**Universidad Nacional Autónoma de México**
Facultad de Ingeniería

# Pronóstico y clasificación

## Enfoques de aprendizaje automático

**Guillermo Molero-Castillo**
guillermo.molero@ingenieria.unam.edu

Diciembre, 2020

# Contexto

## Aprender

Para adquirir conocimientos o comprensión mediante el análisis o la experiencia.

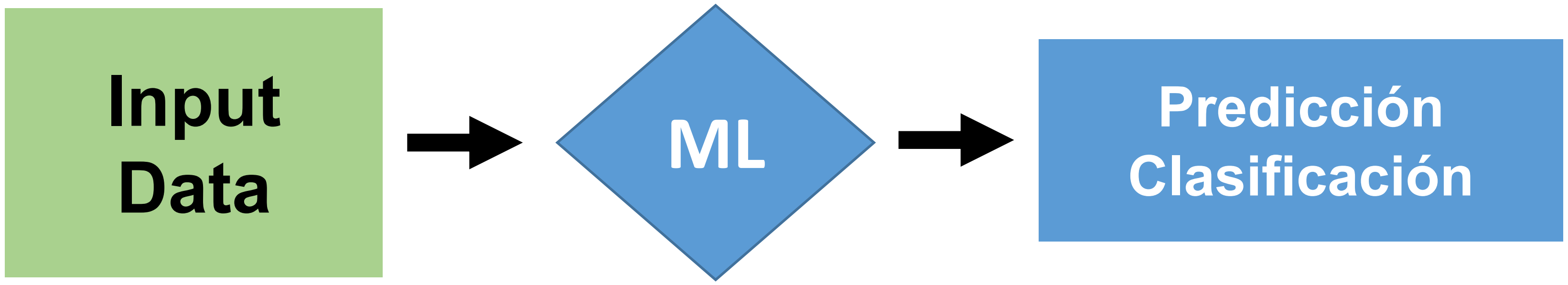
## Aprendizaje automático

- Es aprender automáticamente a partir de los datos, a través de un proceso de inferencia, ajuste de modelo o aprendizaje de ejemplos (reconocimiento de rostros o personas).
- Es ideal para áreas con gran cantidad de datos en ausencia de una teoría general.
- La extracción automática de información es útil mediante la construcción de buenos modelos inferenciales.

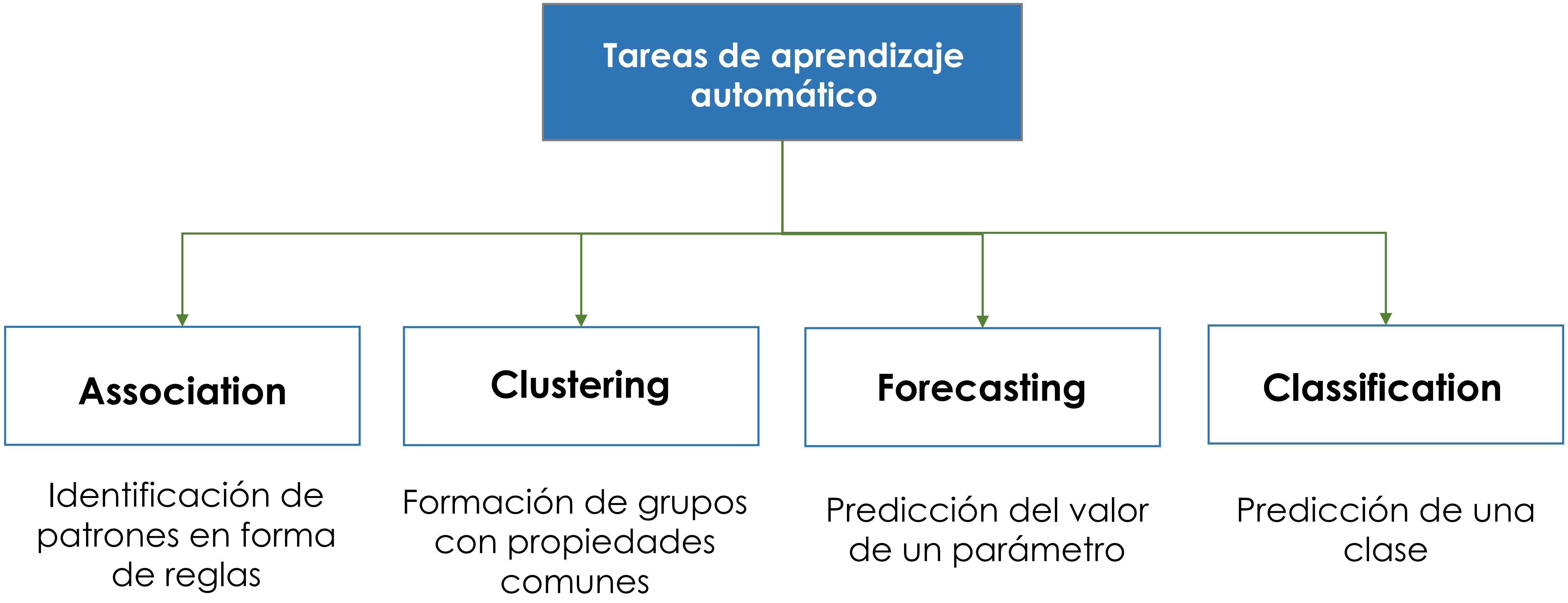
# Contexto

En la actualidad, los modelos de pronóstico se han vuelto predominantes. En diferentes disciplinas se trata de aplicar los algoritmos de aprendizaje automático:

- Los economistas para predecir los precios del mercado, obtener ganancias.
- Los médicos para diagnóstico, por ejemplo, clasificar si un tumor es maligno o benigno.
- Los meteorólogos para predecir el clima.
- Los reclutadores de recursos humanos para verificar si el solicitante cumple con los criterios mínimos para el trabajo.
- Entre otros.

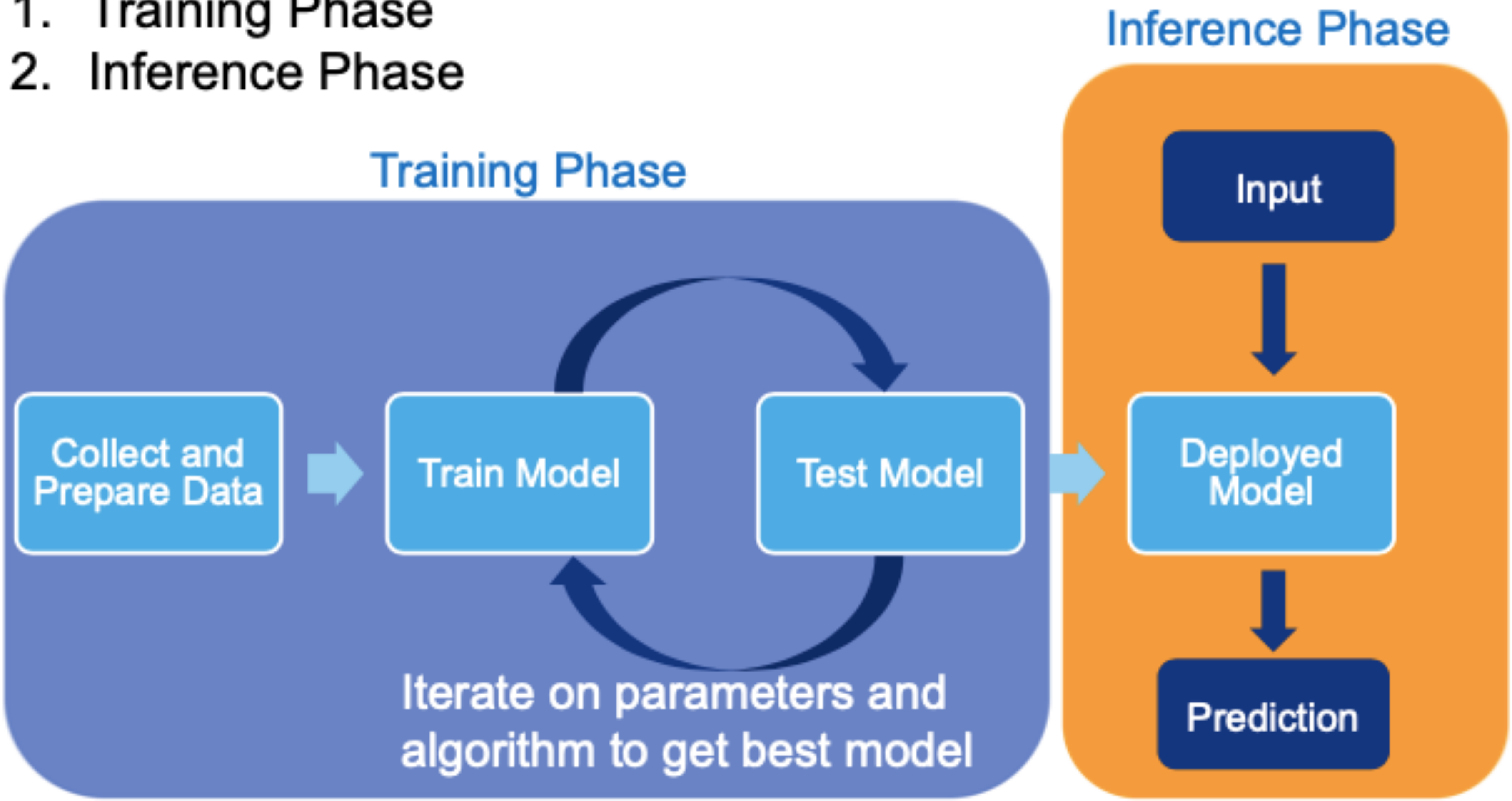


# Contexto


Tareas de aprendizaje automático
Association
Clustering
Forecasting
Classification
Identificación de patrones en forma de reglas
Formación de grupos con propiedades comunes
Predicción del valor de un parámetro
Predicción de una clase

# Contexto

1. Training Phase
2. Inference Phase



## Algoritmos

La predicción es importante, ayuda a automatizar actividades. Sin embargo, solo dice lo que sucederá, pero no lo que se debería hacer.

