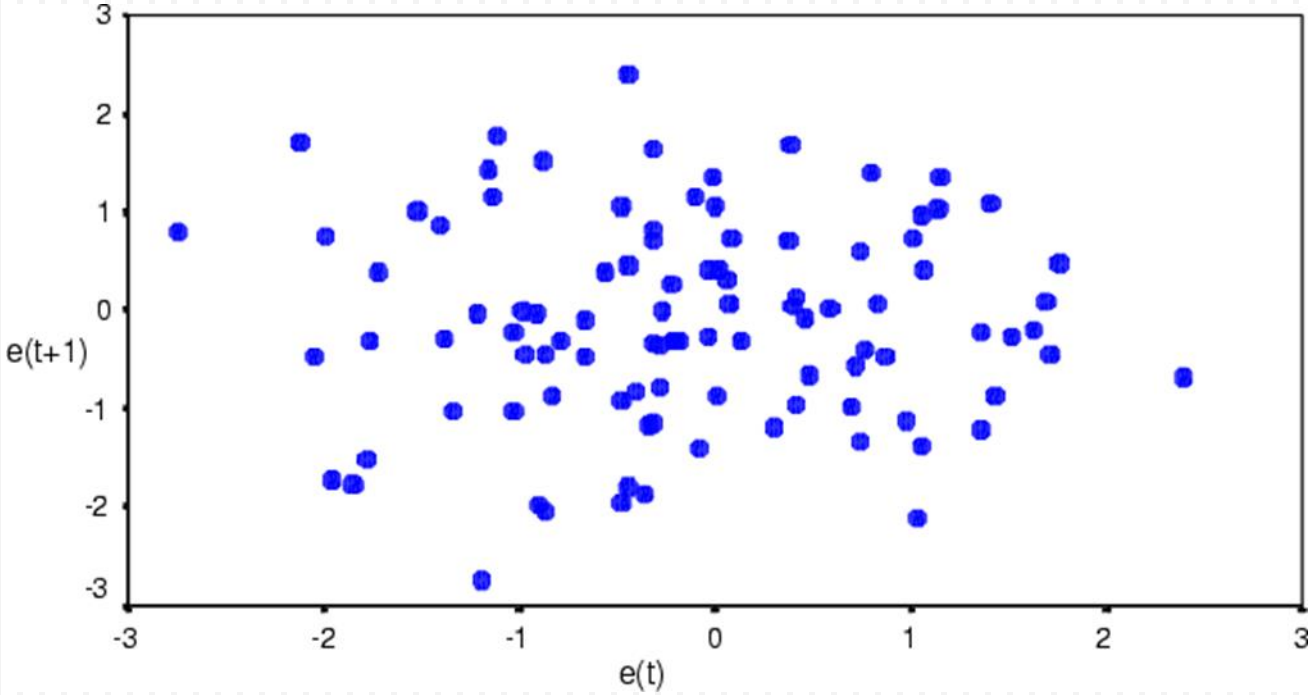
- Construir un modelo que describa cómo influye una variable X sobre otra variable Y
- Obtener estimaciones puntuales de los parámetros de dicho modelo
- Estimar el valor promedio de Y para un valor de X
- Predecir futuros de la variable respuesta Y

Algunos ejemplos

- Estudiar cómo influye la estatura del padre sobre la estatura del hijo.
- Estimar el precio de una vivienda en función de su área.
- Aproximar la calificación obtenida en una materia según el numero de horas de estudio semanal.
- Prever el tiempo de computación de un programa en función de la velocidad del procesador

Diagrama de dispersión

Diagrama matemático que utiliza las coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos.



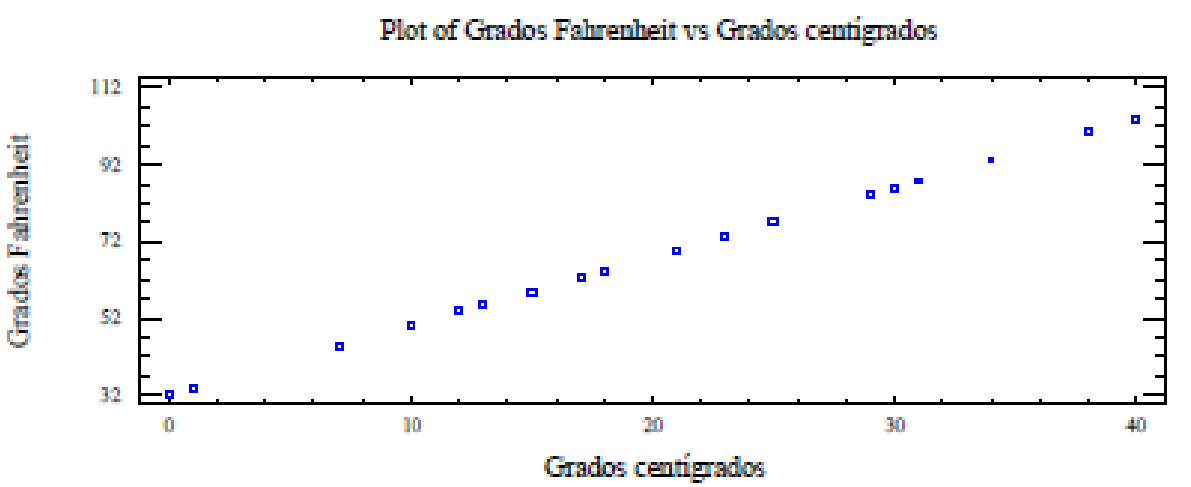
Tipos de relación

- **Determinista:** Conocido el valor de x, el valor de Y queda perfectamente establecido. Son del tipo:

