

# Metodología de Investigación

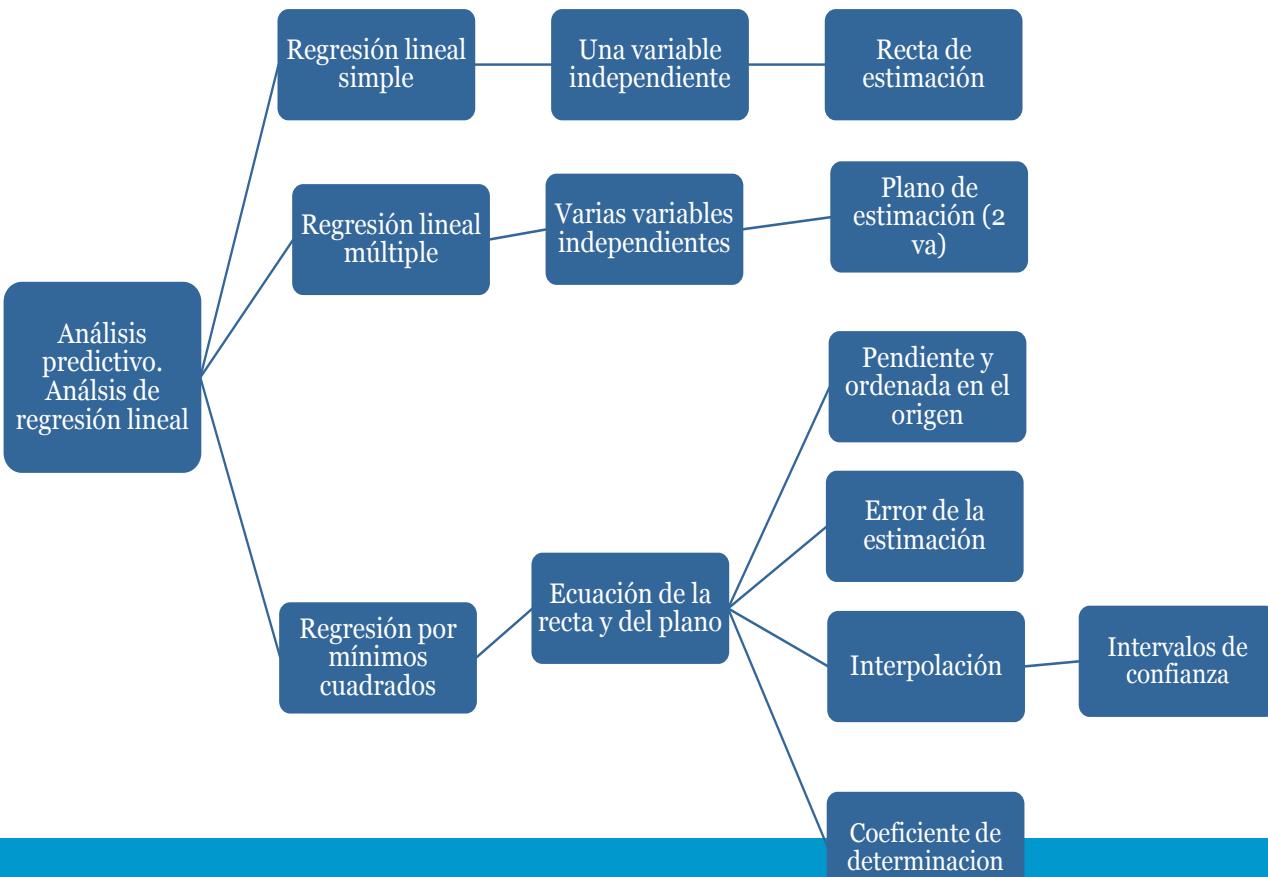
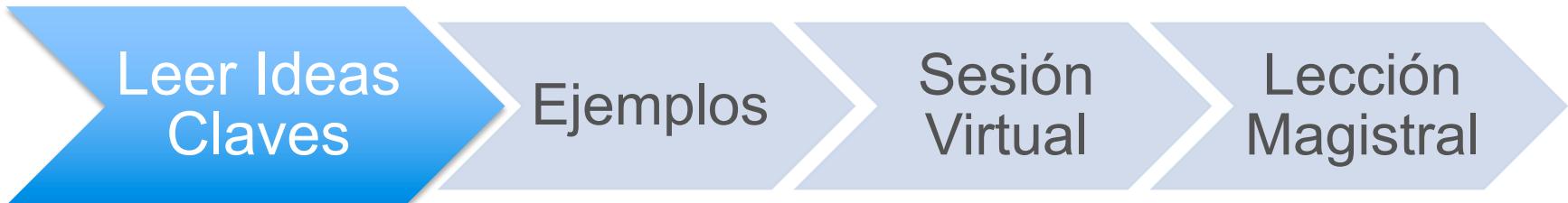
Máster Universitario en Ingeniería Matemática y Computación  
Nadia Gámez