- Linear regression / Logistic regression
- Nearest Neighbor (kNN)
- Decision trees
- Support vector machines
- Artificial Neural Networks
- Bayesian methods
- …

Los diferentes algoritmos tienen diferentes fortalezas y debilidades.
Se debe seleccionar el enfoque de predicción que sea adecuado para el problema.

# Regresión lineal

# Regresión lineal

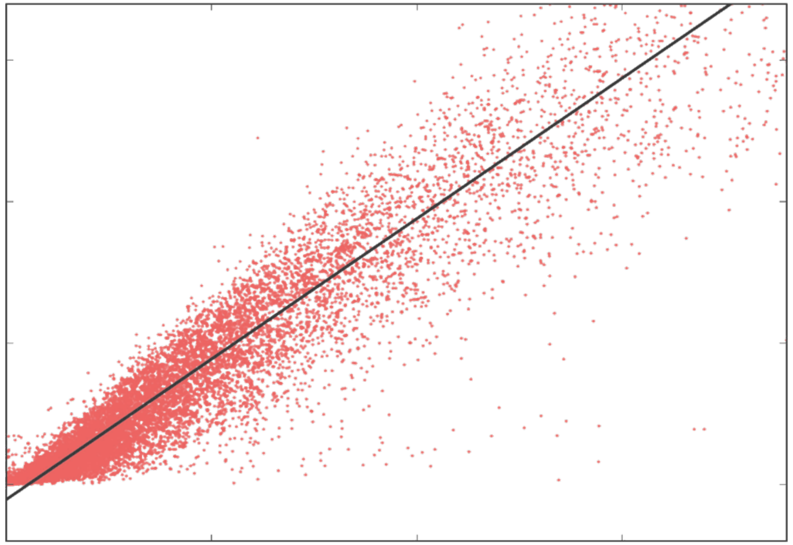
La regresión lineal es un algoritmo básico (rudimentario) de aprendizaje supervisado, cuyo objetivo es calcular una ecuación que minimiza la distancia entre la línea ajustada (recta) y todos los puntos de datos.

El propósito es proporcionar una base para desarrollar y aprender otros algoritmos de ML.

$$Y = a + biXi + u$$

**Y**

Es la variable dependiente (output)



**X**

Es la variable independiente (input)

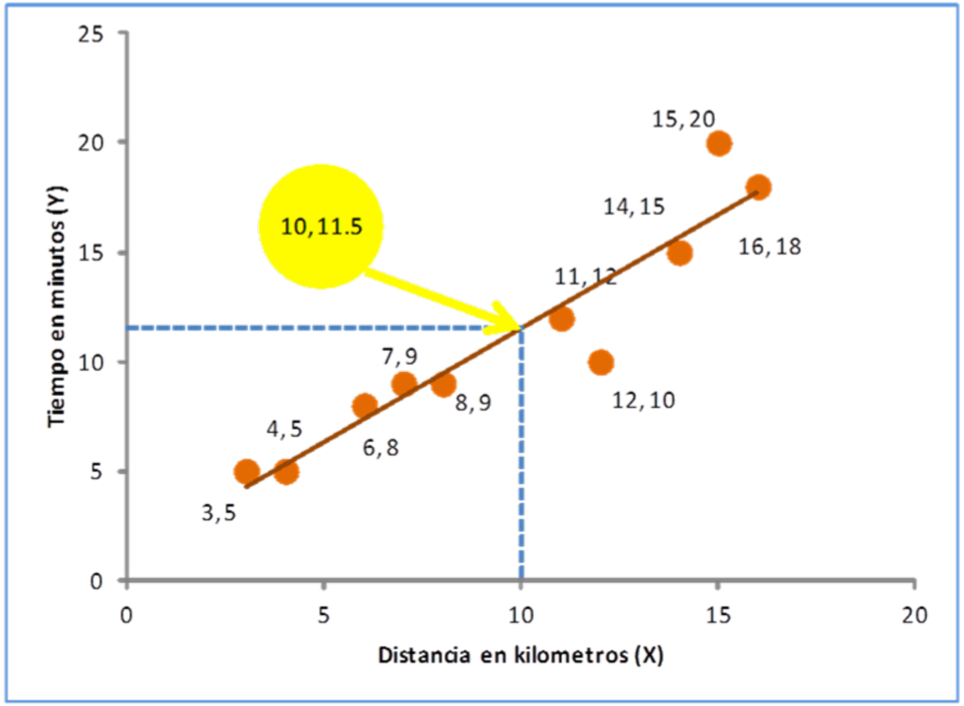
| F1 | F2 | F3 | F4 | Class (Y) |
|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 0.5 |
| 0 | 1 | 1 | 0 | 2 |
| 1 | 0 | 1 | 0 | 0 |
| 0 | 1 | 0 | 1 | 1.2 |

# Regresión lineal

La regresión según el número de variables son:

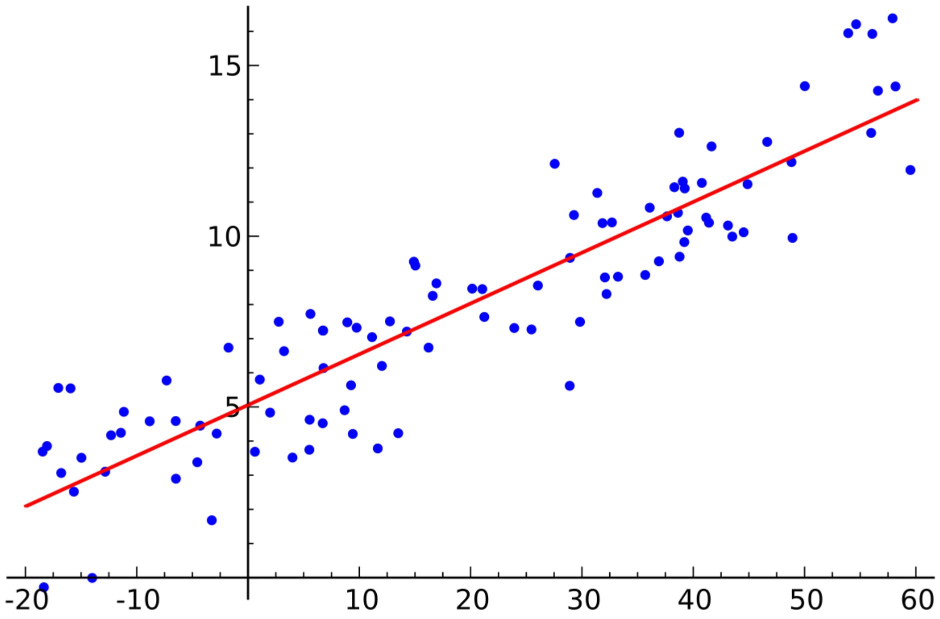
- Si se tiene dos variables (2 dimensiones) se trata de un problema de ***regresión lineal simple***.
- Si se tiene más de dos variables se trata de un problema de ***regresión lineal múltiple***.