$$y = f(x)$$

- Ejemplo: La relación existente entre la temperatura en grados centígrados (X) y grados Fahrenheit (Y) es:

$$y = 1{,}8x + 32$$



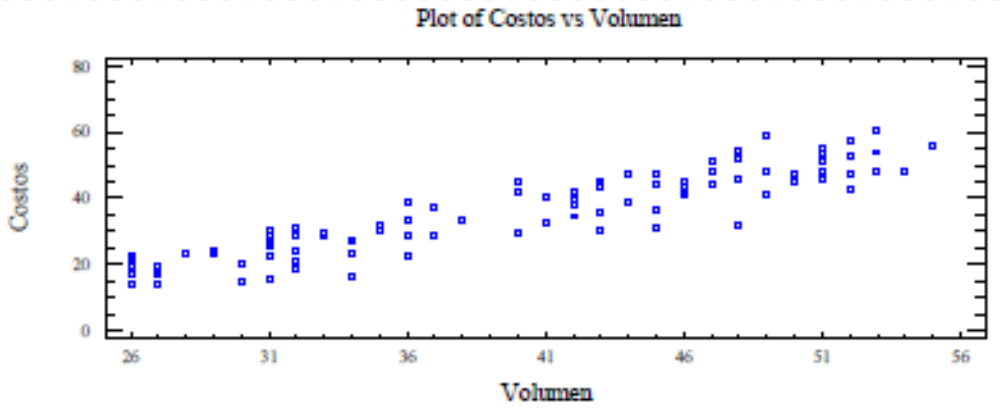
No determinista

Conocido el valor de X, el valor de Y no queda perfectamente establecido. Son del tipo:

$$y = f(x) + \varepsilon$$

donde ε es una perturbación desconocida (variable aleatoria).

Ejemplo: Se tiene una muestra del volumen de producción (X) y el costo total (Y) asociado a un producto en un grupo de empresas, es decir existe una relación pero no es exacta.



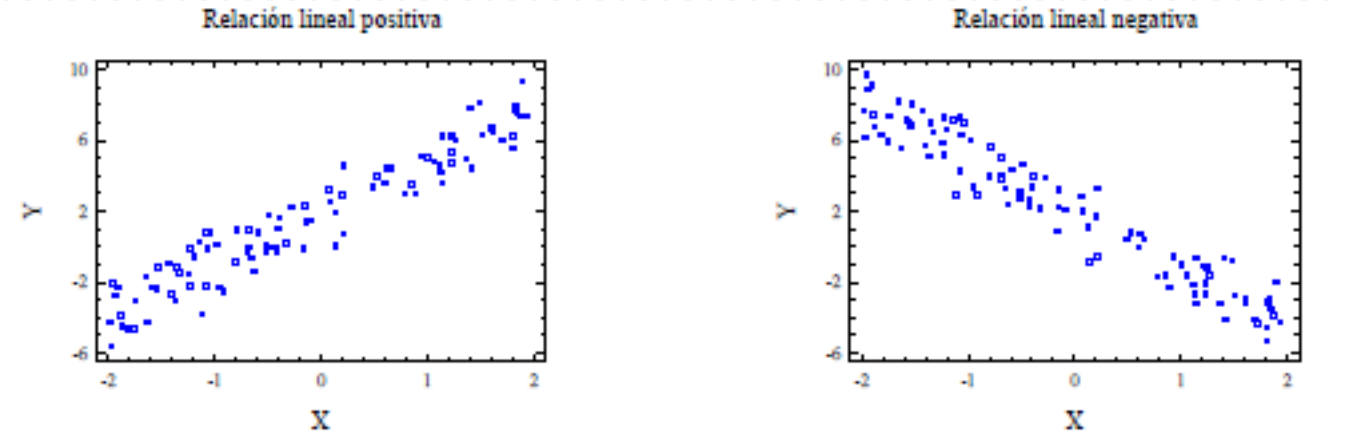
Relación lineal

Cuando la función $f(x)$es lineal,

$$y = \beta_0 + \beta_1 x_i + \varepsilon$$

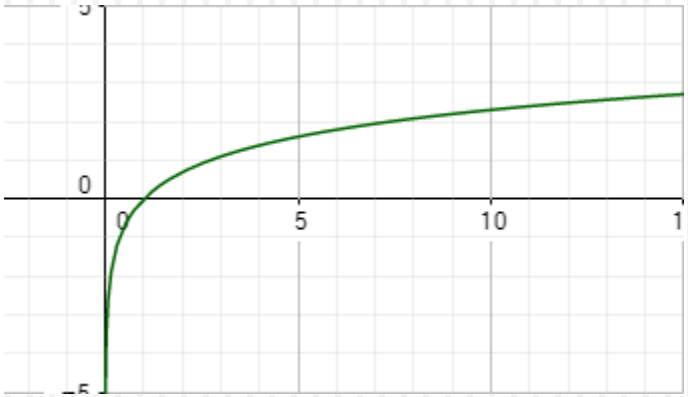
Si $\beta_1 > 0$ hay relación lineal positiva.

