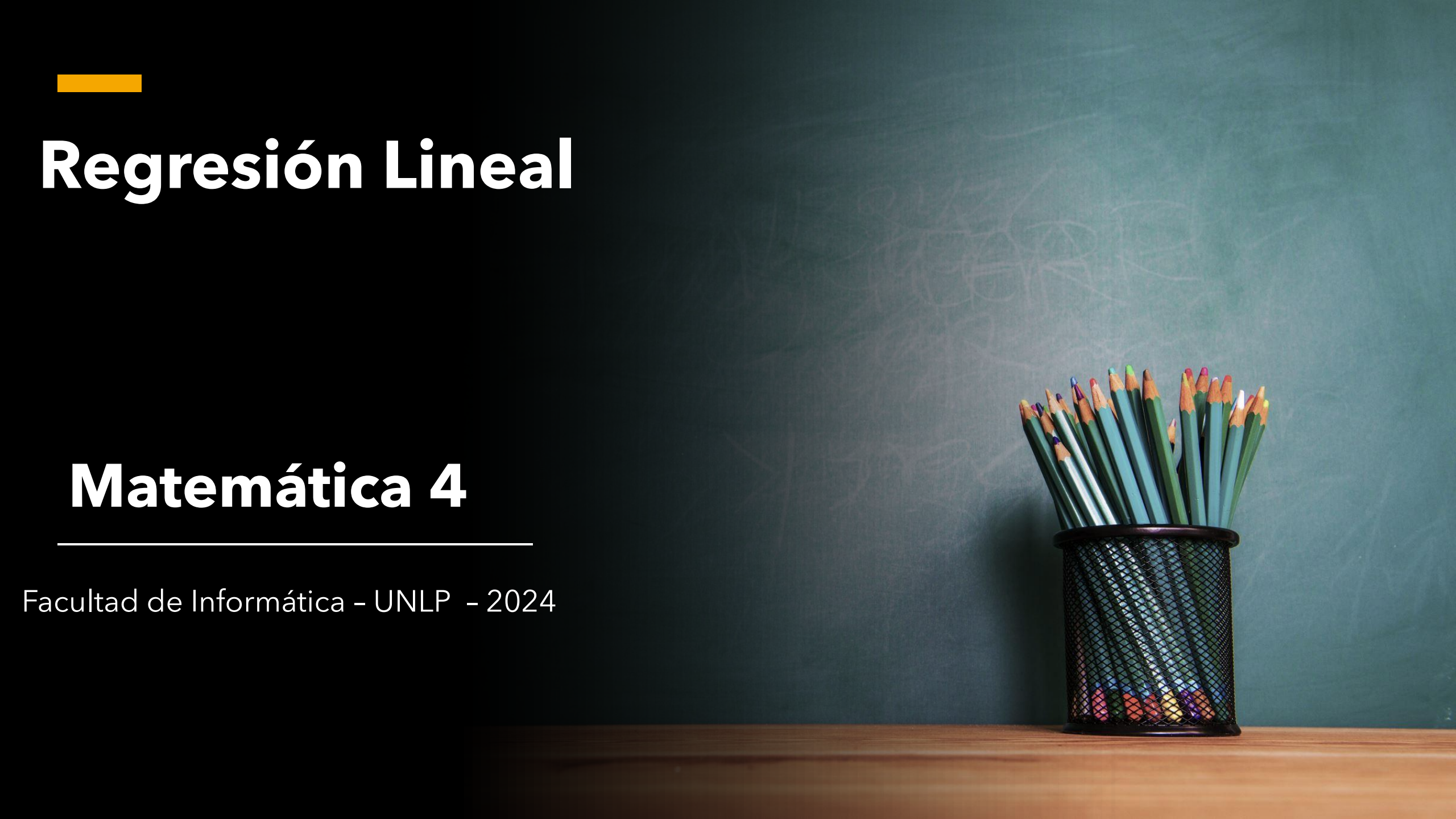
# Regresión Lineal

## Matemática 4

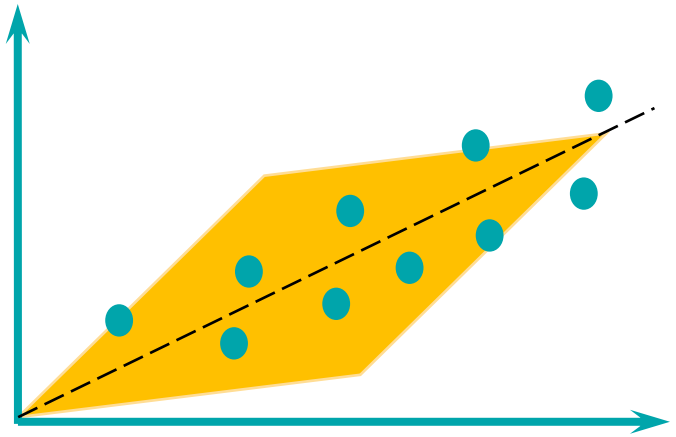
Facultad de Informática – UNLP – 2024



# Regresión Lineal

**Definición:** La regresión lineal es una técnica de modelado estadístico que se emplea para describir una variable de respuesta continua como una función de una o varias variables predictoras. Puede ayudar a comprender y predecir el comportamiento de sistemas complejos o a analizar datos experimentales, Informáticos, financieros y biológicos.

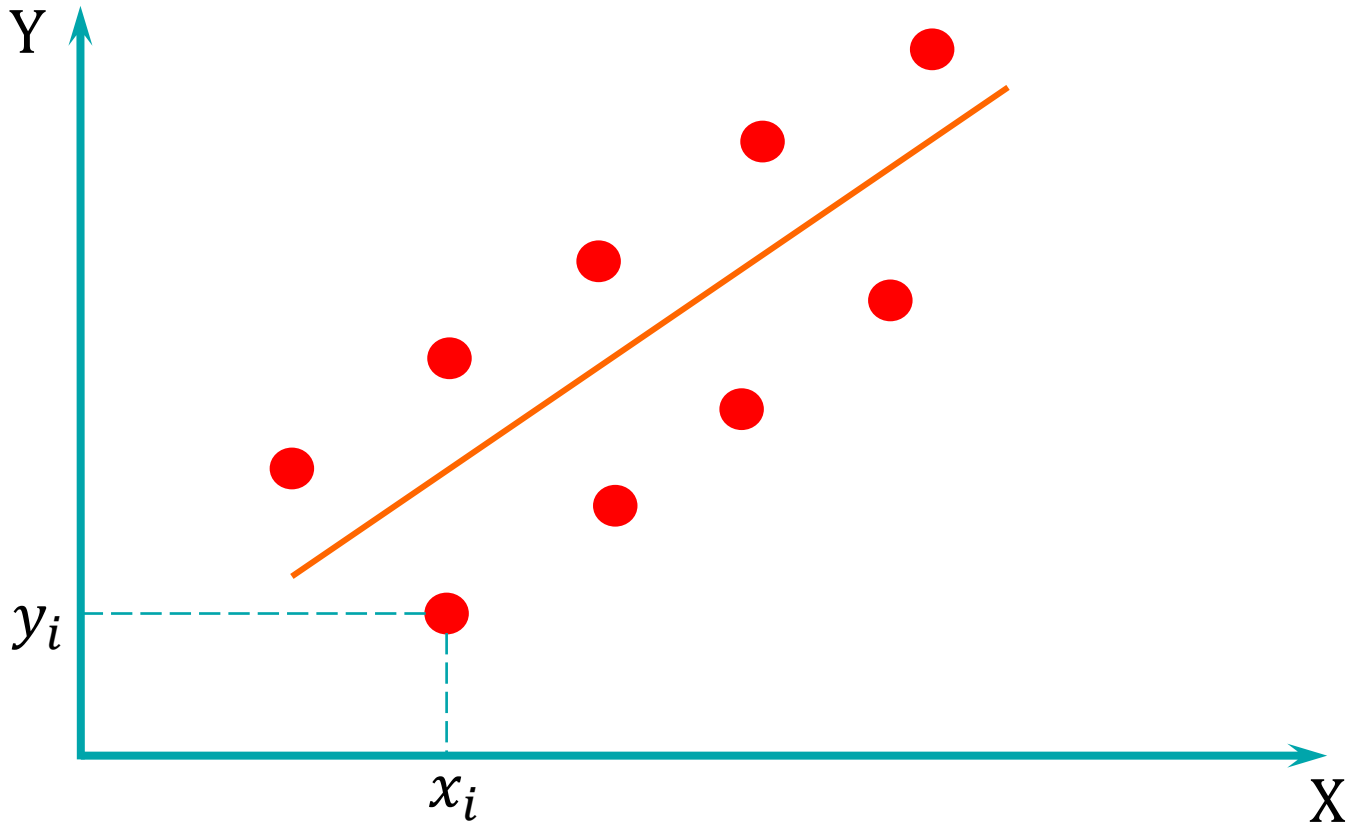
- **Regresión Lineal Simple**



- **Regresión Lineal Múltiple**

# Regresión Lineal Simple

## Gráfico de dispersión



# Regresión Lineal Simple

**Modelo de Regresión Lineal Simple:**

$$Y = \beta_1 X + \beta_0 + \varepsilon$$

Lineal en los parámetros $\beta_1$, $\beta_0$ y simple porque involucra solo una variable predictora.