Se utiliza para pronosticar una variable dependiente (clase): $\boldsymbol{Y}$, en función de una o más variables independientes: $\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, …, $\boldsymbol{X_n}$



# Regresión lineal

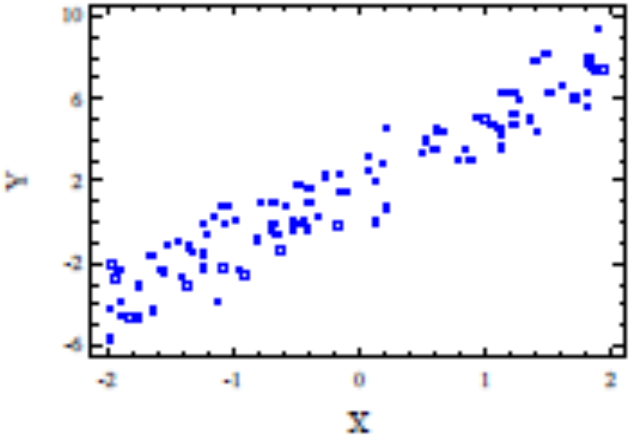
- Un diagrama de dispersión ofrece una aproximación sobre el **tipo de relación** entre las variables.
- Con base en los datos, se traza una recta (línea) que modele mejor los puntos.
- A esto se conoce como la recta de mejor ajuste.



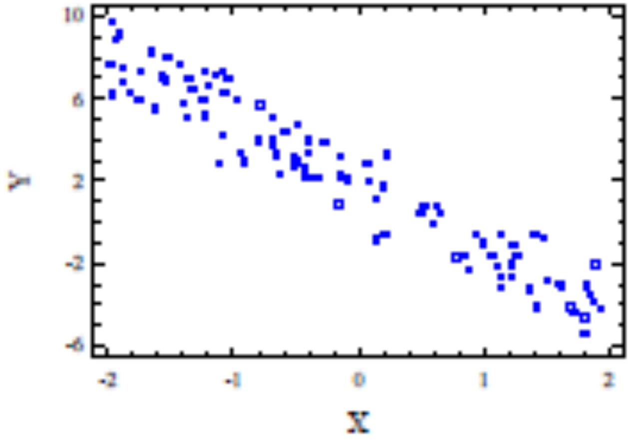

El patrón sugiere una relación positiva

# Tipos de relación

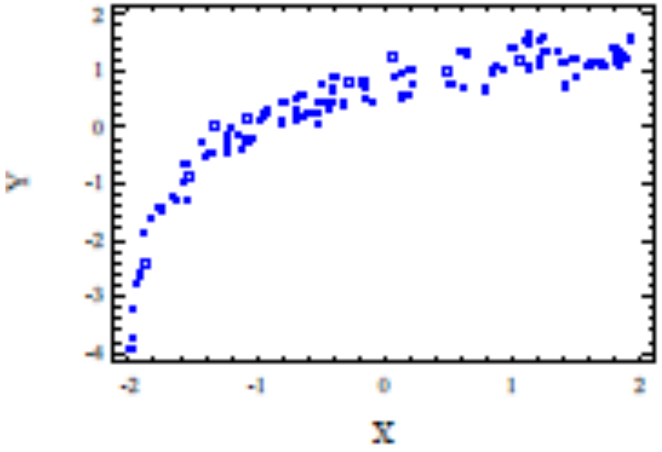
**Relación lineal positiva**



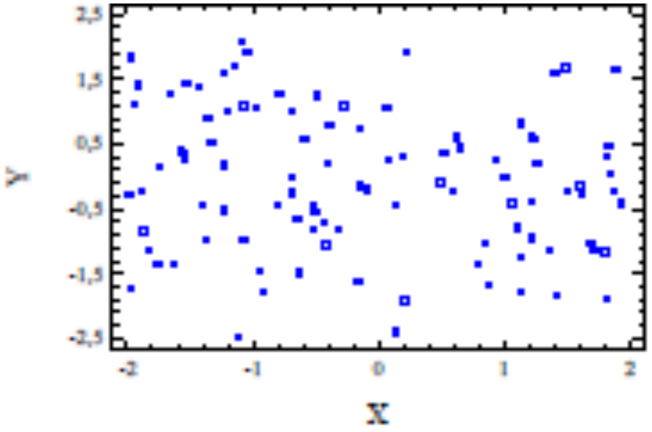

**Relación lineal negativa**




**Relación no lineal**




**Ausencia de relación**



# Regresión lineal simple

# Regresión lineal simple

En una regresión lineal simple se evalúa una sola variable independiente (X), cuya ecuación lineal es:

Valores observados

1. Dada una entrada ***X***, se calcula una salida ***Y***: $Y = a + bX + \ u$
2. Para esto se estiman los parámetros ***a*** y ***b*** (conocidos como coeficientes)

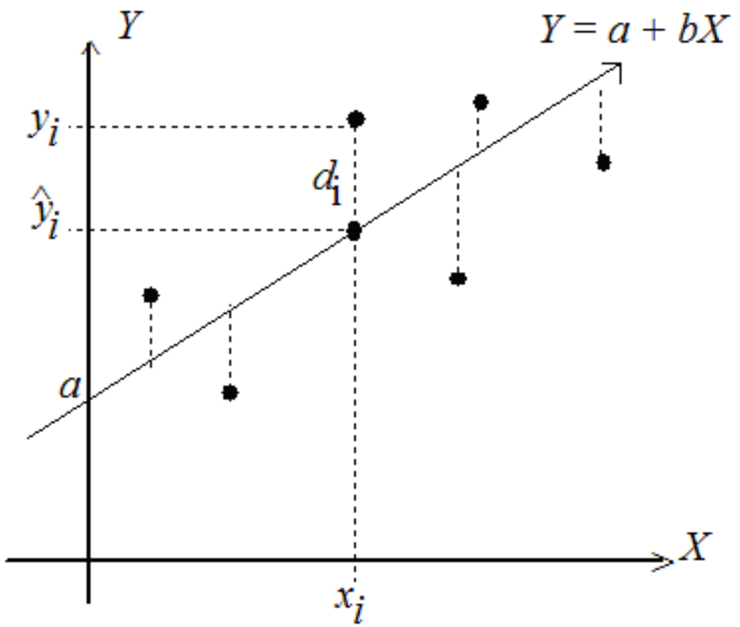
donde:

- ***a*** es el intercepto (corta el eje Y).

  $a = \bar{y} - b\bar{x}$

- ***b*** es la pendiente de la recta.

  $b = \dfrac{\mathrm{Cov(x, y)}}{\mathrm{S}_x^2}$



# Regresión lineal múltiple

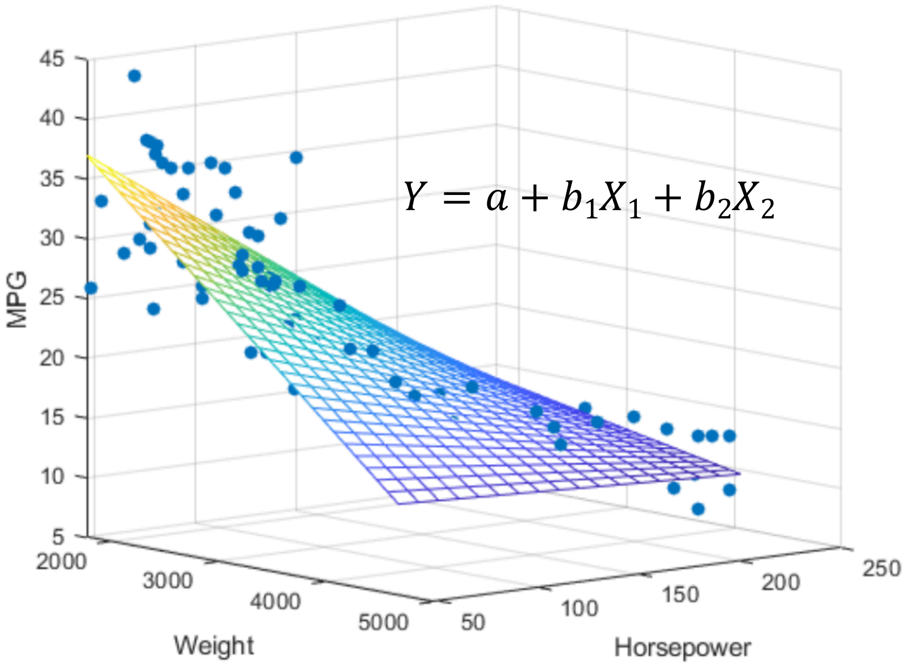
# Regresión lineal múltiple

En una **regresión lineal múltiple** se evalúa dos o más variables independientes ($X_1$, $X_2$, …, $X_n$).