## Tema 8

# Tema 8: Análisis predictivo. Análisis de regresión lineal

- 8.1. ¿Cómo estudiar este tema?
- 8.2. Ajuste de la recta
- 8.3. La regresión por mínimos cuadrados
- 8.4. Otros tipos de regresión
- 8.5. Tipos de residuos

# 8.1 ¿Cómo estudiar este tema?



## 8.2. Ajuste de la recta

### Introducción

- Permite conocer el valor de una variable desconocida a partir de datos de otra variable con la que se asocia en varias observaciones
- **Análisis predictivo**
- **Ecuación de estimación** que relaciona las variables
- **Análisis de correlación** se puede saber el grado en el que se asocian
  - Variables conocidas → independientes
  - Variables a predecir → dependiente
- Tipos : **Regresión Lineal Simple** y Regresión Múltiple

## 8.2. Ajuste de la recta

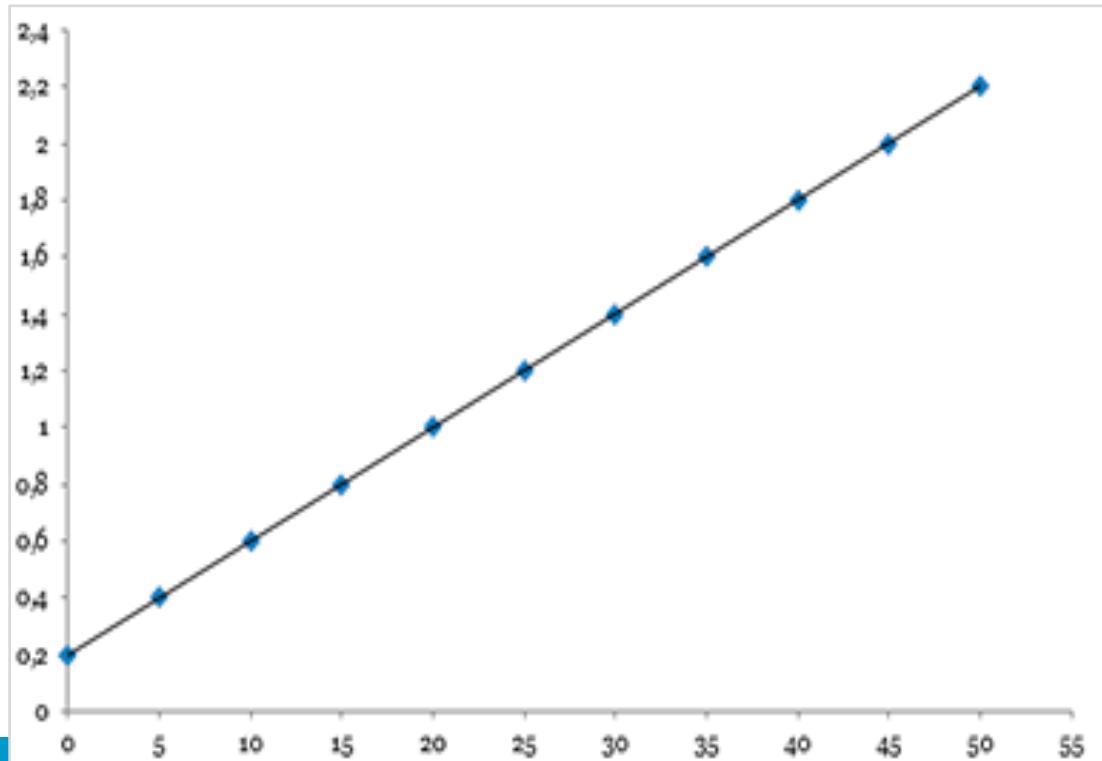