$Y$: Variable Respuesta o Dependiente

$X$: Variable Predictora o Independiente

$\varepsilon$: Variable Error aleatorio

$\beta_1$: Parámetro Pendiente

$\beta_0$: Parámetro Ordenada

**Suposiciones del Modelo:**

- $\varepsilon$ es Independiente
- $\varepsilon \sim N(0, \sigma^2)$

# Regresión Lineal Simple

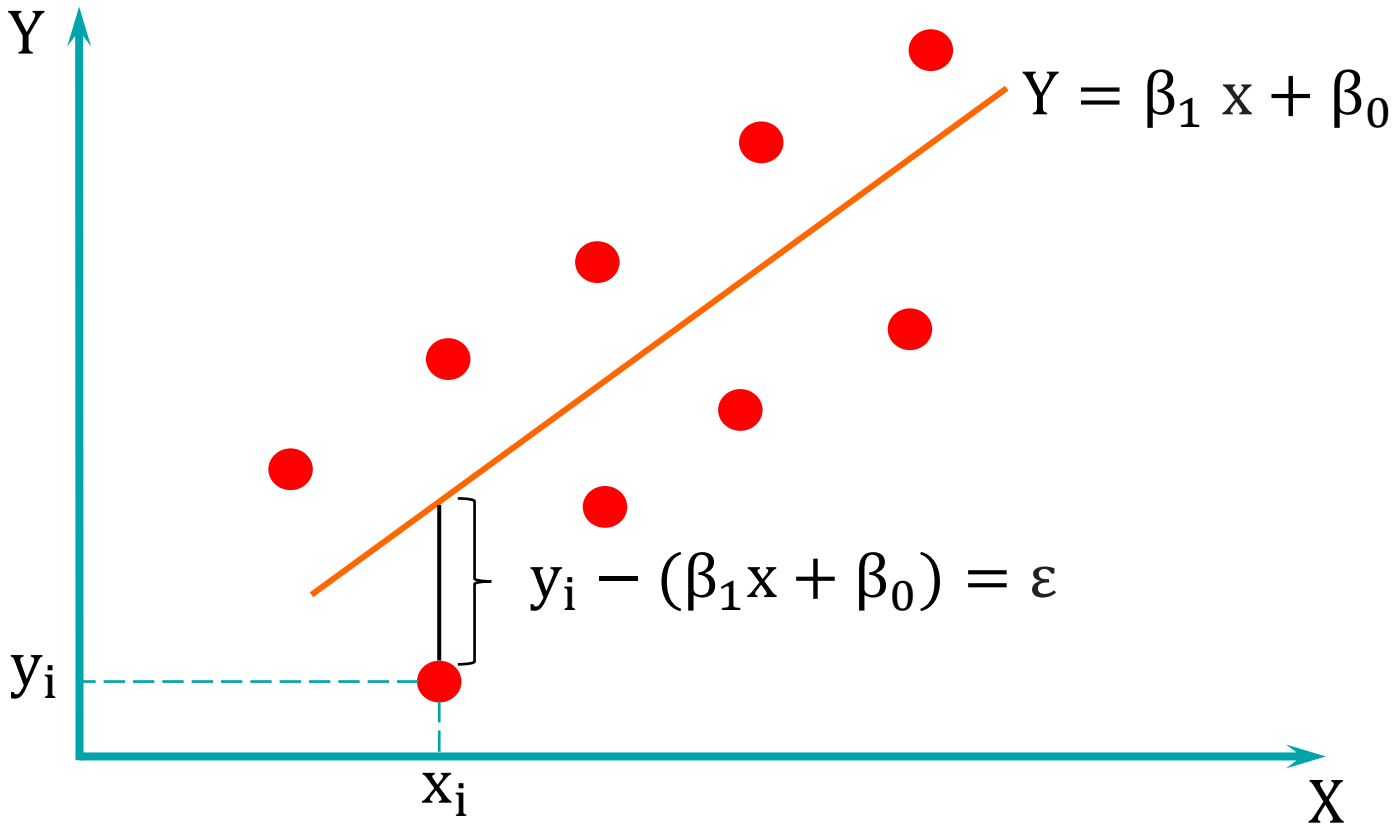
En general podemos decir que al fijar el valor de $\boldsymbol{x}$ observamos el valor de la variable $Y$. Si bien $\boldsymbol{x}$ es fijo, el valor de $\boldsymbol{Y}$ está afectado por el ***error aleatorio*** $\boldsymbol{\varepsilon}$ . Por lo tanto, $\boldsymbol{\varepsilon}$ ***determina las propiedades de*** $Y$.

$$\mathrm{E(Y)} = \mathrm{E}(\beta_1 \mathrm{x} + \beta_0 + \varepsilon) = \mathrm{E}(\beta_1 \mathrm{x}) + \mathrm{E}(\beta_0) + \mathrm{E}(\varepsilon) = \beta_1 \mathrm{x} + \beta_0$$

$$\mathrm{V(Y)} = \mathrm{V}(\beta_1 \mathrm{x} + \beta_0 + \varepsilon) = \mathrm{V}(\beta_1 \mathrm{x}) + \mathrm{V}(\beta_0) + \mathrm{V}(\varepsilon) = \mathrm{V}(\varepsilon) = \sigma^2$$

En consecuencia, el modelo de regresión verdadero $\mathrm{E(Y)} = \beta_1 \mathrm{x} + \beta_0$ es una recta de valores promedio.

# Regresión Lineal Simple




**Objetivo:** Encontrar la recta que mejor se ajuste a los datos observados.

# Regresión Lineal Simple

**Método de Mínimos Cuadrados:** La recta que mejor se ajuste a los datos observados será aquella cuyos valores de $\beta_1$ y $\beta_0$ minimicen la siguiente expresión:

$$f(\beta_1, \beta_0) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

A dichos valores lo denotaremos como $\hat{\beta}_1$, $\hat{\beta}_0$ y los llamaremos estimadores de mínimos cuadrados de la pendiente y la ordenada.

# Regresión Lineal Simple

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_1 x_i - \beta_0)^2$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} 2 \cdot (y_i - \beta_1 x_i - \beta_0) \cdot (-1) = \sum_{i=1}^{n} 2 \cdot (-y_i + \beta_1 x_i + \beta_0) = 0$$

$$\frac{\partial f(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} 2 \cdot (y_i - \beta_1 x_i - \beta_0) \cdot (-x_i) = \sum_{i=1}^{n} 2 \cdot \left(-y_i x_i + \beta_1 x_i^2 + \beta_0 x_i\right) = 0$$

# Regresión Lineal Simple

$$\sum_{i=1}^{n} -y_i + \sum_{i=1}^{n} \beta_1 x_i + \sum_{i=1}^{n} \beta_0 = -\sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i + \beta_0 n = 0$$

$$\sum_{i=1}^{n} -x_i y_i + \sum_{i=1}^{n} \beta_1 x_i^2 + \sum_{i=1}^{n} \beta_0 x_i = -\sum_{i=1}^{n} x_i y_i + \beta_1 \sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i = 0$$

# Regresión Lineal Simple

Reordenando:

$$\beta_1 \sum_{i=1}^{n} x_i + n\beta_0 = \sum_{i=1}^{n} y_i$$

$$\beta_1 \sum_{i=1}^{n} x_i^2 + \beta_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

**Ecuaciones Normales**

# Regresión Lineal Simple

Resolvamos el sistema de ecuaciones:

$$\beta_1 \sum_{i=1}^{n} x_i + n\beta_0 = \sum_{i=1}^{n} y_i \rightarrow \beta_0 = \frac{\sum y_i}{n} - \beta_1 \frac{\sum x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

Reemplazando en la segunda ecuación:

$$\beta_1 \sum_{i=1}^{n} x_i^2 + (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

# Regresión Lineal Simple

$$\beta_1 \sum_{i=1}^{n} x_i^2 + \bar{y} \sum_{i=1}^{n} x_i - \beta_1 \bar{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$\beta_1 \sum_{i=1}^{n} x_i^2 - \beta_1 \bar{x} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i$$

$$\beta_1 \left( \sum_{i=1}^{n} x_i^2 - \bar{x} \sum_{i=1}^{n} x_i \right) = \sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i$$

$$\beta_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\left( \sum x_i^2 - \bar{x} \sum x_i \right)} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

# Regresión Lineal Simple

**Los estimadores de mínimos cuadrados la pendiente y la ordenada seran:**

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

**Y la recta de regresión estimada es:**

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

# Regresión Lineal Simple

Otra forma de expresar el estimador de la pendiente es:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})\,(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_1 = \frac{\sum y_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

Donde:

$$S_{xy} = \sum(x_i - \bar{x})\,(y_i - \bar{y}) = \sum y_i(x_i - \bar{x}) = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n} \rightarrow \text{desviacion } (x_i) \text{ de su media}$$