1. Se ajusta una regresión lineal: $Y = a + b_1X_1 + b_2X_2 \ldots + b_nX_n + u$

Donde:

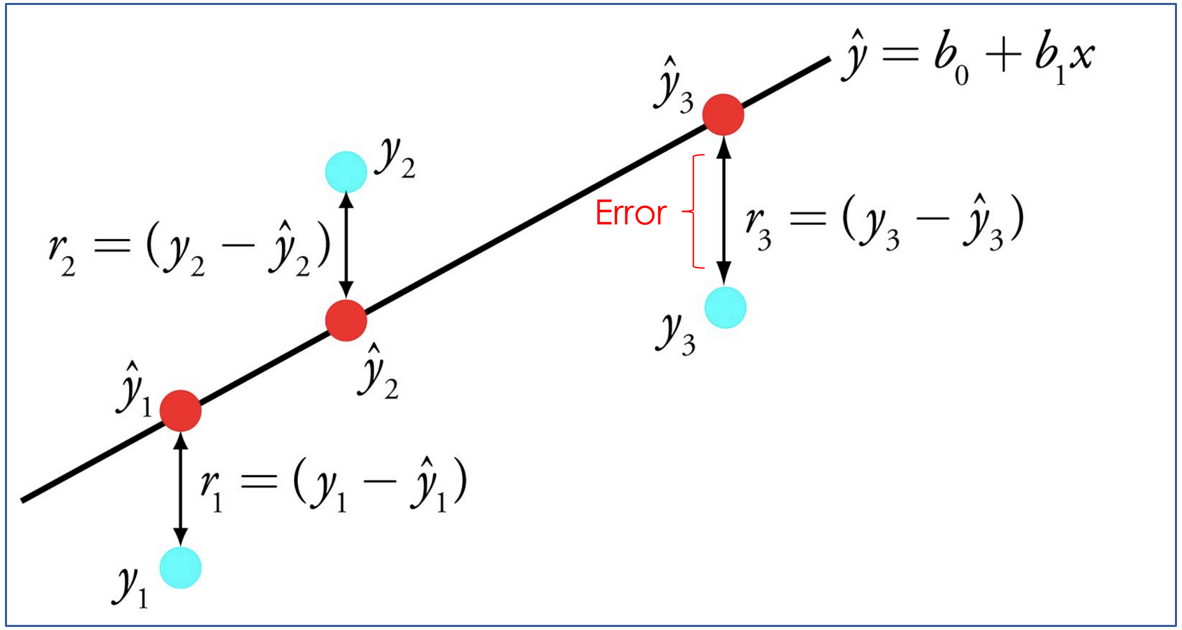
- ***a*** es el intercepto (corta el eje Y).
- ***b₁, b₂,… , bₙ*** son valores de la pendiente.



# Residuo

# Residuo

- Al ajustar los datos en el hiperplano (recta) que "pasa por" los puntos, puede existir una diferencia entre el punto pronosticado y la observación real.
- A esto se conoce como **residuo**.



# Residuo

La diferencia entre $Y - \hat{Y}$ es el error estimado.

- $Y$ *es el* valor real.
- $\hat{Y}$ es el valor pronosticado: $\hat{Y}_i = a + bX_i$

$$Suma\ del\ error\ cuadrático = SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

A partir de ***SSE*** se calcula el **residuo** (**SEE,** error estandar), que mide la dispersión de los valores observados alrededor de la línea de regresión:

$$Residuo = Error\ residual = \underbrace{SEE}_{\downarrow} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

$$Y = a + b_iX_i + u$$

# Bondad de ajuste

# Bondad de ajuste

- La bondad de ajuste, conocido también como coeficiente de determinación ($R^2$ o R2 Score), se utiliza para medir la precisión del modelo de regresión.
- Indica qué tan cerca están los datos de la línea de regresión ajustada. Representa un porcentaje entre 0 y 1.

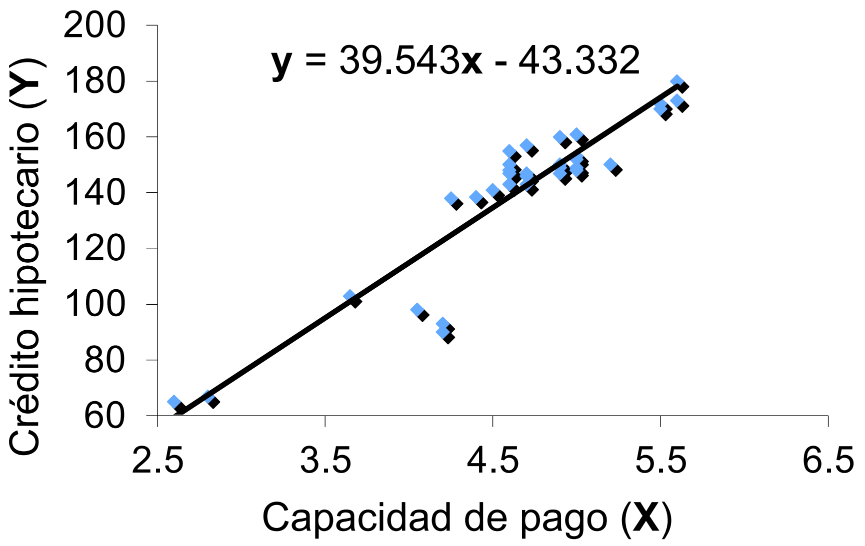
$$\bar{R}^2 = 1 - \frac{N-1}{\underbrace{N-k-1}_{}}[1 - R^2]$$

Número de variables independientes

- 0 cuando las variables son independientes.
- 1 cuando las variables tienen una relación perfecta.

**Por ejemplo,** Si $R^2$ es 0.87, indica:

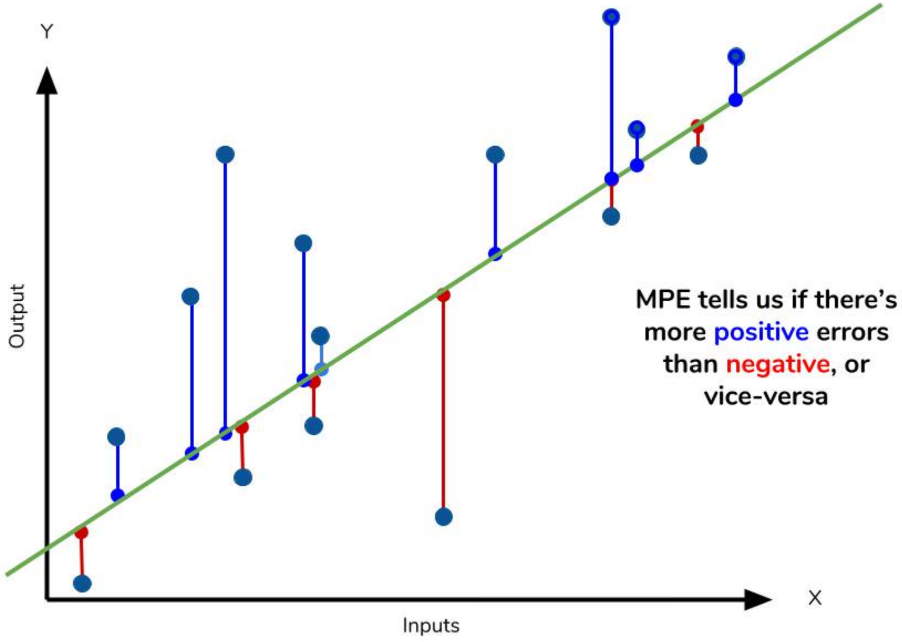
- Si se conoce la **capacidad de pago** (ingresos), entonces se puede lograr, con un 87%, el pronóstico de algún **crédito hipotecario**.



# Bondad de ajuste

**Limitaciones de R-cuadrado**

- Con R-cuadrado no se puede determinar si las estimaciones de los coeficientes están sesgadas.
- R cuadrado puede tener un valor bajo para un buen modelo, o alto para otro que no se ajusta a los datos.



# Ejemplo 1

# Ejemplo

Sean dos variables: Meses del año y Cantidad (coches vendidos)

| Nro | Mes_X | Cantidad_Y |
|---|---|---|
| 1 | 1 | 132 |
| 2 | 2 | 170 |
| 3 | 3 | 230 |
| 4 | 4 | 205 |
| 5 | 5 | 260 |
| 6 | 6 | 302 |
| 7 | 7 | 330 |
| 8 | 8 | 360 |
| 9 | 9 | 480 |
| 10 | 10 | 440 |
| 11 | 11 | 485 |
| 12 | 12 | 540 |
| 13 | | |
| 14 | | |
| 15 | | |
| | 6.50 | 327.83 |

$$Y_i = \boxed{a} + \boxed{b}X_i + \boxed{u}$$

**Variable a predecir**
Variable dependiente
Clase

**Intercepto**
Punto de corte
Alfa

**Pendiente**
Coeficiente

**Variable independiente**
Variable explicativa
Predictora

**Residuo**
Error

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

| Nro | Mes_X | Cantidad_Y | 1 x (varianza) | 2 x*Cantidad_Y | 3 $x^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 |
| 13 | | | | | |
| 14 | | | | | |
| 15 | | | | | |
| | **6.50** | **327.83** | | **5248** | **143** |

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

1 2 3 4

| Nro | Mes_X | Cantidad_Y | x (varianza) | x*Cantidad_Y | $x^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 |
| 13 | | | | | |
| 14 | | | | | |
| 15 | | | | | |
| | **6.50** | **327.83** | | **5248** | **143** |