### Recta de regresión

- La recta puede expresarse mediante la ecuación en forma explícita
- $y = a + bx$  donde  $x$  e  $y$  son las variables independiente y dependiente respectivamente,  $b$  la pendiente y  $a$  la ordenada en el origen
- **análisis predictivo** puesto que para cualquier valor de la variable independiente se puede obtener el valor de la dependiente
- **recta de ajuste a la nube de puntos** y es posible obtener su ecuación a partir del diagrama cuando la asociación es perfecta

## 8.2. Ajuste de la recta

### Recta de regresión

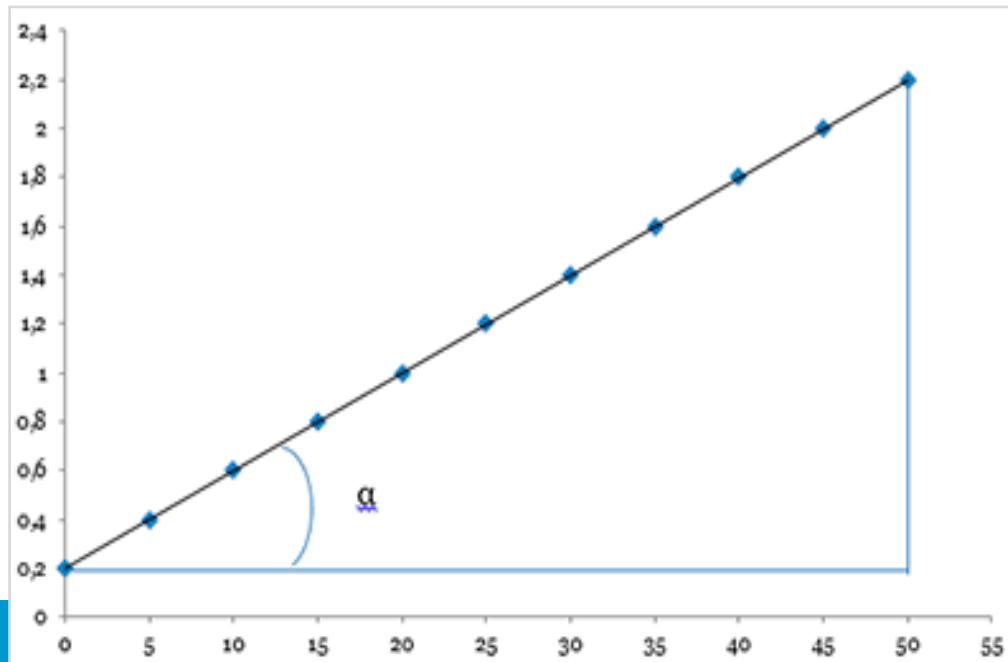
- A partir del diagrama se puede obtener el **valor de la ordenada en el origen**, corte con el eje vertical, es decir **cuando  $x=0$** , en este caso, tal valor es 0,2 en unidades de la variable dependiente