# Regresión Lineal Simple

Demostración $S_{xy}$ :

$$\sum (x_i - \bar{x})(y_i - \bar{y})$$

$$\sum (x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{y}\bar{x})$$

$$\sum x_i y_i - \sum x_i \bar{y} - \sum \bar{x} y_i + \sum \bar{y}\bar{x}$$

$$\sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{y}\bar{x}$$

$$\sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{y}\bar{x}$$

$$\sum x_i y_i - n\bar{y} \frac{\sum x_i}{n} - n\bar{x} \frac{\sum y_i}{n} + n\bar{y}\bar{x}$$

$$\sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y}$$

$$\sum x_i y_i - n\bar{x}\bar{y}$$

$$\sum x_i y_i - n \frac{\sum x_i}{n} \frac{\sum y_i}{n}$$

$$\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

# Regresión Lineal Simple

Demostración $S_{xx}$ :

$$\sum (x_i - \bar{x})^2$$

$$\sum \left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right)$$

$$\sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$\sum x_i^2 - 2n\bar{x}\frac{\sum x_i}{n} + n\bar{x}^2$$

$$\sum x_i^2 - 2n\bar{x}\frac{\sum x_i}{n} + n\bar{x}^2$$

$$\sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$\sum x_i^2 - n\bar{x}^2$$

$$\sum x_i^2 - n\frac{(\sum x_i)^2}{n^2}$$

$$\sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

# Regresión Lineal Simple

**Nota 1:** La recta de regresión estimada:

$$\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$$

No debe ser utilizada para pronosticar valores de la variable independiente que no se encuentren dentro del rango de la misma. Porque se corre el **peligro de extrapolación**.

# Regresión Lineal Simple

**Nota 2:** Si la variable independiente pasa a tener la siguiente expresión:

$$ax + c$$

con a y c constantes cualesquiera. Entonces la nueva recta de regresión estimada es:

$$\hat{y}' = \hat{\beta}_1(ax + c) + \hat{\beta}_0$$
$$\hat{y}' = \hat{\beta}_1 ax + \hat{\beta}_1 c + \hat{\beta}_0$$
$$\hat{y}' = (\hat{\beta}_1 a)x + (\hat{\beta}_1 c + \hat{\beta}_0)$$

Denotando $\hat{\beta}_1' = \hat{\beta}_1 a$ y $\hat{\beta}_0' = \hat{\beta}_1 c + \hat{\beta}_0$ nos queda:

$$\hat{y}' = \hat{\beta}_1' x + \hat{\beta}_0'$$

# Regresión Lineal Simple

**Propiedades de los estimadores de mínimos cuadrados:**

Como $\hat{\beta}_1$ y $\hat{\beta}_0$ son estimadores de $\beta_1$ y $\beta_0$ respectivamente, son variables aleatorias, por lo tanto, podemos calcular su esperanza y varianza. Como estamos asumiendo que $x$ no es v.a. entonces $\hat{\beta}_1$ y $\hat{\beta}_0$ son funciones de la variable aleatoria $Y$ .

$$E(\hat{\beta}_1) = \beta_1 \quad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$E(\hat{\beta}_0) = \beta_0 \quad V(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}}\right)$$

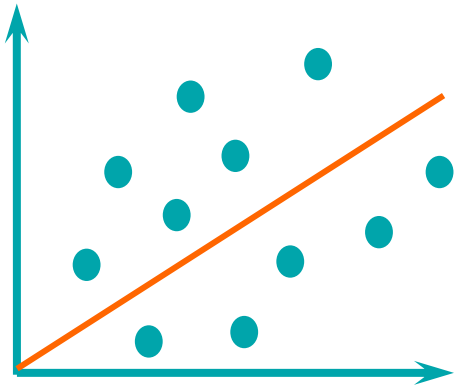
$\hat{\beta}_1$ y $\hat{\beta}_0$ son estimadores insesgado de $\beta_1$ y $\beta_0$

# Regresión Lineal Simple

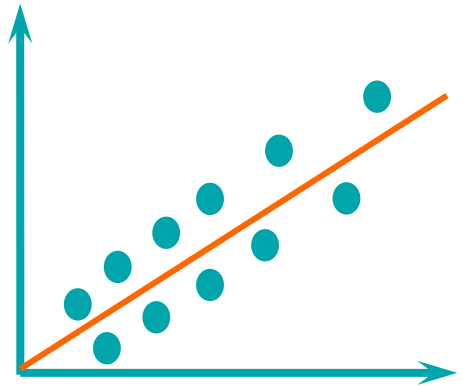
**Estimación de la varianza $\sigma^2$**

El parámetro $\sigma^2$ determina la cantidad de variabilidad en el modelo de regresión.

- Un valor grande de $\sigma^2$ conducirá a $(x_i, y_i)$ observados que están bastante dispersos entorno a la recta de regresión verdadera.
- Mientras que $\sigma^2$ sea pequeña los puntos observados tenderán a quedar cerca de la recta de regresión verdadera.



$\sigma^2$



$\sigma^2$

# Regresión Lineal Simple

Para estimar la varianza vamos a usar las desviaciones verticales entre los valores observados y valores estimados. Las cuales nombraremos **Residuos**:

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_1 x_i - \hat{\beta}_0$$

Y a la suma al cuadrado de la misma llamaremos **Suma de los Residuos al Cuadrado** y lo denotamos como $SS_R$:

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

# Regresión Lineal Simple

Entonces el estimador insesgado de $\sigma^2$ es:

$$\hat{\sigma}^2 = \frac{SS_R}{n-2}$$

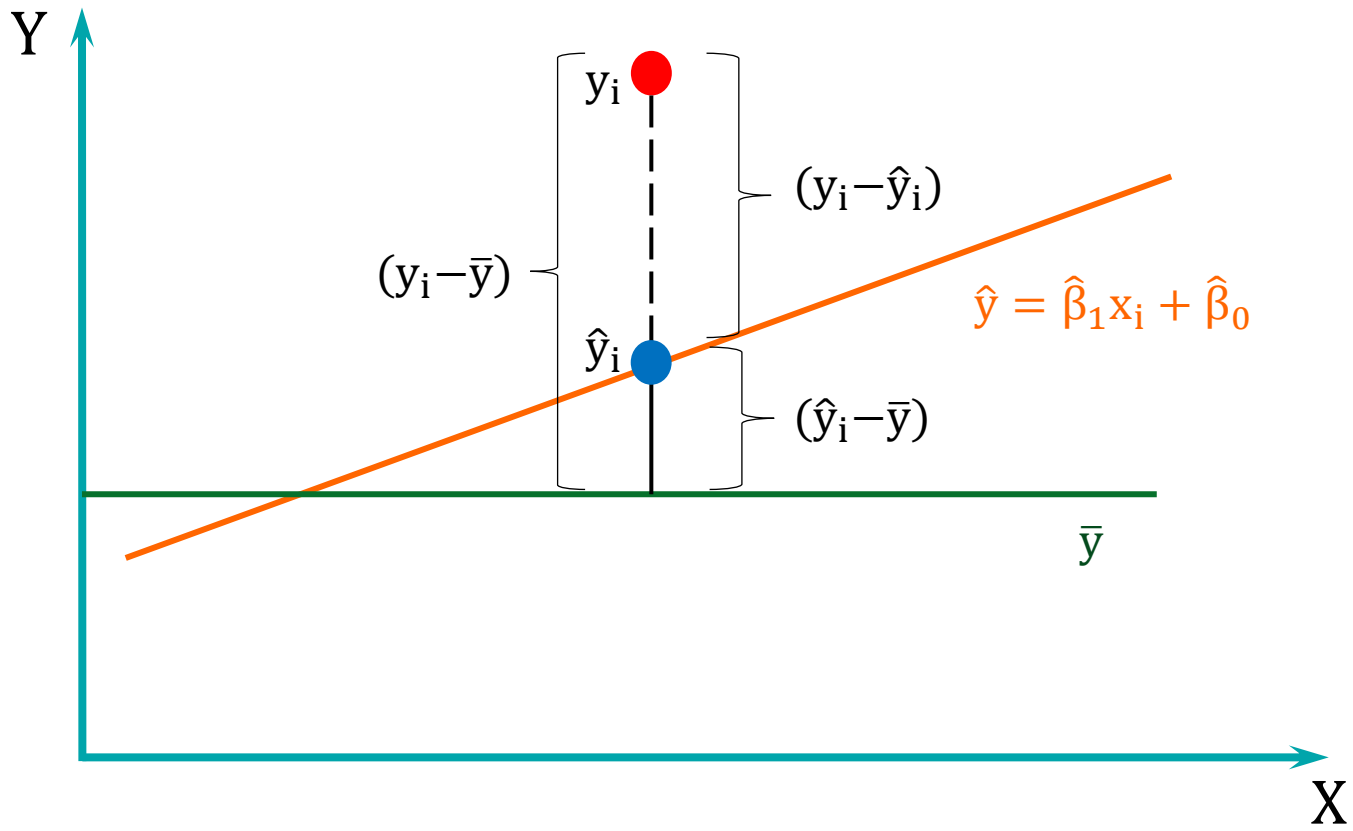
Otra forma de expresar la suma de los residuos al cuadrado es:

$$SS_R = S_{yy} - \frac{(S_{xy})^2}{S_{xx}} = S_{yy} - \hat{\beta}_1 S_{xy}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \rightarrow \text{desviacion } (y_i) \text{ de su media}$$

# Regresión Lineal Simple

## Coeficiente de determinación




$(y_i-\hat{y}_i)$: Desviación no explicada del valor observado, de su media.