**Pendiente (*b*)**

$$b = \frac{\mathrm{Cov(x,y)}}{\mathrm{S}_x^2}$$

$$b = \frac{\sum(x * \mathrm{Cantidad\_Y})}{\sum x^2}$$

$b = 5248 / 143$

$\boldsymbol{b = 36.7}$

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

| Nro | Mes_X | Cantidad_Y | 1 x (varianza) | 2 x*Cantidad_Y | 3 $x^2$ |
|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 |
| 13 | | | | | |
| 14 | | | | | |
| 15 | | | | | |
| | 6.50 | 327.83 | | 5248 | 143 |

4 **Pendiente: *b***

$b = 5248 / 143$

$\boldsymbol{b = 36.7}$

5 **Intercepto: *a***

$Y = a + bX$

$a = \bar{y} - b\bar{x}$

$a = 327.83 - (36.7 * 6.5)$

$a = 327.83 - 238.55$

$\boldsymbol{a = 89.29}$

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

| Nro | Mes_X | Cantidad_Y | x (varianza) | x*Cantidad_Y | $x^2$ | Ŷ (pronóstico) |
|---|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 | 125.99 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 | 162.69 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 | 199.39 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 | 236.09 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 | 272.78 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 | 309.48 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 | 346.18 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 | 382.88 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 | 419.58 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 | 456.28 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 | 492.98 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 | 529.68 |
| 13 | | | | | | |
| 14 | | | | | | |
| 15 | | | | | | |
| | **6.50** | **327.83** | | **5248** | **143** | |

6 **Pronóstico: Ŷ**

$$Y_i = a + bX_i$$

$\hat{Y}_1 = 89.29 + 36.7(1)$
$\hat{Y}_1 = 89.29 + 36.7$
$\boldsymbol{\hat{Y}_1 = 125.99}$

$\hat{Y}^2 = 89.29 + 36.69(2)$
$\hat{Y}_2 = 89.29 + 73.38$
$\boldsymbol{\hat{Y}_2 = 162.69}$

…

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

| Nro | Mes_X | Cantidad_Y | x (varianza) | x*Cantidad_Y | $x^2$ | Ŷ (pronóstico) | $(Y-\hat{Y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 | 125.99 | 36.15 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 | 162.69 | 53.49 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 | 199.39 | 937.23 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 | 236.09 | 966.28 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 | 272.78 | 163.44 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 | 309.48 | 56.01 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 | 346.18 | 261.89 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 | 382.88 | 523.60 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 | 419.58 | 3650.38 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 | 456.28 | 265.07 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 | 492.98 | 63.68 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 | 529.68 | 106.51 |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| | 6.50 | 327.83 | | 5248 | 143 | | 7083.74 |

7 **Error cuadrático**

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$SSE_1 = (132 - 125.99)^2$
$SSE_1 = (6.01)^2$
$\boldsymbol{SSE_1 = 36.15}$

$SSE_2 = (170 - 162.69)^2$
$SSE_2 = (7.31)^2$
$\boldsymbol{SSE_2 = 53.49}$

# Ejemplo

$$Y_i = a + bX_i + u$$

Solución

| Nro | Mes_X | Cantidad_Y | x (varianza) | x*Cantidad_Y | $x^2$ | Ŷ (pronóstico) | $(Y\text{-}\hat{Y})^2$ |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 132 | -5.50 | -726 | 30.25 | 125.99 | 36.15 |
| 2 | 2 | 170 | -4.50 | -765 | 20.25 | 162.69 | 53.49 |
| 3 | 3 | 230 | -3.50 | -805 | 12.25 | 199.39 | 937.23 |
| 4 | 4 | 205 | -2.50 | -513 | 6.25 | 236.09 | 966.28 |
| 5 | 5 | 260 | -1.50 | -390 | 2.25 | 272.78 | 163.44 |
| 6 | 6 | 302 | -0.50 | -151 | 0.25 | 309.48 | 56.01 |
| 7 | 7 | 330 | 0.50 | 165 | 0.25 | 346.18 | 261.89 |
| 8 | 8 | 360 | 1.50 | 540 | 2.25 | 382.88 | 523.60 |
| 9 | 9 | 480 | 2.50 | 1200 | 6.25 | 419.58 | 3650.38 |
| 10 | 10 | 440 | 3.50 | 1540 | 12.25 | 456.28 | 265.07 |
| 11 | 11 | 485 | 4.50 | 2183 | 20.25 | 492.98 | 63.68 |
| 12 | 12 | 540 | 5.50 | 2970 | 30.25 | 529.68 | 106.51 |
| 13 | | | | | | | |
| 14 | | | | | | | |
| 15 | | | | | | | |
| | 6.50 | 327.83 | | 5248 | 143 | | 7083.74 |

8 **Error residual**

$$SEE = \sqrt{\frac{SSE}{n-2}}$$

$$SEE = \sqrt{\frac{7083.74}{12-2}}$$

$$\boldsymbol{SEE = 26.62}$$

## Ejemplo

Función de estimación: $\boldsymbol{Y_i = a + bX_i + u}$ $\quad Y_i = 89.29 + 36.7X_i + 26.62$

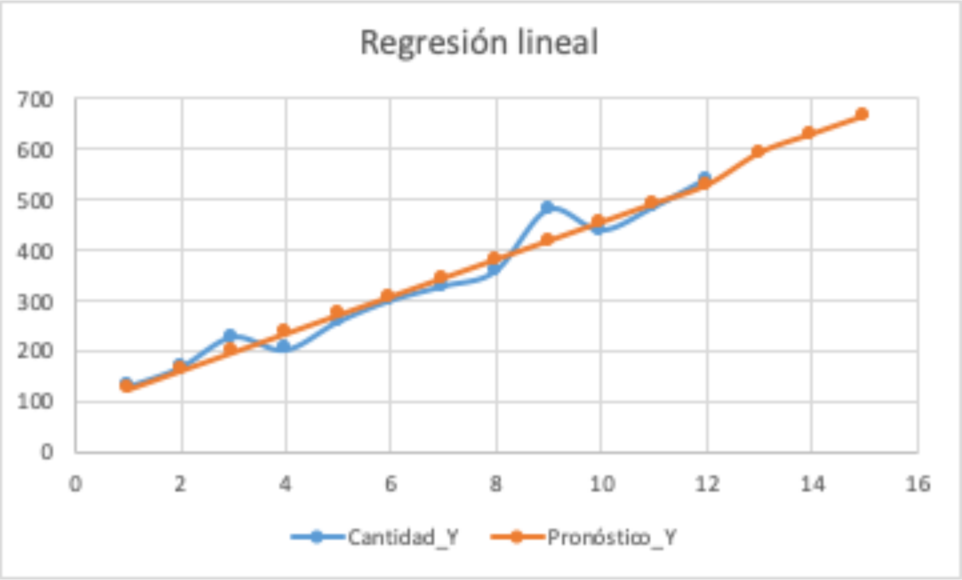
| Nro | Mes_X | Cantidad_Y |
|---|---|---|
| 1 | 1 | 132 |
| 2 | 2 | 170 |
| 3 | 3 | 230 |
| 4 | 4 | 205 |
| 5 | 5 | 260 |
| 6 | 6 | 302 |
| 7 | 7 | 330 |
| 8 | 8 | 360 |
| 9 | 9 | 480 |
| 10 | 10 | 440 |
| 11 | 11 | 485 |
| 12 | 12 | 540 |
| 13 | 13 | 593 |
| 14 | 14 | 630 |
| 15 | 15 | 666 |

**Pronóstico de nuevos casos: Ŷ**

$\hat{Y}_{13} = 89.29 + 36.7(13) + 26.62 = 89.29 + 477.1 + 26.62 = 593.01$

$\hat{Y}_{14} = 89.29 + 36.7(14) + 26.62 = 89.29 + 513.8 + 26.62 = 629.71$

$\hat{Y}_{15} = 89.29 + 36.7(15) + 26.62 = 89.29 + 550.5 + 26.62 = 666.41$



# Ejemplo

## 9 Bondad de ajuste

$$SEE = \sqrt{\frac{SSE}{n-2}}$$

| | |
|---|---|
| Error estandar del residuo | 26.62 |
| Pendiente (b) | 36.70 |
| Intercepto (a) | 89.29 |
| R (coef. Correl) | 0.98 |
| R cuadrado | 0.96 |
| Total Observaciones | 12 |
| R cuadrado ajustado | 0.96 |