## 8.2. Ajuste de la recta

### Recta de regresión

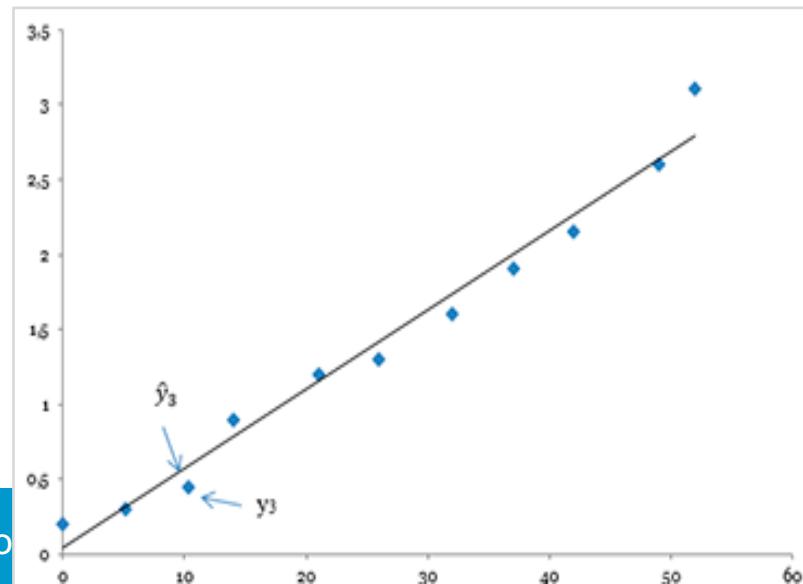
- La **pendiente** se calcula como la **tangente** del ángulo que la recta de ajuste forma con el eje horizontal
- $b = \operatorname{tg} \alpha = \frac{\text{lado opuesto}}{\text{lado contiguo}}$
- $b = \frac{y_{j+n}-y_j}{x_{i+n}-x_i}$   $x_i$  e  $y_j$  han de ser las coordenadas de un mismo punto



## 8.3. La regresión por mínimos cuadrados

### Obtención de la recta de regresión

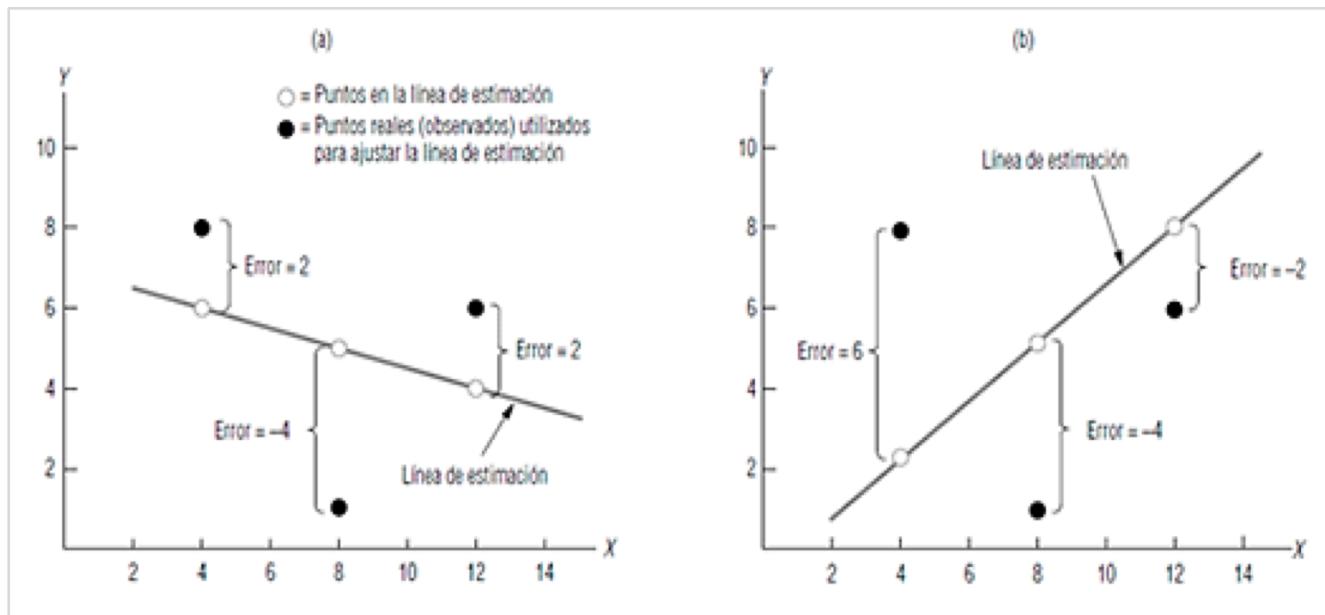
- Ajustar los puntos de un diagrama de dispersión → **método de mínimos cuadrados**
- Encontrar una recta que **minimice el error** entre los puntos dispersado (mejor ajuste entre dos variables)