$(\hat{y}_i-\bar{y})$: Desviación explicada del valor observado, de su media.

$(y_i-\bar{y})$ : Desviación total del valor observado, de su media.

# Regresión Lineal Simple

Si generalizamos lo anterior para $n$ observaciones tenemos y sumando:

$$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i-\hat{y}_i)^2$$

Reordenando:

$$\sum(\hat{y}_i - \bar{y})^2 = \sum(y_i-\bar{y})^2 - \sum(y_i-\hat{y}_i)^2$$

# Regresión Lineal Simple

Dividiendo ambos miembros por la desviación total de los valores observados:

$$\underbrace{\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}}_{\text{Proporción Explicada}} = 1 - \underbrace{\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}}_{\text{Proporción No Explicada}}$$

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

$$R^2 = 1 - \frac{SS_R}{S_{yy}}$$

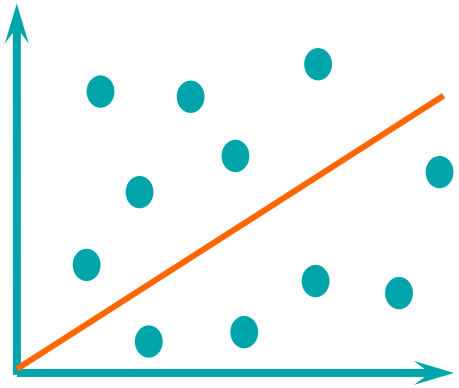
Se interpreta como la proporción de variación de $y$ observada que puede ser explicada por el modelo de regresión lineal simple.

# Regresión Lineal Simple

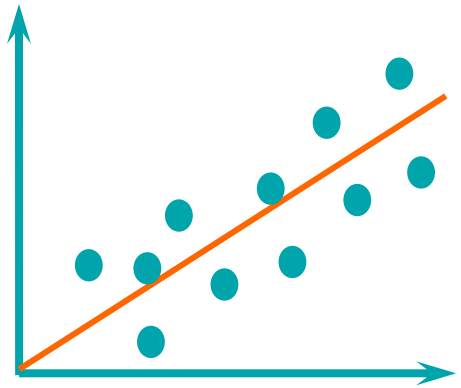
El coeficiente toma los siguientes valores:

$$0 \leq R^2 \leq 1$$

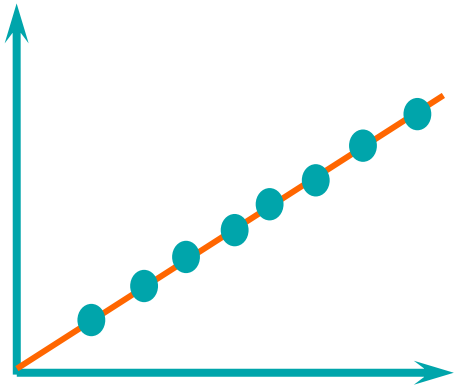
Mientras más próximo a la unidad el valor de $R^2$, mejor será nuestro modelo de regresión lineal simple al explicar la variación de $y$.



$R^2 = 0{,}2$



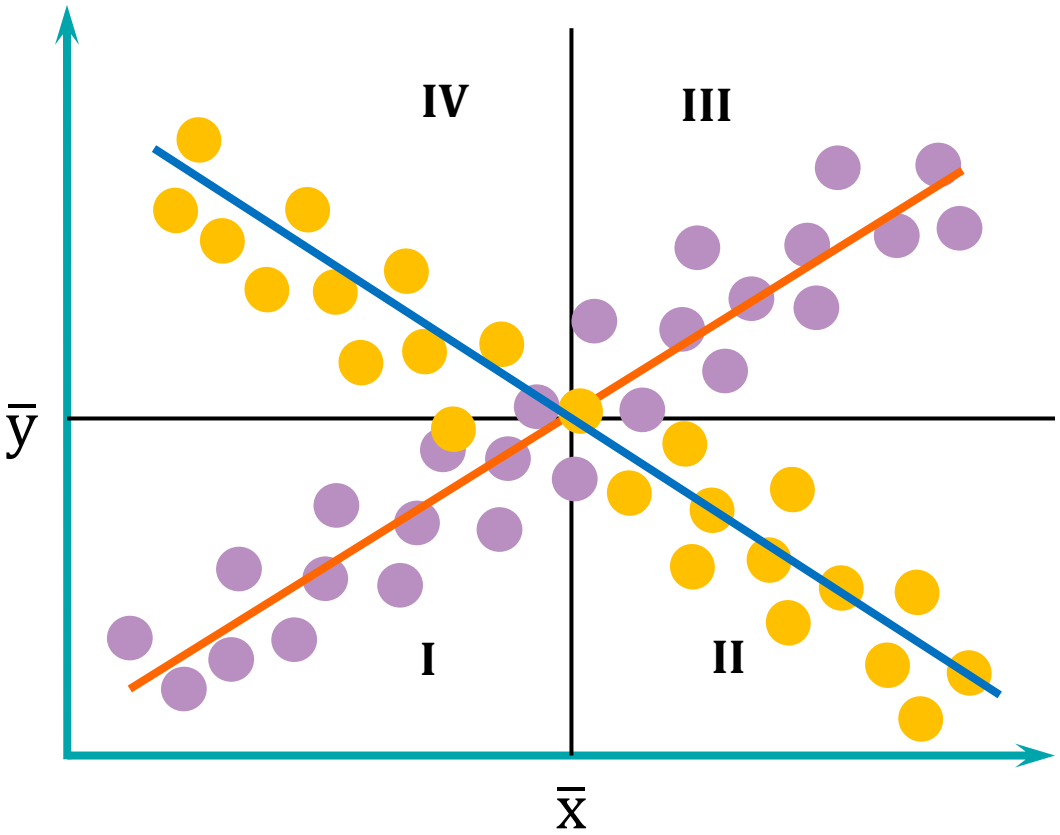
$R^2 = 0{,}8$



$R^2 = 1$

# Regresión Lineal Simple

## Covarianza




I) $\left.\begin{matrix}(x_i-\bar{x})<0\\(y_i-\bar{y})<0\end{matrix}\right\}\rightarrow(x_i-\bar{x})(y_i-\bar{y})>0$

III) $\left.\begin{matrix}(x_i-\bar{x})>0\\(y_i-\bar{y})>0\end{matrix}\right\}\rightarrow(x_i-\bar{x})(y_i-\bar{y})>0$

II) $\left.\begin{matrix}(x_i-\bar{x})>0\\(y_i-\bar{y})<0\end{matrix}\right\}\rightarrow(x_i-\bar{x})(y_i-\bar{y})<0$

IV) $\left.\begin{matrix}(x_i-\bar{x})<0\\(y_i-\bar{y})>0\end{matrix}\right\}\rightarrow(x_i-\bar{x})(y_i-\bar{y})<0$

# Regresión Lineal Simple

Entonces queremos una medida que sea positiva, si es que los datos se concentran en torno a la recta naranja y negativa si es que los datos se concentran entorno a la recta azul. Dicha medida se define como:

$$\mathrm{Cov}(x,y) = S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

La covarianza en general no es una función acotada, es decir no tiene un máximo y por lo tanto es difícil comparar relativamente para diferentes pares de variables.