Si $\beta_1 < 0$ hay relación lineal negativa

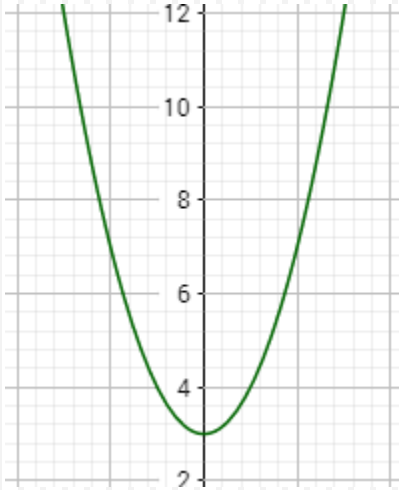


Relación no lineal

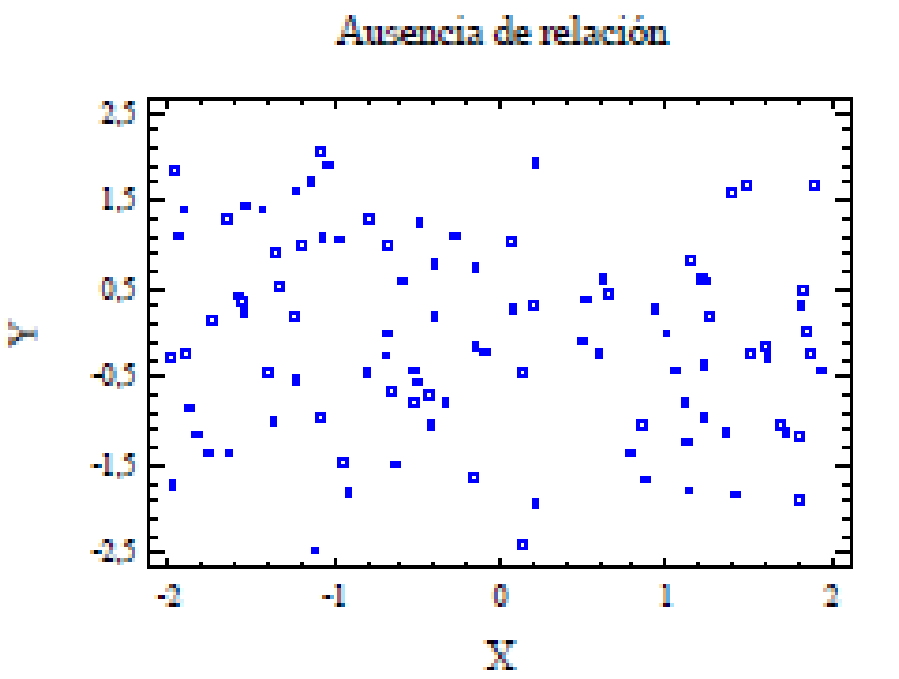
- Los datos no tienen un aspecto recto
- $f(x) = \log(x)$



$f(x) = x^2 + 3$



Sin relación $f(x)=0$



Medidas de dependencia lineal

La covarianza indica el grado de variación conjunta de dos variables aleatorias respecto a sus medias

$$cov\,(x,y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})\,(y_i - \bar{y})}{n-1}$$

- Si hay relación lineal positiva, la covarianza será positiva y grande.
- Si hay relación lineal negativa, la covarianza será negativa y grande en valor absoluto.
- Si no hay relación entre las variables la covarianza será próxima a cero.
- La covarianza depende de las unidades de medida de las variables.