Cantidad_Y = $\boxed{89.29}$ $\boxed{+\ 36.7}$Mes_X + $\boxed{26.62}$

$$Y_i = \boxed{a} + \boxed{bX_i} + \boxed{u_i}$$

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1-R^2]$$

Número de variables independientes

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1-R^2]$$

$$\bar{R}^2 = 1 - \frac{12-1}{12-1-1}[1-0.96]$$

$$\bar{R}^2 = 1 - 0.044$$

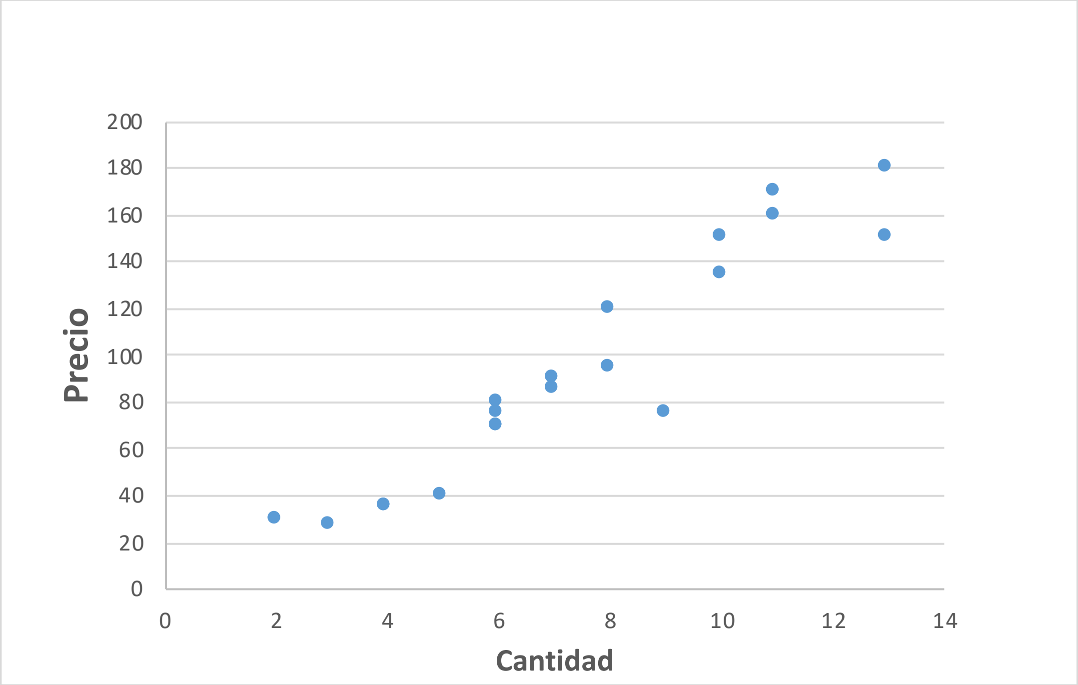
$$\bar{R}^2 = 0.956$$

# Ejemplo 2

# Ejemplo 2

Sean dos variables: Precio y Cantidad (productos)

| Nro | Precio | Cantidad |
|---|---|---|
| 1 | 75 | 9 |
| 2 | 170 | 11 |
| 3 | 70 | 6 |
| 4 | 35 | 4 |
| 5 | 85 | 7 |
| 6 | 120 | 8 |
| 7 | 160 | 11 |
| 8 | 70 | 6 |
| 9 | 40 | 5 |
| 10 | 80 | 6 |
| 11 | 150 | 10 |
| 12 | 95 | 8 |
| 13 | 30 | 2 |
| 14 | 27 | 3 |
| 15 | 90 | 7 |
| 16 | 180 | 13 |
| 17 | 135 | 10 |
| 18 | 35 | 4 |
| 19 | 75 | 6 |
| 20 | 150 | 13 |
| | **93.60** | **7.45** |



# Ejemplo 2

## Solución

| Nro | Precio | Cantidad | vP | vP*Cantidad | $vP^2$ | Ŷ (Pronóstico) |
|---|---|---|---|---|---|---|
| 1 | 75 | 9 | -18.60 | -167 | 345.96 | 6.32 |
| 2 | 170 | 11 | 76.40 | 840 | 5836.96 | 12.08 |
| 3 | 70 | 6 | -23.60 | -142 | 556.96 | 6.02 |
| 4 | 35 | 4 | -58.60 | -234 | 3433.96 | 3.90 |
| 5 | 85 | 7 | -8.60 | -60 | 73.96 | 6.93 |
| 6 | 120 | 8 | 26.40 | 211 | 696.96 | 9.05 |
| 7 | 160 | 11 | 66.40 | 730 | 4408.96 | 11.47 |
| 8 | 70 | 6 | -23.60 | -142 | 556.96 | 6.02 |
| 9 | 40 | 5 | -53.60 | -268 | 2872.96 | 4.20 |
| 10 | 80 | 6 | -13.60 | -82 | 184.96 | 6.63 |
| 11 | 150 | 10 | 56.40 | 564 | 3180.96 | 10.87 |
| 12 | 95 | 8 | 1.40 | 11 | 1.96 | 7.53 |
| 13 | 30 | 2 | -63.60 | -127 | 4044.96 | 3.60 |
| 14 | 27 | 3 | -66.60 | -200 | 4435.56 | 3.41 |
| 15 | 90 | 7 | -3.60 | -25 | 12.96 | 7.23 |
| 16 | 180 | 13 | 86.40 | 1123 | 7464.96 | 12.68 |
| 17 | 135 | 10 | 41.40 | 414 | 1713.96 | 9.96 |
| 18 | 35 | 4 | -58.60 | -234 | 3433.96 | 3.90 |
| 19 | 75 | 6 | -18.60 | -112 | 345.96 | 6.32 |
| 20 | 150 | 13 | 56.40 | 733.20 | 3180.96 | 10.87 |
| | **93.60** | **7.45** | | **2835** | **46785** | |

1 2 3

4 **Pendiente: *b***

*b* = Suma(vP*Cantidad) / Suma($vP^2$)
*b* = 2835 /46785
***b* = 0.06**

5 **Intercepto: *a***

*a* = Prom(Cantidad) - (***b****Prom(Precio))
*a* = 7.45 – (0.06*93.6)
*a* = 7.45 – 5.616
***a* = 1.78**

6 **Pronóstico: Ŷ**

$\hat{Y}_1$ = 1.78 + 0.06 (75)
$\hat{Y}_1$ = 1.78 + 4.5
$\hat{Y}_1$ **= 6.32**

# Ejemplo 2

## Solución

| Nro | Precio | Cantidad | Ŷ (Pronóstico) | $(Y-\hat{Y})^2$ |
|---|---|---|---|---|
| 1 | 75 | 9 | 6.32 | 7.17 |
| 2 | 170 | 11 | 12.08 | 1.16 |
| 3 | 70 | 6 | 6.02 | 0.00 |
| 4 | 35 | 4 | 3.90 | 0.01 |
| 5 | 85 | 7 | 6.93 | 0.01 |
| 6 | 120 | 8 | 9.05 | 1.10 |
| 7 | 160 | 11 | 11.47 | 0.22 |
| 8 | 70 | 6 | 6.02 | 0.00 |
| 9 | 40 | 5 | 4.20 | 0.64 |
| 10 | 80 | 6 | 6.63 | 0.39 |
| 11 | 150 | 10 | 10.87 | 0.75 |
| 12 | 95 | 8 | 7.53 | 0.22 |
| 13 | 30 | 2 | 3.60 | 2.55 |
| 14 | 27 | 3 | 3.41 | 0.17 |
| 15 | 90 | 7 | 7.23 | 0.05 |
| 16 | 180 | 13 | 12.68 | 0.10 |
| 17 | 135 | 10 | 9.96 | 0.00 |
| 18 | 35 | 4 | 3.90 | 0.01 |
| 19 | 75 | 6 | 6.32 | 0.10 |
| 20 | 150 | 13 | 10.87 | 4.55 |
| | **93.60** | **7.45** | | **19.21** |

7 **Error cuadrático: $(Y-\hat{Y})^2$**

$SSE = (\text{Cantidad} - \hat{Y})^2$

$SSE_1 = (9 - 6.32)^2$

$\boldsymbol{SSE_1 = 7.17}$

…

$$SSE = \sum_{i=1}^{n} [Y_i - \hat{Y}_i]^2$$

***SSE* = 19.21**

8 **Error residual**

$$SEE = \sqrt{\frac{SSE}{n-2}}$$

$$SEE = \sqrt{\frac{19.21}{20-2}} = \sqrt{\frac{19.21}{18}} = \mathbf{1.03}$$

# Ejemplo 2

Función de estimación: $\boldsymbol{Y_i = 1.78 + 0.06X_i + 1.03}$

| *Nro* | Precio | Cantidad |
|---|---|---|
| *21* | 82 | 8 |
| *22* | 102 | 9 |
| *23* | 162 | 13 |
| *24* | 48 | 6 |
| *25* | 64 | 7 |