# Regresión Lineal Simple

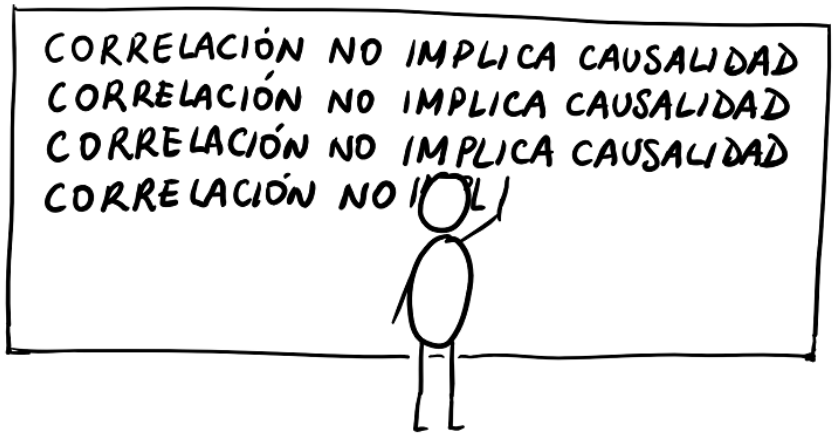
## Coeficiente de Correlación Lineal

$$\left|\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right| \leq \sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \cdot \sum_{i=1}^{n}(y_i-\bar{y})^2}$$

$$|S_{xy}| \leq \sqrt{S_{xx} \cdot S_{yy}}$$

$$\frac{|S_{xy}|}{\sqrt{S_{xx} \cdot S_{yy}}} \leq 1$$

$$|r| \leq 1$$



# Regresión Lineal Simple

**Coeficiente de Correlación Lineal:** El coeficiente de correlación es la medida para describir qué tan fuerte es la relación lineal entre dos variables.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

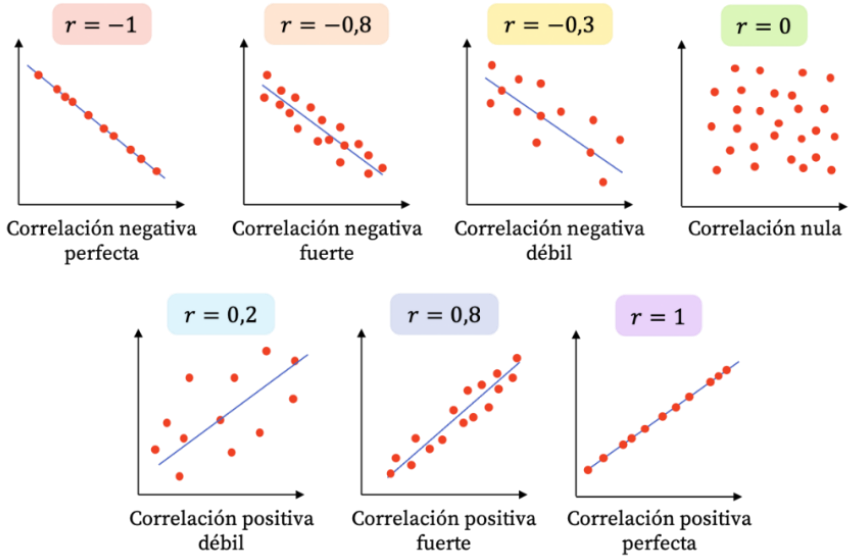
Otra forma de expresar el coeficiente de correlación lineal es:

$$r = \sqrt{R^2}$$

# Regresión Lineal Simple

**Propiedades del coeficiente de correlación lineal:**

1) $-1 \leq r \leq 1$
2) $r = 0 \rightarrow$ X e Y no están correlacionadas
3) $r = 1 \rightarrow$ Correlación directa perfecta
4) $r = -1 \rightarrow$ Correlación inversa perfecta
5) r Solo mide la correlación lineal entre X e Y



# Inferencias sobre los parámetros de Regresión


## Inferencias sobre $\beta_1$

Para poder hacer inferencia sobre $\beta_1$ necesitamos conocer lo siguiente sobre el estimador $\hat{\beta}_1$

$$E(\hat{\beta}_1) = \beta_1 \qquad V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \qquad S_{\hat{\beta}_1} = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

**Teorema:** La suposición del modelo de regresión lineal simple implica la variable estándar:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

# Inferencias sobre los parámetros de Regresión


## Intervalo de confianza para $\beta_1$

$$P\left(-t_{\frac{\alpha}{2},n-2} \leq T \leq t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \qquad \left(-t_{\frac{\alpha}{2},n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \leq t_{\frac{\alpha}{2},n-2}\right)$$

Despejando $\beta_1$ obtenemos un intervalo de nivel de confianza $1-\alpha$ para $\beta_1$

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2/S_{xx}}\,;\, \hat{\beta}_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2/S_{xx}}\right)$$

# Inferencias sobre los parámetros de Regresión


## Test de Hipótesis sobre $\beta_1$

$$H_0: \beta_1 \begin{cases} = \beta_{10} \\ = \beta_{10} \\ = \beta_{10} \end{cases} \qquad H_1: \beta_1 \begin{cases} > \beta_{10} \\ \neq \beta_{10} \\ < \beta_{10} \end{cases}$$

## Estadístico de prueba:

$$T = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2} \text{ bajo } H_0$$

## Regla de decisión:

Si $H_1: \beta_1 \neq \beta_{10}$ se rechaza si $|T| > t_{\frac{\alpha}{2},n-2}$

Si $H_1: \beta_1 > \beta_{10}$ se rechaza si $T > t_{\alpha,n-2}$

Si $H_1: \beta_1 < \beta_{10}$ se rechaza si $T < -t_{\alpha,n-2}$

# Inferencias sobre los parámetros de Regresión


## Test de Hipótesis sobre $\beta_1$

Un caso especial importante es cuando $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

Estas hipótesis están relacionadas con la ***significancia de la regresión***.

- Aceptar $H_0: \beta_1 = 0$ es equivalente a concluir que no hay ninguna relación lineal entre x e Y.
- Si $H_0: \beta_1 = 0$ se rechaza implica que x tiene importancia al explicar la variabilidad en Y.

También puede significar que el modelo lineal es adecuado, o que, aunque existe efecto lineal pueden obtenerse mejores resultados agregando términos polinomiales de mayor grado en x.

# Inferencias sobre los parámetros de Regresión


## Inferencias sobre $\beta_0$

Para poder hacer inferencia sobre $\beta_0$ necesitamos conocer lo siguiente sobre el estimador $\hat{\beta}_0$

$$E(\hat{\beta}_0) = \beta_0 \qquad V(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \qquad S_{\hat{\beta}_0} = \sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

**Teorema:** La suposición del modelo de regresión lineal simple implica la variable estándar:

$$T = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

# Inferencias sobre los parámetros de Regresión


## Intervalo de confianza para $\beta_0$

$$P\left(-t_{\frac{\alpha}{2},n-2} \leq T \leq t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \qquad \left(-t_{\frac{\alpha}{2},n-2} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}} \leq t_{\frac{\alpha}{2},n-2}\right)$$

Despejando $\beta_0$ obtenemos un intervalo de nivel de confianza $1-\alpha$ para $\beta_0$

$$\left(\hat{\beta}_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\,;\hat{\beta}_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}\right)$$

# Inferencias sobre los parámetros de Regresión


## Test de Hipótesis sobre $\beta_0$

$$H_0: \beta_0 \begin{cases} = \beta_{00} \\ = \beta_{00} \\ = \beta_{00} \end{cases} \qquad H_1: \beta_0 \begin{cases} > \beta_{00} \\ \neq \beta_{00} \\ < \beta_{00} \end{cases}$$

## Estadístico de prueba:

$$T = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2} \text{ bajo } H_0$$

## Regla de decisión:

Si $H_1: \beta_0 \neq \beta_{00}$ se rechaza si $|T| > t_{\frac{\alpha}{2},n-2}$

Si $H_1: \beta_0 > \beta_{00}$ se rechaza si $T > t_{\alpha,n-2}$

Si $H_1: \beta_0 < \beta_{00}$ se rechaza si $T < -t_{\alpha,n-2}$

# Inferencias sobre los parámetros de Regresión


## Intervalo de confianza para la respuesta media

Buscamos un intervalo de confianza para estimar $\beta_0 + \beta_1 x^*$, es decir estimar la media $E(Y)$ para un valor especifico $x^*$

Sea $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ , donde $x^*$ es algún valor fijo de $x$. Entonces:

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \beta_0 + \beta_1 x^*$$

$$V(\hat{Y}) = V(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

$$S_{\hat{Y}} = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

$\hat{Y}$ tiene distribución normal con esperanza y varianza anteriores

# Inferencias sobre los parámetros de Regresión


**Teorema:** La variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{\sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}}$$

tiene distribución student con $n - 2$ grados de libertad

# Inferencias sobre los parámetros de Regresión


Un **intervalo de confianza** de $(1-\alpha)$**para** $\boldsymbol{\beta_0 + \beta_1 x^*}$ de la línea de regresión verdadera es

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x^* - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}\right]}\,;\, \hat{\beta}_0 + \hat{\beta}_1 x^* + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x^*-\bar{x})^2}{S_{xx}}\right]}\right]$$

**Observaciones:**

- El ancho del intervalo de confianza para $\beta_0 + \beta_1 x^*$ depende de $x^*$. El ancho del intervalo es mínimo cuando $x^* = \bar{x}$ y crece a medida que $|x^* - \bar{x}|$ aumenta
- Al repetir los cálculos anteriores para varios valores diferentes de $x^*$ pueden obtenerse intervalos de confianza para cada valor correspondiente de $\beta_0 + \beta_1 x^*$

# Inferencias sobre los parámetros de Regresión


## Intervalo de Predicción para futuras observaciones de $Y^*$

Un valor futuro de **Y** no es un parámetro sino una variable aleatoria; por eso se hace referencia a un intervalo de valores factibles para un valor **Y** futuro como **Intervalo de Predicción** en lugar de intervalo de confianza.