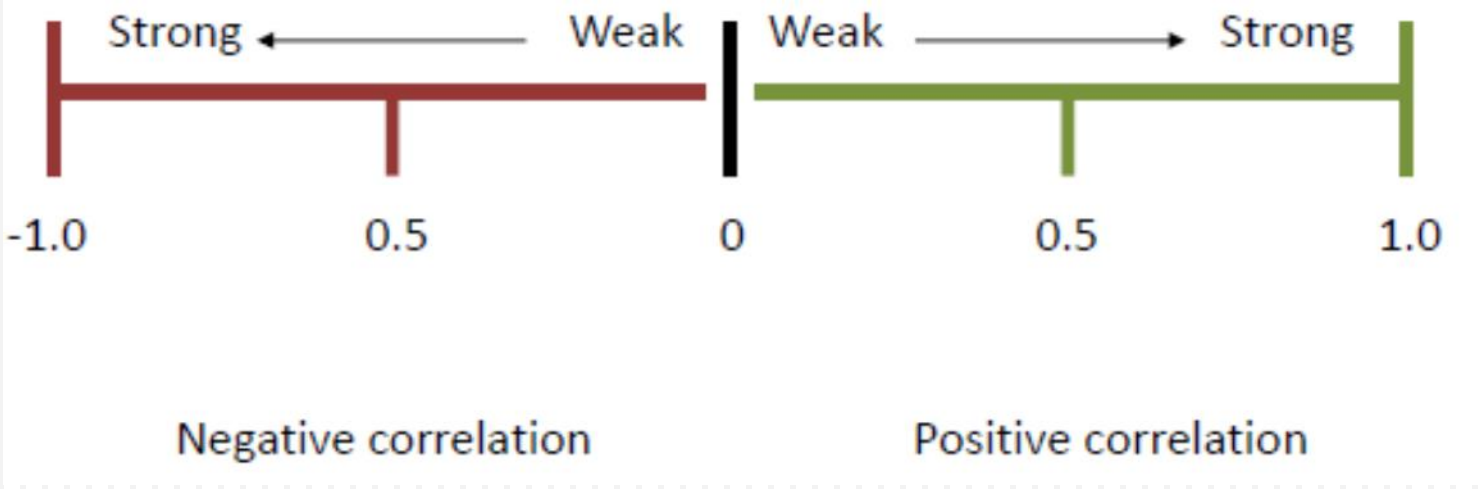
Coeficiente de correlación

Indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables cuantitativas estadísticas.

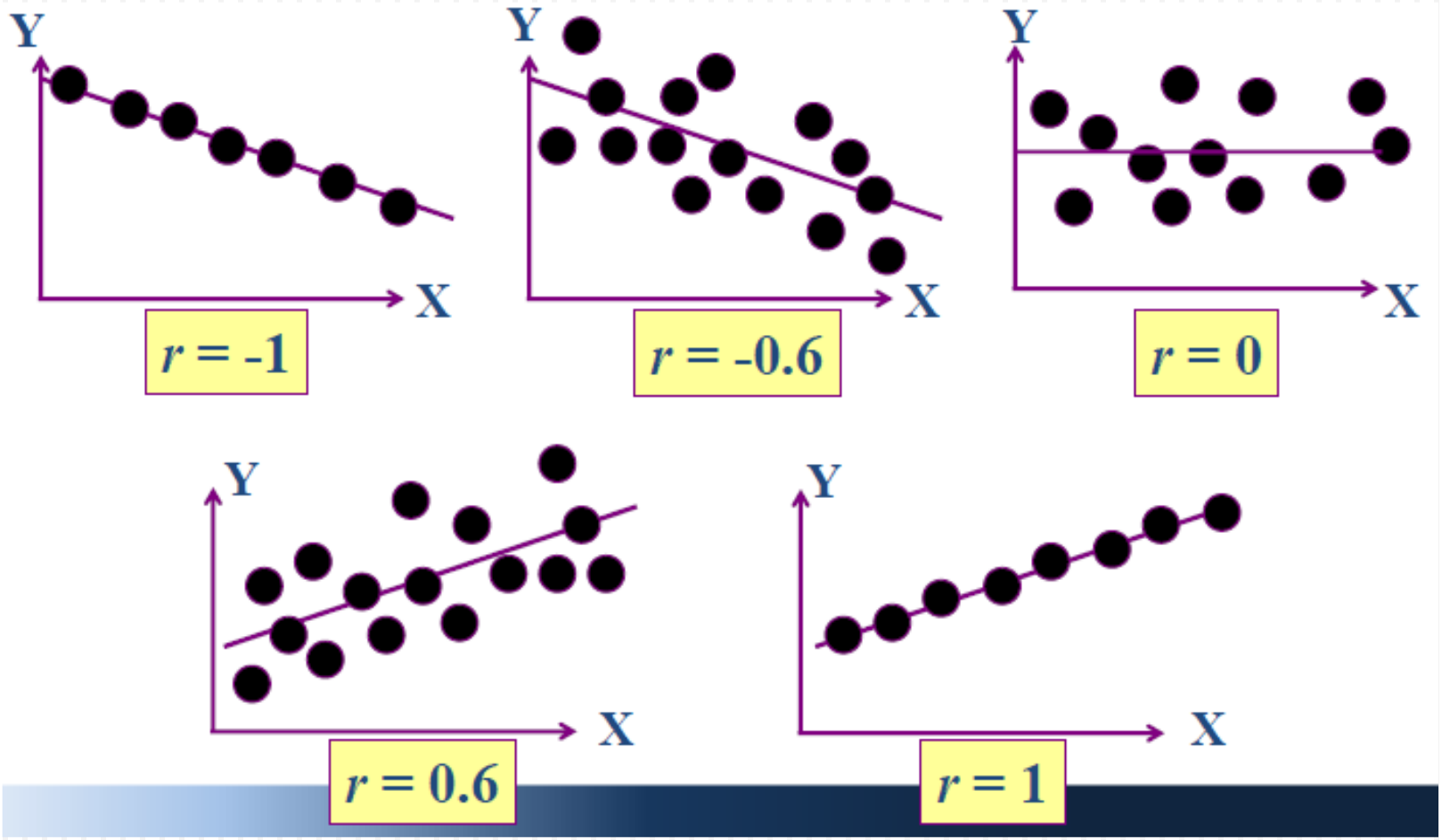
$$r = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}}$$