## 8.3. La regresión por mínimos cuadrados

### Obtención de la recta de regresión

- **Residuos** → diferencias entre los valores observados y los estimados por la recta
- a)  $y_1 - \hat{y}_1 = 8 - 6 = 2$     $y_2 - \hat{y}_2 = 1 - 5 = -4$     $y_3 - \hat{y}_3 = 6 - 4 = 2$
- b)  $y_1 - \hat{y}_1 = 8 - 2 = 6$     $y_2 - \hat{y}_2 = 1 - 5 = -4$     $y_3 - \hat{y}_3 = 6 - 8 = -2$



## 8.3. La regresión por mínimos cuadrados

### Regresión por mínimos cuadrados

- El **error total** no se puede calcular así :
  - Sumando los residuos porque se anulan
  - Sumando los valores absolutos de los residuos porque no considera puntos muy alejados.
- Solución, **elevar al cuadrado los residuos** → método de mínimos cuadrados
- Hallar la **recta que minimice los cuadrados de los residuos**
- $\sum_1^n (y_j - \hat{y}_j)^2$  Mínimo →  $\sum_1^n (y_j - a - bx_i)^2$  Mínimo
- $\sum_1^n y_j = na + b \sum_1^n x_i$   $a = \bar{y} - b\bar{x}$
- $\sum_1^n y_j x_i = (\bar{y} - b\bar{x}) \sum_1^n x_i + b \sum_1^n x_i^2$
- $y - \bar{y} = \frac{s_{xy}}{s_x^2} (x - \bar{x})$  Donde  $\bar{x}$  e  $\bar{y}$  son los valores medios de x y de y respectivamente,  $S_{xy}$  es la covarianza y  $s_x^2$  es la varianza de x.

## 8.3. La regresión por mínimos cuadrados

### Ejemplo 2. Ecuación de la recta por mínimos cuadrados

Un jugador de baloncesto ha encestado las siguientes canastas desde diferentes distancias, escribir la recta de regresión obtenida por el método de mínimos cuadrados.

| Distancia | Canastas |
|-----------|----------|
| 2,5       | 23       |
| 3,2       | 21       |
| 4,7       | 19       |
| 6,2       | 15       |
| 8,7       | 12       |
| 13,2      | 8        |
| 17        | 4        |
| 18,5      | 1        |

| x                | $x - \bar{x}$ | $(x - \bar{x})^2$ | y                  | $y - \bar{y}$ | $(x - \bar{x})(y - \bar{y})$ |
|------------------|---------------|-------------------|--------------------|---------------|------------------------------|
| 2,50             | -6,75         | 45,56             | 23                 | 10,125        | -68,34                       |
| 3,20             | -6,05         | 36,60             | 21                 | 8,125         | -49,16                       |
| 4,70             | -4,55         | 20,70             | 19                 | 6,125         | -27,87                       |
| 6,20             | -3,05         | 9,30              | 15                 | 2,125         | -6,48                        |
| 8,70             | -0,55         | 0,30              | 12                 | -0,875        | 0,48                         |
| 13,20            | 3,95          | 15,60             | 8                  | -4,875        | -19,26                       |
| 17,00            | 7,75          | 60,06             | 4                  | -8,875        | -68,78                       |
| 18,50            | 9,25          | 85,56             | 1                  | -11,875       | -109,84                      |
| $\bar{x} = 9,25$ |               | 273,70            | $\bar{y} = 12,875$ |               | -349,25                      |

La covarianza  $S_{xy}$  es:  $S_{xy} = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{8-1} \cdot (-349,25) = -49,89$

La varianza de  $x$  es:  $S_x^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{273,70}{8-1} = 39,1$

Sustituyendo en la ecuación de la recta se obtiene:

$$y - 12,87 = \frac{-49,89}{39,1} (x - 9,25)$$

Operando se saca la ecuación de la recta:

$$y = -1,28x + 24,68$$

Debido a que la covarianza es negativa la relación entre las variables es inversa y la pendiente es negativa.