**Pronóstico de nuevos casos: $\hat{Y}$**

$\hat{Y}_{21}$ = 1.78 + 0.06(82) + 1.03 = 7.73 = 8

$\hat{Y}_{22}$ = 1.78 + 0.06(102) + 1.03 = 8.93 = 9

$\hat{Y}_{23}$ = 1.78 + 0.06(162) + 1.03 = 12.53 = 13

$\hat{Y}_{24}$ = 1.78 + 0.06(48) + 1.03 = 5.69 = 6

$\hat{Y}_{25}$ = 1.78 + 0.06(64) + 1.03 = 6.65 = 7

# Ejemplo 2

## 9 Bondad de ajuste

$$SEE = \sqrt{\frac{SSE}{n-2}}$$

| | |
|---|---|
| ***Error estandar del residuo*** | **1.03** |
| ***Pendiente (b)*** | **0.06** |
| ***Intercepto (a)*** | **1.78** |
| *R (coef. Correl)* | 0.95 |
| *R cuadrado* | 0.90 |
| *Total Observaciones* | 20 |
| *R cuadrado ajustado* | 0.89 |

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1-R^2]$$

*Número de variables independientes*

Cantidad = $\boxed{1.78}$ $\boxed{+\ 0.06}$ Precio $\boxed{+1.03}$

$$Y_i = \boxed{a} + \boxed{b}X_i + \boxed{u_i}$$

$$\bar{R}^2 = 1 - \frac{N-1}{N-k-1}[1-R^2]$$

$$\bar{R}^2 = 1 - \frac{20-1}{20-1-1}[1-0.90]$$

$$\bar{R}^2 = 1 - 0.1055$$

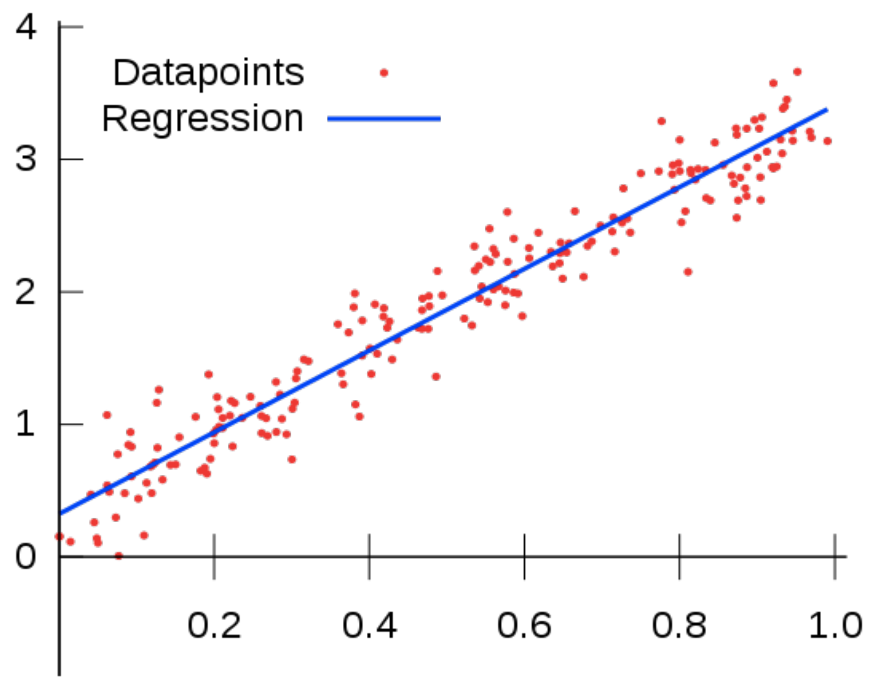
$$\bar{R}^2 = 0.89$$

# Consideraciones

La regresión lineal es un algoritmo útil. Sin embargo, obliga a ajustar los datos en forma de línea, que en ocasiones no encaja.

- Pronóstico de ventas.
- Pronóstico del consumo eléctrico.
- Pronóstico de la demanda de productos y servicios.
- Estimación de precios.
- Pronóstico de la bolsa de valores.
- Entre otros.

Cuando el problema incluye una no linealidad, entonces se necesita emplear otros algoritmos.



## Consideraciones

Una solución a la no linealidad eserías usar polinomios, pero sigue siendo un problema de regresión.

**Entonces, ¿qué tipo de funciones se pueden usar?**

Polinomial: $\hat{y} = a + b_1x + b_2x^2 + u$ $\qquad$ $\hat{y} = a + b_1x + b_2x^2 + b_3x^3 + u$

Exponencial: $\hat{y} = a + e^{bx}$

Logarítmico: $\hat{y} = a + b\log\mathrm{x}$

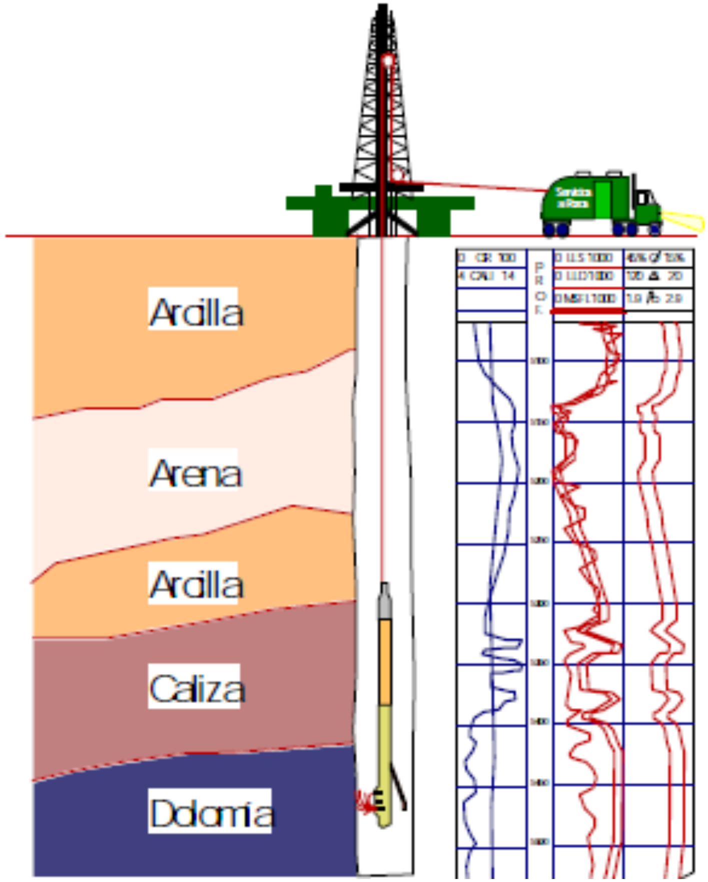
Sigmoide: $\hat{y} = \dfrac{1}{1 + e^{-(a+bX)}}$

Gaussiana: $\hat{y} = \dfrac{(x - \mu_j)}{2\sigma_j^2}$

# Caso práctico

# Caso de estudio

- Se tiene 4 mediciones de registros geofísicos: RC1, RC2, RC3 y RC4
- Se desea obtener el pronóstico (valores integrados) para estas 4 mediciones.




- RC1 = Registro Neutrón
- RC2 = Registro Sónico
- RC3 = Registro Densidad-Neutrón
- RC4 = Registro Densidad (corregido por arcilla)

- Para la toma de registros se cuenta con cables electromecánicos, sensores, dispositivos eléctricos y sistemas computarizados.
- Se procesan los datos a través de los sensores, para luego ser enviados a la superficie por medio del cable.

# Caso de estudio

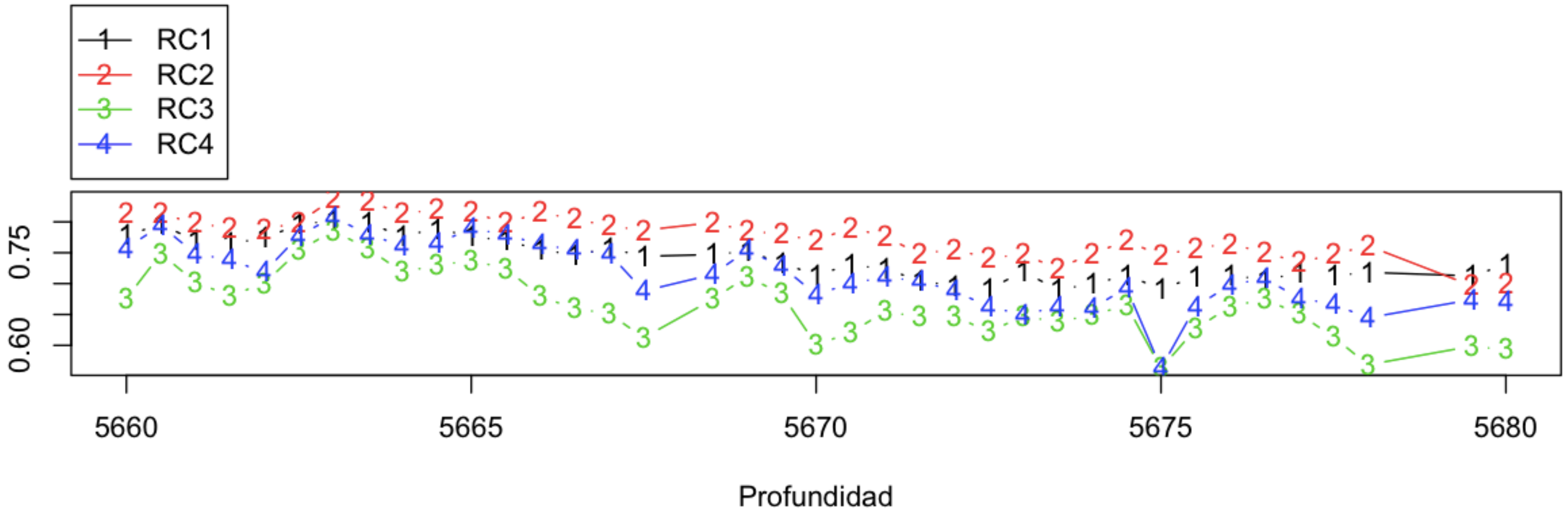
```
RegGeo <- read.table("/Users/guille/Documents/1 FI-UNAM/1 Cursos/2021-1/1 IA2021-1/2 CasosPracticos/5 RLineal/RGeofisicos.txt", header=TRUE, sep="\t")

RegGeo
```