Características del coeficiente de correlación

- Rango entre -1 y 1
- Valores cercanos a -1 la relación es fuertemente negativa.
- Valores cercanos a 1 la relación es fuertemente positiva.
- Valores cercanos a 0 la relación es débil, es decir no hay una relación lineal



Varios coeficientes de correlación



Hipótesis del modelo de regresión lineal simple

- **Linealidad:** La relación existente entre X e Y es lineal,
$$f(x) = \beta_0 + \beta_1 x_i$$

- **Homogeneidad:** El valor promedio del error es cero.
$$E[\varepsilon_i] = 0$$

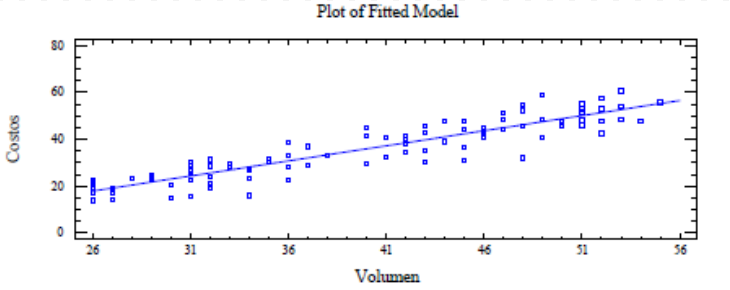
- **Homocedasticidad:** La varianza de los errores es constante
$$Var(\varepsilon_i) = \sigma^2$$

- **Independencia:** Las observaciones son independientes,
$$E[x, y] = 0$$

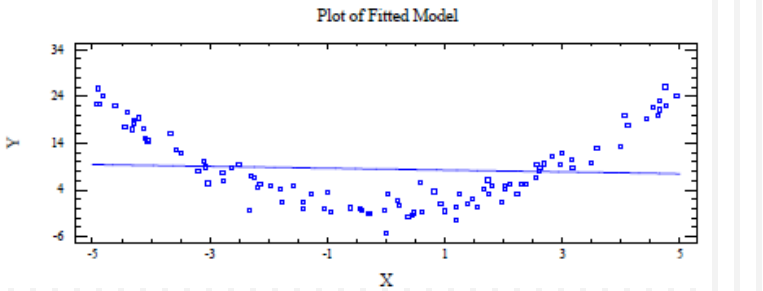
- **Normalidad:** Los errores siguen una distribución normal,
$$\varepsilon_i \sim N(0, \sigma^2)$$

Linealidad

- Los datos deben ser rectos.



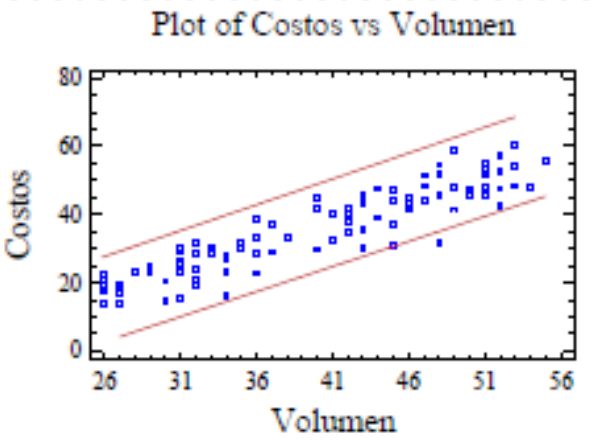
- Si no, la recta de regresión no representa la estructura de los datos.

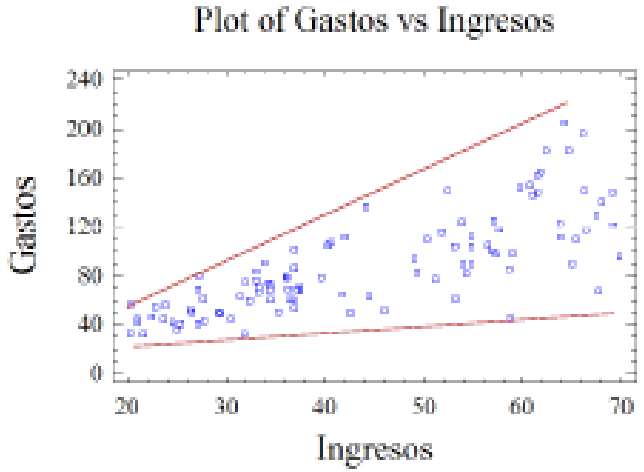


Homocedasticidad

La dispersión de los datos debe ser constante para que los datos sean **homocedásticos.**