***El error de predicción es $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$ una diferencia entre dos variables aleatorias. Existe por lo tanto más incertidumbre en la predicción que, en la estimación, así que un intervalo de predicción será más ancho que un intervalo de confianza.***

$$E(Y^* - \hat{Y}^*) = 0 \qquad V(Y^* - \hat{Y}^*) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \qquad S_{Y^* - \hat{Y}^*} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}$$

Por lo tanto, $Y^* - \hat{Y}^*$ tiene distribución normal con esperanza y varianza anteriores

# Inferencias sobre los parámetros de Regresión


**Teorema:** La variable

$$T = \frac{Y^* - \hat{Y}^*}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}}$$

tiene distribución student con $n - 2$ grados de libertad

Por el argumento usual llegamos al siguiente intervalo de predicción de nivel $(1 - \alpha)$ para $Y^*$

$$\left[\hat{Y}^* - t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}; \hat{Y}^* + t_{\frac{\alpha}{2},n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}\right]$$

# Inferencias sobre los parámetros de Regresión


**Cálculo del Valor Critico $t_{\frac{\alpha}{2},n-2}$ o $t_{\alpha,n-2}$ en Probability Distributions :**

$$X \sim t(v)$$

$$v = n - 2$$

$$x = t_{\frac{\alpha}{2},n-2} \text{ o } t_{\alpha,n-2}$$

$$P(X > x) = \alpha/2 \text{ o } \alpha$$

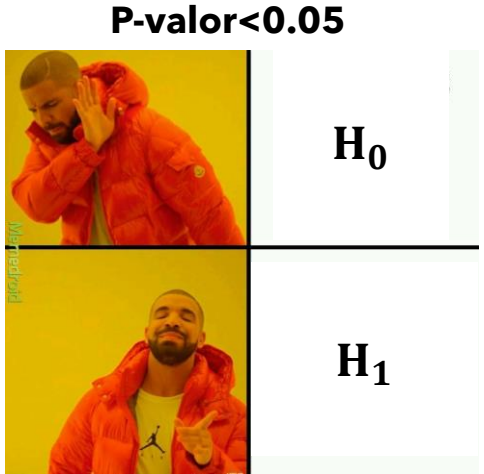
# Inferencias sobre los parámetros de Regresión


Un P-valor muy chico significa mucha evidencia en contra de $H_0$ ; un P-valor alto significa que no hay evidencia en contra $H_0$ .

**Regla de decisión:**

Si $P-\text{valor} > 0.05$ entonces no se rechaza $H_0$ con nivel de significancia $0.05$

Si $P-\text{valor} < 0.05$ entonces se rechaza $H_0$ con nivel de significancia $0.05$



# Inferencias sobre los parámetros de Regresión


## Cálculo de P-valor para t

Si las hipótesis son $H_0: \beta_1 = \beta_{10}$ vs $H_1: \beta_1 \neq \beta_{10}$ :

$$P - valor = P(|T| > |t_0|)$$

Si las hipótesis son $H_0: \beta_1 = \beta_{10}$ vs $H_1: \beta_1 > \beta_{10}$:

$$P - valor = P(T > t_0)$$

Si las hipótesis son $H_0: \beta_1 = \beta_{10}$ vs $H_1: \beta_1 < \beta_{10}$:

$$P - valor = P(T < t_0)$$

De la misma forma para $\beta_0$

# Inferencias sobre los parámetros de Regresión


**Calculo $P-valor = P(|T| > |t_0|)$ en Probability Distributions:**

$$X \sim t(v)$$

$$v = n - 2$$

$$x = |t_0|$$

$$2P(X > |x|) = P - valor$$

# Inferencias sobre los parámetros de Regresión


**Calculo $P - valor = P(T > t_0)$ en Probability Distributions:**

$$X \sim t(v)$$

$$v = n - 2$$

$$x = t_0$$

$$P(X > x) = P - valor$$

# Inferencias sobre los parámetros de Regresión


**Calculo $P - valor = P(T < t_0)$ en Probability Distributions:**

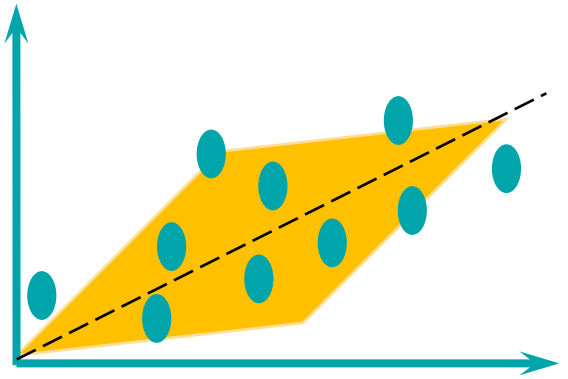
$$X \sim t(v)$$

$$v = n - 2$$

$$x = t_0$$

$$P(X < x) = P - valor$$

# Regresión Lineal Múltiple



## Modelo de Regresión Lineal Múltiple

$$y = \beta_0 + \beta_{1i}x_1 + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon$$

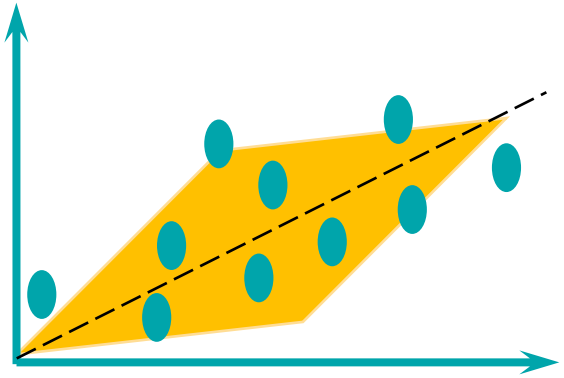
Las Suposiciones son las mismas que el modelo de regresión lineal simple y la estimación de los parámetros se obtendrán de la misma forma, por el método de mínimos cuadrados.

Función a minimizar:

$$f(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki})]^2$$

$$f(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_k x_{ki})^2$$

# Regresión Lineal Múltiple



Haciendo las respectivas derivadas igualadas a cero y despejando, se obtiene las **ecuaciones normales:**

$$n\beta_0 + \beta_1 \sum x_{1i} + \beta_2 \sum x_{2i} + \cdots + \beta_k \sum x_{ki} = \sum y_i$$

$$\beta_0 \sum x_{1i} + \beta_1 \sum x_{1i}^2 + \beta_2 \sum x_{1i}x_{2i} + \cdots + \beta_k \sum x_{1i}x_{ki} = \sum x_{1i}y_i$$

$$\beta_0 \sum x_{2i} + \beta_1 \sum x_{1i}x_{2i} + \beta_2 \sum x_{2i}^2 + \cdots + \beta_k \sum x_{2i}x_{ki} = \sum x_{2i}y_i$$

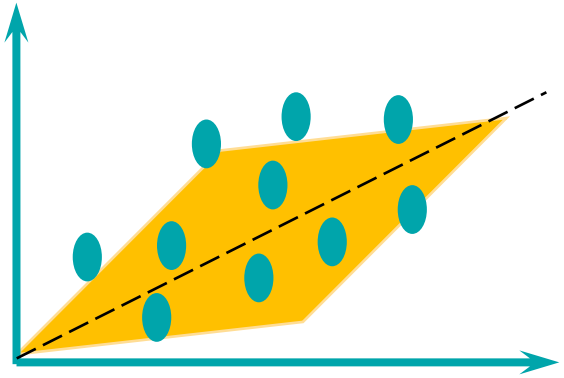
. . .

. . .

. . .

$$\beta_0 \sum x_{ki} + \beta_1 \sum x_{ki}x_{1i} + \beta_2 \sum x_{ki}x_{2i} + \cdots + \beta_k \sum x_{ki}^2 = \sum x_{ki}y_i$$

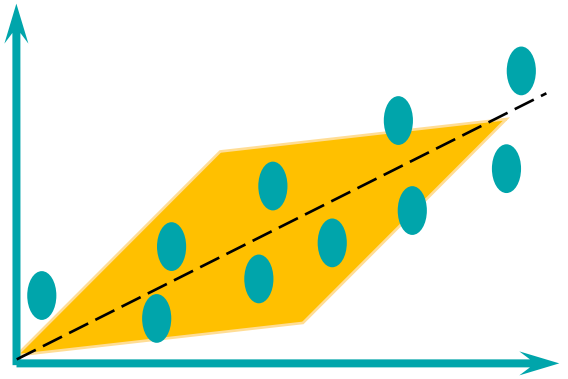
# Regresión Lineal Múltiple



El sistema de ecuaciones obtenidas se dispone en forma matricial de la siguiente manera:

$$\underbrace{\begin{pmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{2i}x_{ki} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki}x_{1i} & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}^2 \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \\ \vdots \\ \sum x_{ki}y_i \end{pmatrix}}_{Y}$$