| | Profundidad | RC1 | RC2 | RC3 | RC4 |
|---|---|---|---|---|---|
| 1 | 5660.0 | 0.7779242 | 0.8140291 | 0.6756982 | 0.7578417 |
| 2 | 5660.5 | 0.7962395 | 0.8131670 | 0.7486699 | 0.7938719 |
| 3 | 5661.0 | 0.7692308 | 0.7975620 | 0.7022851 | 0.7483621 |
| 4 | 5661.5 | 0.7647744 | 0.7903653 | 0.6802887 | 0.7384512 |
| 5 | 5662.0 | 0.7738128 | 0.7881841 | 0.7002481 | 0.7184618 |
| 6 | 5662.5 | 0.7956272 | 0.7988501 | 0.7534721 | 0.7775373 |
| 7 | 5663.0 | 0.8021554 | 0.8377173 | 0.7854410 | 0.8079569 |
| 8 | 5663.5 | 0.7978784 | 0.8338511 | 0.7568469 | 0.7796415 |
| 9 | 5664.0 | 0.7772059 | 0.8131166 | 0.7187127 | 0.7614542 |
| 10 | 5664.5 | 0.7886036 | 0.8200411 | 0.7295821 | 0.7656005 |
| 11 | 5665.0 | 0.7769245 | 0.8159172 | 0.7373500 | 0.7886884 |
| 12 | 5665.5 | 0.7690031 | 0.7979401 | 0.7247360 | 0.7796748 |
| 13 | 5666.0 | 0.7553049 | 0.8151496 | 0.6791887 | 0.7629721 |
| 14 | 5666.5 | 0.7460951 | 0.8047134 | 0.6596015 | 0.7546896 |
| 15 | 5667.0 | 0.7570502 | 0.7931801 | 0.6513744 | 0.7483805 |
| 16 | 5667.5 | 0.7441865 | 0.7864764 | 0.6124304 | 0.6880620 |

# Caso de estudio

- Variables independientes: Profundidad, RC1, RC2 y RC3
- Variable dependiente: RC4 (Registro Densidad -corregido por arcilla-)

```
library(Rcmdr)
with(RegGeo, lineplot(Profundidad, RC1, RC2, RC3, RC4))
```



# Caso de estudio

Regresion <- lm(RegGeo, formula = RC4 ~ Profundidad+RC1+RC2+RC3)

Regresion

```
Call:
lm(formula = RC4 ~ Profundidad + RC1 + RC2 + RC3, data = RegGeo)

Coefficients:
(Intercept)  Profundidad          RC1          RC2          RC3
  2.624e-01   -7.506e-05    5.066e-01    2.275e-01    4.891e-01
```

$$\widehat{\mathrm{RC}4} = a + b_1 \mathrm{Profundidad} + b_2 \mathrm{RC1} + b_3 \mathrm{RC2} + b_4 \mathrm{RC3} + u$$

# Caso de estudio

summary(Regresion)

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.624e-01  7.640e+00   0.034 0.972813
Profundidad -7.506e-05  1.320e-03  -0.057 0.954992
RC1          5.066e-01  2.296e-01   2.207 0.034397 *
RC2          2.275e-01  2.171e-01   1.048 0.302269
RC3          4.891e-01  1.145e-01   4.272 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02097 on 33 degrees of freedom
Multiple R-squared:  0.8581, Adjusted R-squared:  0.8409
F-statistic:  49.9 on 4 and 33 DF,  p-value: 1.539e-13
```

```
Coefficients:
               Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.26237022  7.64044786   0.034 0.972813
Profundidad -0.00007506  0.00131985  -0.057 0.954992
RC1          0.50661905  0.22957422   2.207 0.034397 *
RC2          0.22747126  0.21705964   1.048 0.302269
RC3          0.48909133  0.11449548   4.272 0.000155 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.02097 on 33 degrees of freedom
Multiple R-squared:  0.8581, Adjusted R-squared:  0.8409
F-statistic:  49.9 on 4 and 33 DF,  p-value: 1.539e-13
```

$$\widehat{\mathrm{RC}4} = a + b_1\mathrm{Profundidad} + b_2\mathrm{RC1} + b_3\mathrm{RC2} + b_4\mathrm{RC3} + u$$

$$\widehat{\mathrm{RC}4} = 0.2623 - 0.000075\mathrm{Profundidad} + 0.5066\mathrm{RC1} + 0.2274\mathrm{RC2} + 0.4890\mathrm{RC3} + 0.02097$$

# Caso de estudio

```
Pronostico <- predict(Regresion, type = "response", newdata = RegGeo[, c("Profundidad", "RC1", "RC2", "RC3")])
Pronostico
```

$$\widehat{RC}4 = 0.2623 - 0.000075\text{Profundidad} + 0.5066\text{RC1} + 0.2274\text{RC2} + 0.4890\text{RC3} + 0.02097$$

```
        1         2         3         4         5         6         7         8         9        10        11        12        13 
0.7472942 0.7920293 0.7520726 0.7373820 0.7511894 0.7906610 0.8184077 0.8013388 0.7674605 0.7800886 0.7769953 0.7626860 0.7373466 
       14        15        16        17        18        19        20        21        22        23        24        25        26 
0.7206894 0.7195546 0.6924280 0.7269750 0.7446542 0.7199351 0.6672274 0.6903277 0.7000343 0.6800535 0.6777213 0.6594727 0.6872664 
       27        28        29        30        31        32        33        34        35        36        37        38 
0.6635396 0.6779189 0.6983861 0.6312838 0.6751625 0.6957009 0.6951390 0.6845189 0.6673696 0.6512601 0.6482442 0.6553729 
```

# Caso de estudio

write.csv(cbind(RegGeo, Pronostico), file="/Users/guille/Documents/1 FI-UNAM/1 Cursos/2021-1/1 IA2021-1/2 CasosPracticos/5 RLineal/Resultados.csv", row.names=FALSE)

Resultados

| Profundidad | RC1 | RC2 | RC3 | RC4 | Pronostico |
|---|---|---|---|---|---|
| 5660 | 0.77792425 | 0.814029063 | 0.67569817 | 0.757841697 | 0.747294237159165 |
| 5660.5 | 0.796239499 | 0.81316698 | 0.748669946 | 0.793871855 | 0.792029326006498 |
| 5661 | 0.769230786 | 0.797562036 | 0.702285088 | 0.748362073 | 0.752072559621215 |
| 5661.5 | 0.76477436 | 0.790365252 | 0.680288728 | 0.738451206 | 0.737382029266921 |
| 5662 | 0.773812818 | 0.788184093 | 0.700248127 | 0.718461836 | 0.751189372948178 |
| 5662.5 | 0.79562719 | 0.798850051 | 0.753472127 | 0.777537338 | 0.79066101602287 |
| 5663 | 0.802155394 | 0.837717345 | 0.785440951 | 0.807956871 | 0.818407666080092 |
| 5663.5 | 0.797878424 | 0.833851101 | 0.756846932 | 0.779641454 | 0.801338795826379 |
| 5664 | 0.777205898 | 0.813116563 | 0.718712688 | 0.761454166 | 0.767460531117542 |
| 5664.5 | 0.788603618 | 0.820041076 | 0.729582106 | 0.765600467 | 0.780088569588179 |
| 5665 | 0.776924464 | 0.815917152 | 0.737349992 | 0.788688434 | 0.776995289745656 |
| 5665.5 | 0.769003107 | 0.797940103 | 0.724735975 | 0.77967479 | 0.762685981567906 |
| 5666 | 0.755304855 | 0.815149612 | 0.679188716 | 0.762972062 | 0.737346555588098 |

# Caso de estudio

with(RegGeo, plot(Profundidad, Pronostico, pch=18, col=3, type="b"))