Si no se cumple, los datos son **heterocedásticos.**

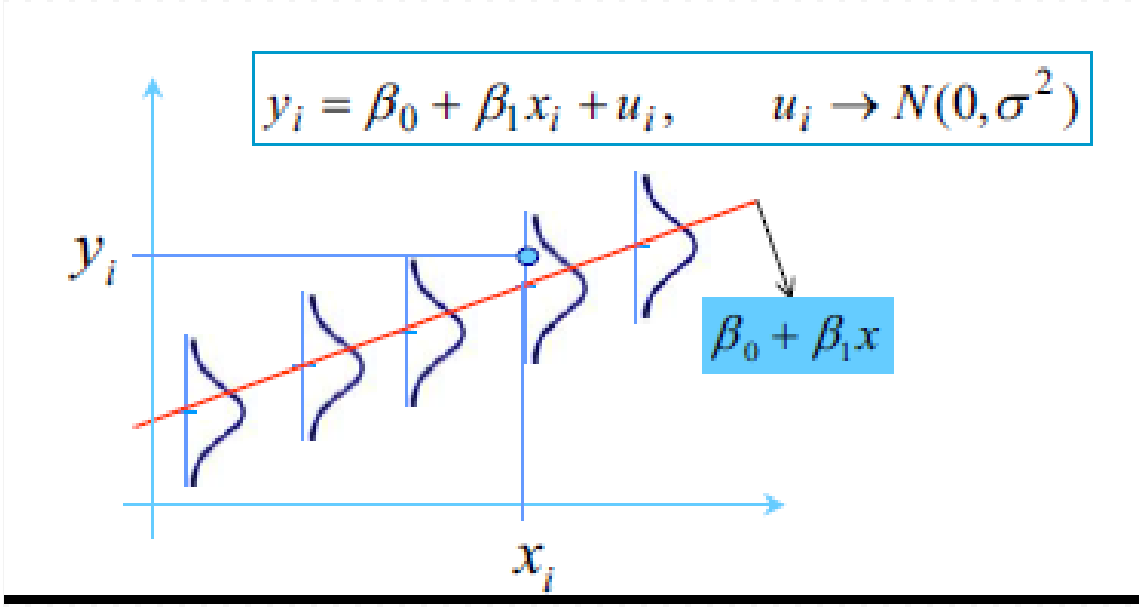




Independencia

- Los datos deben ser independientes.
- Una observación no debe dar información sobre las demás.
- Las series temporales no cumplen la hipótesis de independencia.

Normalidad



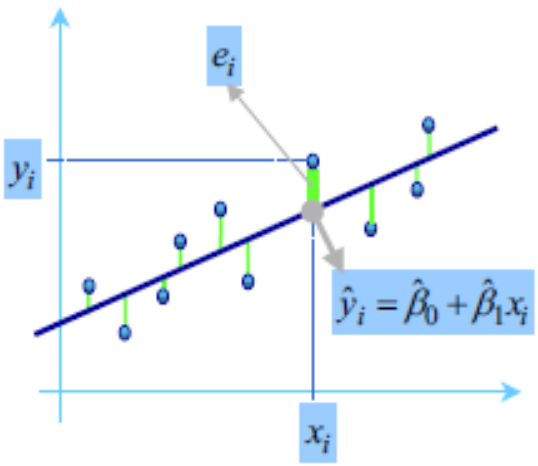
Estimadores de mínimos cuadrados

Gauss propuso en 1809 el método de mínimos cuadrados para obtener los valores $\widehat{\beta_0}$ y $\widehat{\beta_1}$ que mejor se ajustan a los datos:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, minimizar la suma de los residuos al cuadrado:

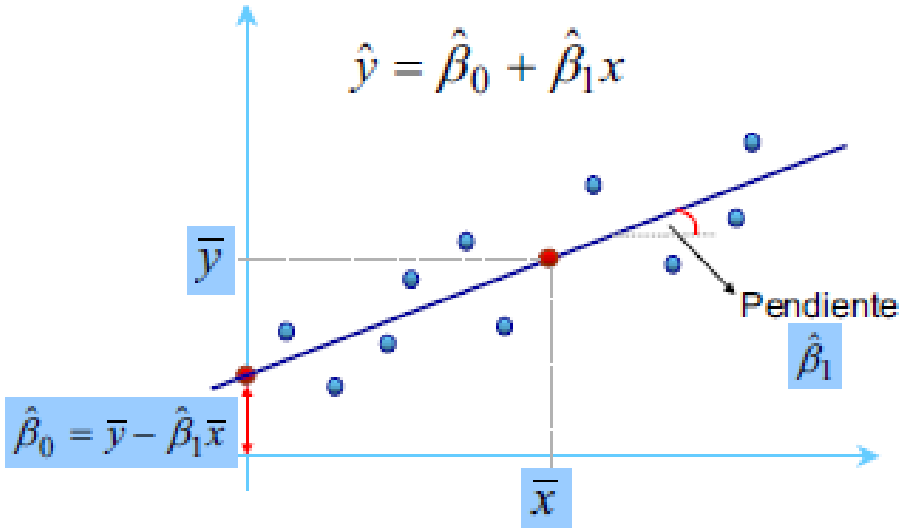
$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$



El resultado que se obtiene es:

$$\hat{\beta}_1 = \frac{cov(x,y)}{s_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$



Estimación de la varianza

Para estimar la varianza de los errores, σ^2, podemos utilizar,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n}$$

que es el estimador máximo verosímil de σ^2, pero es un estimador sesgado.