# Regresión Lineal Múltiple



Entonces queda el siguiente sistema matricial a resolver:

$$Y = X\beta$$

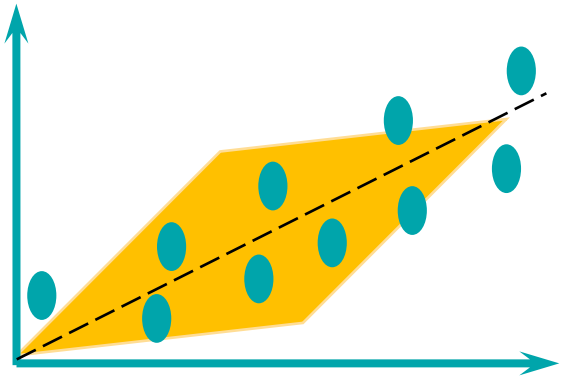
**1°) Forma:**

$\hat{\beta} = (X^T X)^{-1} X^T Y$ Donde $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)$

Entonces:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

# Regresión Lineal Múltiple

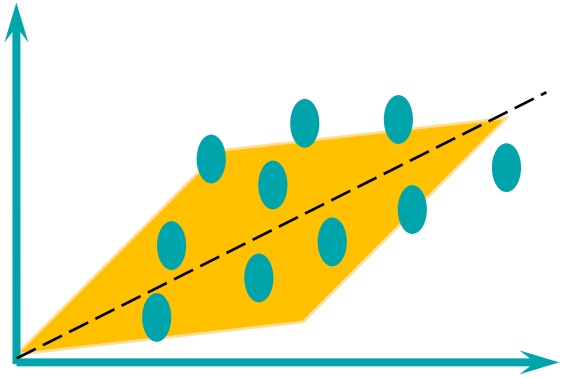


**2°) Forma:** Usando determinantes

Denotamos $\mathbf{D}$ (con $D \neq 0$) al determinante de la matriz $X$:

$$\begin{pmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{2i}x_{ki} \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki}x_{1i} & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}^2 \end{pmatrix}$$

# Regresión Lineal Múltiple

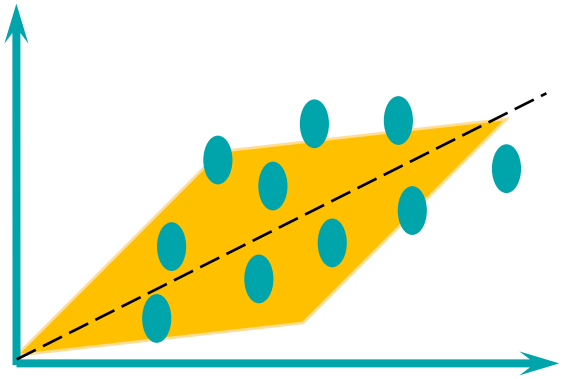


Denotamos $\mathbf{D_0}$ al determinante de la matriz:

$$\begin{pmatrix} \sum y_i & \sum x_{1i} & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i}y_i & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{2i}x_{ki} \\ \sum x_{2i}y_i & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki}y_i & \sum x_{ki}x_{1i} & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}^2 \end{pmatrix}$$

Lo que hicimos fue reemplazar la primera columna de la matriz $\mathbf{X}$ por $\mathbf{Y}$

# Regresión Lineal Múltiple

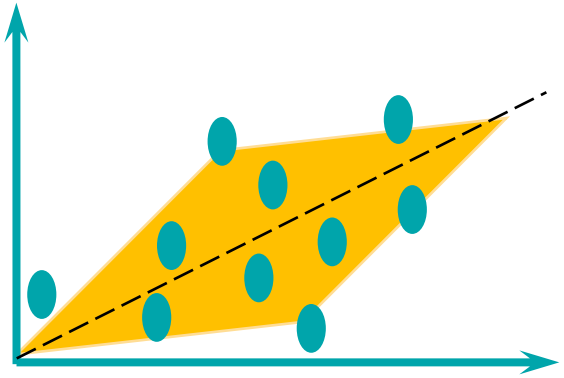


Denotamos $\mathbf{D_1}$ al determinante de la matriz:

$$\begin{pmatrix} n & \sum y_i & \sum x_{2i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}y_i & \sum x_{1i}x_{2i} & \cdots & \sum x_{2i}x_{ki} \\ \sum x_{2i} & \sum x_{2i}y_i & \sum x_{2i}^2 & \cdots & \sum x_{2i}x_{ki} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki}y_i & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}^2 \end{pmatrix}$$

Lo que hicimos fue reemplazar la segunda columna de la matriz $X$ por $Y$

# Regresión Lineal Múltiple

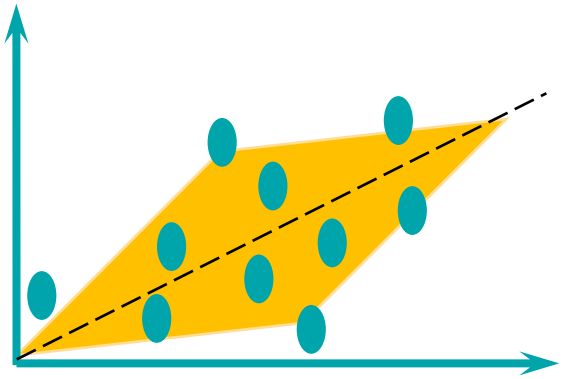


Denotamos $\mathbf{D_k}$ al determinante de la matriz :

$$\begin{pmatrix} n & \sum x_{1i} & \sum x_{2i} & \cdots & \sum y_i \\ \sum x_{1i} & \sum x_{1i}^2 & \sum x_{1i}x_{2i} & \cdots & \sum x_{1i}y_i \\ \sum x_{2i} & \sum x_{1i}x_{2i} & \sum x_{2i}^2 & \cdots & \sum x_{2i}y_i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki}x_{1i} & \sum x_{ki}x_{2i} & \cdots & \sum x_{ki}y_i \end{pmatrix}$$

Lo que hicimos fue reemplazar la última columna de la matriz $\mathbf{X}$ por $\mathbf{Y}$

# Regresión Lineal Múltiple



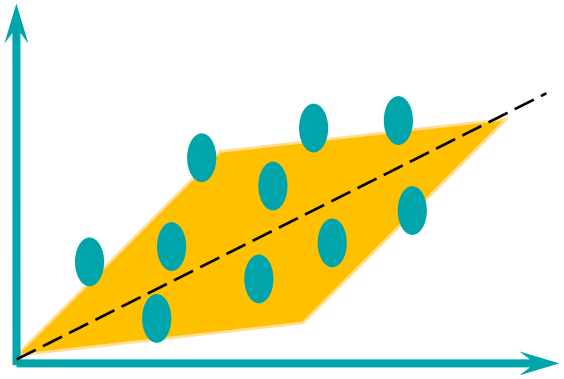
Los estimadores de los coeficientes se calcularán como:

$$\hat{\beta}_0 = \frac{D_0}{D} \quad \hat{\beta}_1 = \frac{D_1}{D} \quad \hat{\beta}_2 = \frac{D_2}{D} \quad \cdots\cdots \quad \hat{\beta}_k = \frac{D_k}{D}$$

Entonces:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

# Regresión Lineal Múltiple



## Estimación de la varianza

$$\hat{\sigma}^2 = \frac{SS_R}{n-k-1} = \frac{\Sigma(y_i - \hat{y}_i)^2}{n-k-1}$$

## Coeficiente de determinación ajustado:

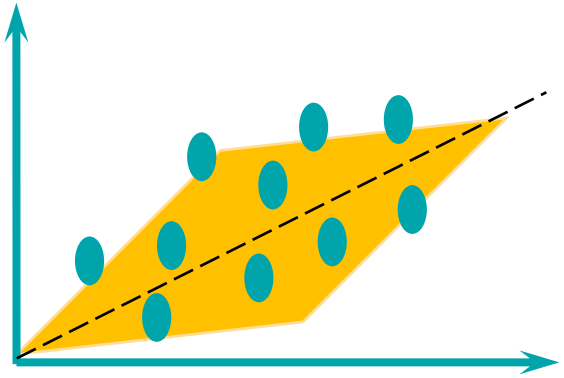
$$R_a^2 = 1 - (1 - R^2)\left(\frac{n-1}{n-k-1}\right) \rightarrow 0 \leq R_a^2 \leq 1$$

## Coeficiente de correlación:

$$r_a = \sqrt{R_a^2} \rightarrow -1 \leq r_a \leq 1$$

Siendo k el número de variables independientes

# Regresión Lineal Múltiple

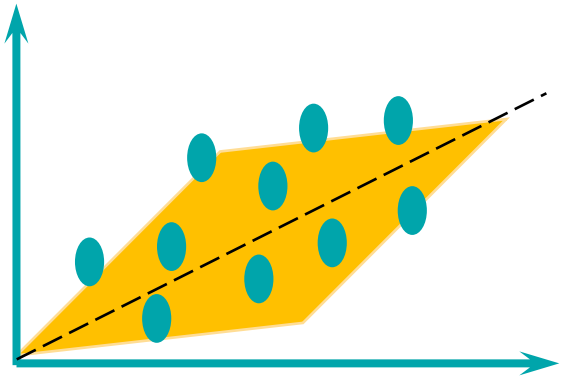


**Ejemplo:** Se realizo un experimento para determinar si era posible predecir el peso de un animal después de un periodo determinado con base en su peso inicial y la cantidad de alimento que consumía. Se registran los siguientes datos en kilogramos:

| N° de animal | Peso final del animal (Y) | Peso inicial del animal (X1) | Peso del Alimento (X2) |
|---|---|---|---|
| 1 | 95 | 40 | 280 |
| 2 | 75 | 35 | 225 |
| 3 | 80 | 30 | 260 |
| 4 | 100 | 45 | 290 |
| 5 | 95 | 40 | 310 |
| 6 | 70 | 35 | 185 |
| 7 | 50 | 30 | 175 |
| 8 | 80 | 45 | 230 |
| 9 | 90 | 45 | 235 |
| 10 | 85 | 35 | 230 |

a) Estime la ecuación de regresión correspondiente para el problema.
b) Prediga cuanto pesara un animal que comienza pesando 35 kg después de consumir 250 kg de alimentos.
c) Calcule los indicadores para la misma.