Un estimador insesgado de σ^2 es la varianza residual,

$$\boxed{s_R^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}}$$

EJEMPLO

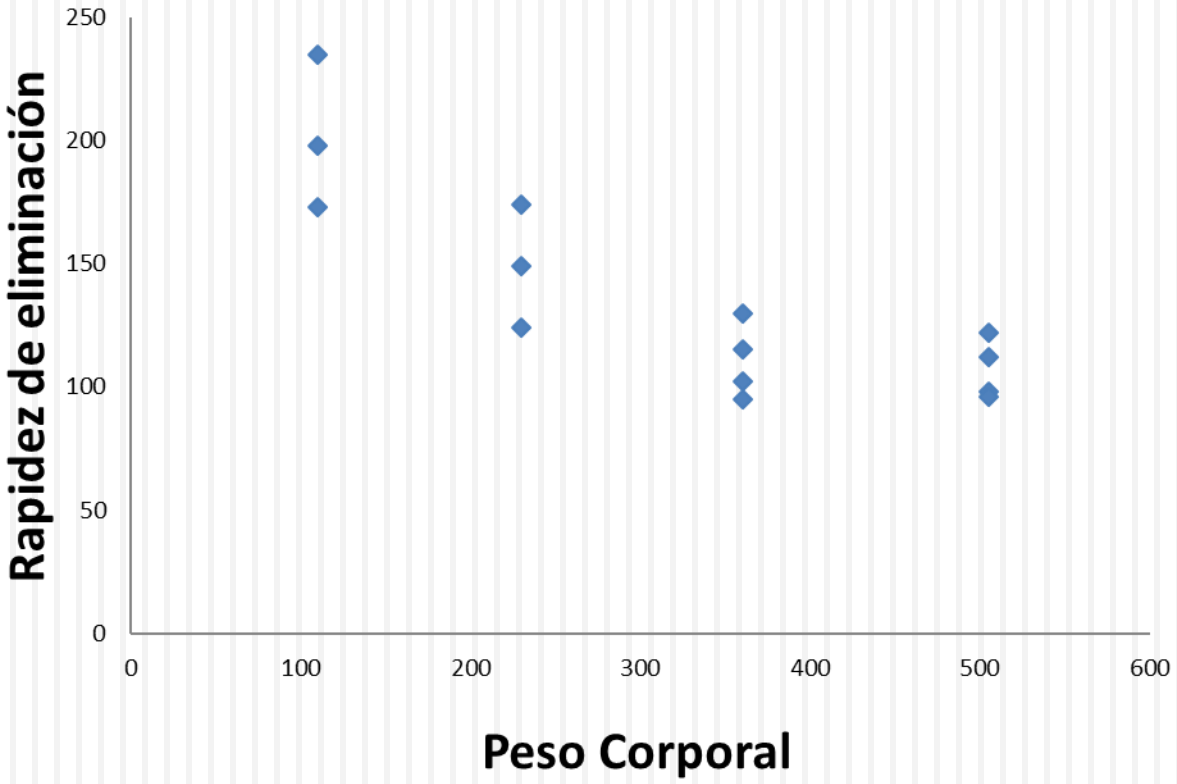
Los siguientes datos se recolectaron con el fin de determinar la relación existente entre el peso corporal del ganado vacuno (X), y la rapidez de eliminación metabólica/peso corporal (Y).

Los datos que aparecen a continuación son el resultado de varias realizaciones del experimento, en distintos niveles del peso.

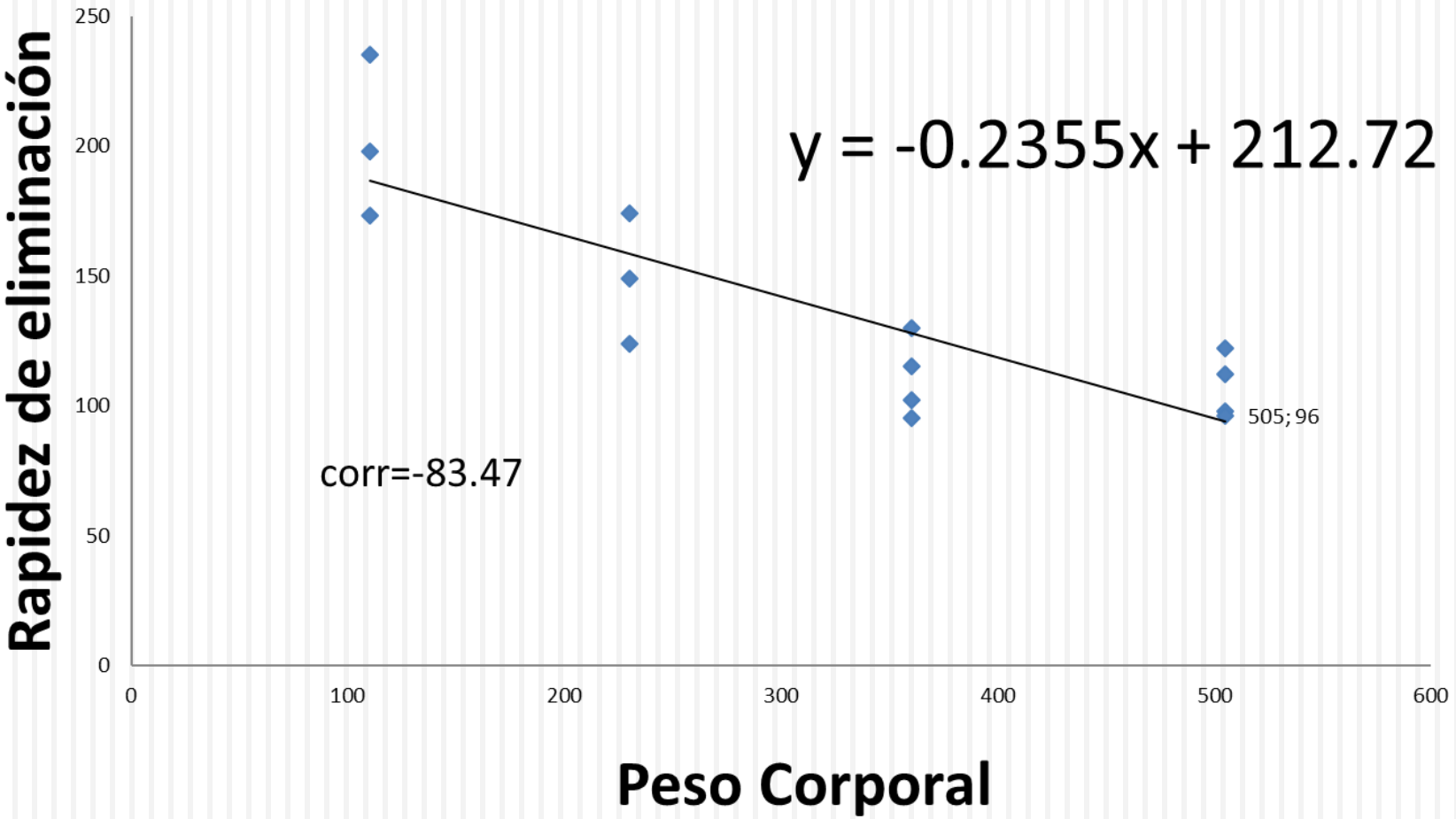
Peso corporal (x)	Rapidez de eliminación (y)
110	235
110	198
110	173
230	174
230	149
230	124
360	115
360	130
360	102
360	95
505	122
505	112
505	98
505	96

Diagrama de dispersión

¿Qué correlación hay entre estas variables?
¿Cuál es el modelo de regresión lineal?



	Peso	Eliminación	$X_i - \bar{X}$	$(X_i - \bar{X})^2$	$y_i - \bar{y}$	$(X_i - \bar{X})(y_i - \bar{y})$
	110	235				
	110	198				
	110	173				
	230	174				
	230	149				
	230	124				
	360	115	40	1600	-22,36	-894,29
	360	130	40	1600	-7,36	-294,29
	360	102	40	1600	-35,36	-1414,29
	360	95	40	1600	-42,36	-1694,29
	505	122	185	34225	-15,36	-2841,07
	505	112	185	34225	-25,36	-4691,07
	505	98	185	34225	-39,36	-7281,07
	505	96	185	34225	-41,36	-7651,07
$\bar{X}$						
TOTAL				299900		-70630


Rapidez de eliminación
250
200
150
100
50
0
0
100
200
300
400
500
600
y = -0.2355x + 212.72
505; 96
corr=-83.47
Peso Corporal

Significado de $\widehat{\beta_0}$ y $\widehat{\beta_1}$

- $\widehat{\beta_0}$ el intercepto es en el valor de 212.72, que corresponde a la eliminación metabólica

- $\widehat{\beta_1}$ es el valor de la pendiente, es decir que por cada kilo que aumenta una res, la eliminación metabólica se reduce en 0.2355 unidades

Correlación

- r=0.8347

- Obtenga una estimación puntual de la eliminación del peso para una res que pesa 550 kilos.

Cómo realizar un modelo de regresión lineal en la calculadora

- https://www.youtube.com/watch?v=4_WO31Dapv0&t=7s
- https://www.youtube.com/watch?v=4cQe6J7RzAI&t=41s

¿Cómo realizar un modelo de regresión lineal en Excel?

- Una sustancia empleada en investigación médica y biológica es transportada por carga aérea en cajas de cartón conteniendo 1000 ampollas de la sustancia. En la siguiente tabla se presentan los datos obtenidos para 10 embarques y corresponden a número de veces que las cajas son transferidas de un avión a otro en la ruta de embarque y número de ampollas que fueron halladas quebradas a la llegada.

Transferencias	Ampollas quebradas
1	16
0	9
2	17
0	12
3	22
1	13
0	8
1	15
2	19
0	11

¿Cómo realizar un modelo de regresión lineal en R?