# Regresión Lineal Múltiple

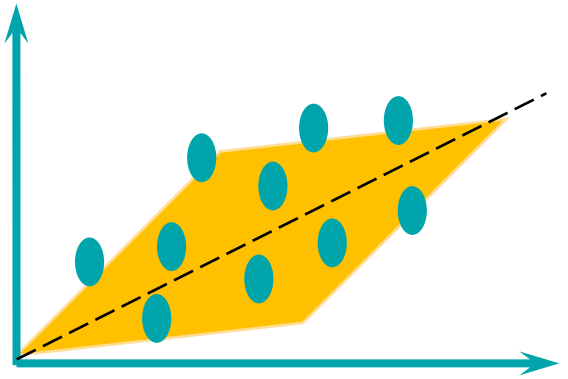


**Resolución**

**a)** Como son dos las variables independientes $X_1$ y $X_2$ que se consideran la ecuación a estimar será:

$$\begin{cases} n\beta_0 + \beta_1 \sum x_{1i} + \beta_2 \sum x_{2i} = \sum y_i \\ \beta_0 \sum x_{1i} + \beta_1 \sum x_{1i}^2 + \beta_2 \sum x_{1i}x_{2i} = \sum x_{1i}y_i \\ \beta_0 \sum x_{2i} + \beta_1 \sum x_{1i}x_{2i} + \beta_2 \sum x_{2i}^2 = \sum x_{2i}y_i \end{cases}$$

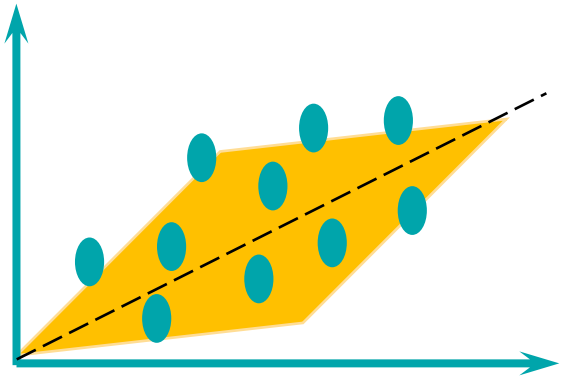
# Regresión Lineal Múltiple



Calculamos las sumatorias correspondientes a partir de los datos:

| N° de animal | Peso final del animal $(Y)$ | Peso inicial del animal $(X_1)$ | Peso del Alimento $(X_2)$ | $(X_1^2)$ | $(X_2^2)$ | $X_1 \cdot X_2$ | $X_1 \cdot Y$ | $X_2 \cdot Y$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 95 | 40 | 280 | 1600 | 78400 | 11200 | 3800 | 26600 |
| 2 | 75 | 35 | 225 | 1225 | 50625 | 7875 | 2625 | 16875 |
| 3 | 80 | 30 | 260 | 900 | 67600 | 7800 | 2400 | 20800 |
| 4 | 100 | 45 | 290 | 2025 | 84100 | 13050 | 4500 | 29000 |
| 5 | 95 | 40 | 310 | 1600 | 96100 | 12400 | 3800 | 29450 |
| 6 | 70 | 35 | 185 | 1225 | 34225 | 6475 | 2450 | 12950 |
| 7 | 50 | 30 | 175 | 900 | 30625 | 5250 | 1500 | 8750 |
| 8 | 80 | 45 | 230 | 2025 | 52900 | 10350 | 3600 | 18400 |
| 9 | 90 | 45 | 235 | 2025 | 55225 | 10575 | 4050 | 21150 |
| 10 | 85 | 35 | 230 | 1225 | 52900 | 8050 | 2975 | 19550 |
| **Sumatoria** | **820** | **380** | **2420** | **14750** | **602700** | **93025** | **31700** | **203525** |

# Regresión Lineal Múltiple



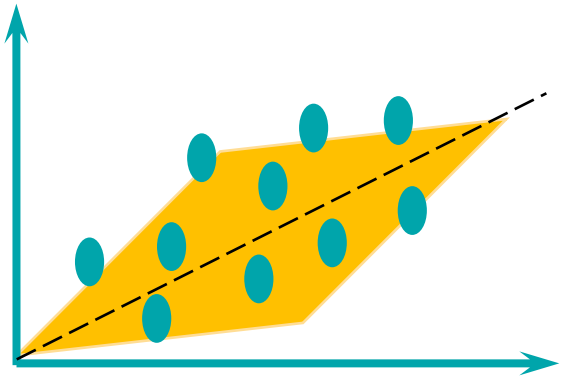
Reemplazando los valores en nuestra ecuación:

$$\begin{cases} 10\beta_0 + 380\beta_1 + 2420\beta_2 = 820 \\ 380\beta_0 + 14750\beta_1 + 93025\beta_2 = 31700 \\ 2420\beta_0 + 93025\beta_1 + 602700\beta_2 = 203525 \end{cases}$$

En forma matricial:

$$\begin{pmatrix} 10 & 380 & 2420 \\ 380 & 14750 & 93025 \\ 2420 & 93025 & 602700 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} 820 \\ 31700 \\ 203525 \end{pmatrix}$$

# Regresión Lineal Múltiple



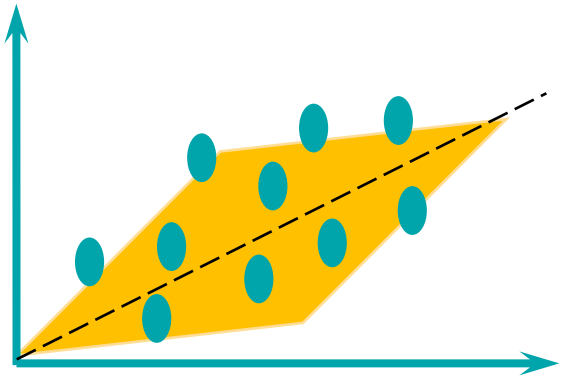
Resolviendo el producto matricial:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -11{,}0546 \\ 0{,}9139 \\ 0{,}2410 \end{pmatrix}$$

Entonces la ecuación de regresión estimada será:

$$\hat{y} = -11{,}0546 + 0{,}9139x_1 + 0{,}2410x_2$$

# Regresión Lineal Múltiple



**b)** Predicción del peso de un animal que comienza pesando 35 kg después de consumir 250 kg de alimentos:

$$\hat{y} = -11{,}0546 + 0{,}9139 \cdot (35) + 0{,}2410 \cdot (250) = 81{,}2 \text{ kg}$$

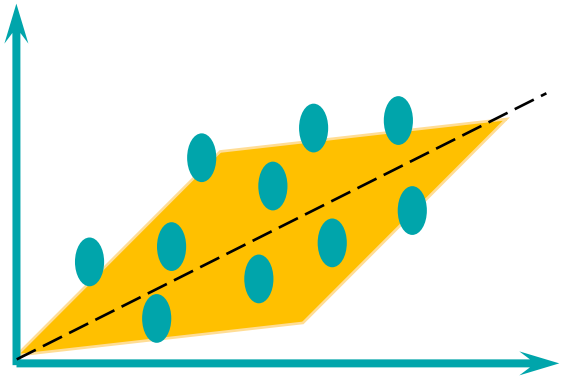
**c)** Calculo de los indicadores

Necesitamos:

$$SS_R = \sum (y_i - \hat{y}_i)^2 = 240{,}9284 \qquad R^2 = 1 - \frac{SS_R}{S_{yy}} = 0{,}8771$$

$$Syy = \sum (y_i - \bar{y})^2 = 1960$$

# Regresión Lineal Múltiple



Entonces:

$$\hat{\sigma}^2 = \frac{SS_R}{n-k-1} = \frac{240{,}9284}{10-2-1} = 34{,}4183$$

$$R_a^2 = 1-(1-R^2)\left(\frac{n-1}{n-k-1}\right) = 1-(1-0{,}8771)\left(\frac{10-1}{10-2-1}\right) = 0{,}8420$$

$$r_a = \sqrt{R_a^2} = \sqrt{0{,}8420} = 0{,}91$$


Fin
¡Muchas Gracias !