## 8.3. La regresión por mínimos cuadrados

### Mínimos cuadrados con la calculadora

**1.** Borrar los datos memorizados:

Pulsar *SHIFT + CLR + tecla 1 (SCL)* = o tecla 3 (*All*).

**2.** Regresión lineal

Pulse *MODE* dos veces. Pulsar 2 (*REG*) y 1 (*lin*).

**3.** Introducir los datos:

Escribir el dato  $x_1$  y pulsar “,” meter el dato  $y_1$  y pulsar *M+ (data)*. Repetir para los demás datos. En *REPLAY* se pueden comprobar los datos y las frecuencias.

**4.** Obtener el valor de a y b:

Pulsar *SHIFT + S-VAR* (tecla del número 2) + *REPLAY* dos veces derecha. Ordenada en el origen tecla 1, pendiente tecla 2.

## 8.3. La regresión por mínimos cuadrados

### Error Estándar de Estimación

- Ver si la **estimación es buena**
- Mediante el **error estándar** de la estimación
- **Significado** similar a la desviación estándar que evalúa lo que los datos se dispersan de la recta de regresión obtenida
- Su cálculo se lleva a cabo así:
- $S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-2}}$  **donde**  $y - \hat{y}$  son los residuos de la variable dependiente  $y$  y  $n$  el total de puntos.
- Cuando se eleva el error de la estimación al cuadrado se obtiene el **error cuadrático medio**,  $S_e^2$

## 8.3. La regresión por mínimos cuadrados

### Coeficiente de determinación

- Al hacer un ajuste por regresión lineal se puede calcular de un **coeficiente que nos dé información sobre el grado de asociación** entre las variables
- $r^2 = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$
- $r^2$  siempre es positivo y 1 cuando se trata de correlación perfecta entre variables.
- Tal caso de correlación perfecta será cuando sea cero, es decir, cuando los valores observados caigan sobre la recta obtenidas

## 8.3. La regresión por mínimos cuadrados

### Ejemplo 5. Coeficiente de determinación

La siguiente tabla recoge las ganancias obtenidas por un hotel en función de los huéspedes de lunes a domingo excluyendo el miércoles.

| Día de la semana | Ganancias | Número de huéspedes |
|------------------|-----------|---------------------|
| Lunes            | 1200      | 12                  |
| Martes           | 1500      | 16                  |
| Jueves           | 3670      | 31                  |
| Viernes          | 7200      | 69                  |
| Sábado           | 10090     | 102                 |
| Domingo          | 5800      | 56                  |

**A.** Escribir la recta de regresión y el coeficiente de correlación.

| x                | $x - \bar{x}$ | $(x - \bar{x})^2$ | y                | $y - \bar{y}$ | $(x - \bar{x})(y - \bar{y})$ |
|------------------|---------------|-------------------|------------------|---------------|------------------------------|
| 12               | -35,7         | 1272,3            | 1200             | -3710         | 132335,7                     |
| 16               | -31,7         | 1003              | 1500             | -3410         | 107994,7                     |
| 31               | -16,7         | 277,9             | 3670             | -1240         | 20670,8                      |
| 69               | 21,3          | 455               | 7200             | 2290          | 48845,7                      |
| 102              | 54,3          | 2951,7            | 10090            | 5180          | 281429,4                     |
| 56               | 8,3           | 69,4              | 5800             | 890           | 7413,7                       |
| $\bar{x} = 47,7$ |               | 6029,3            | $\bar{y} = 4910$ |               | 598690                       |

La covarianza  $S_{xy}$  es:

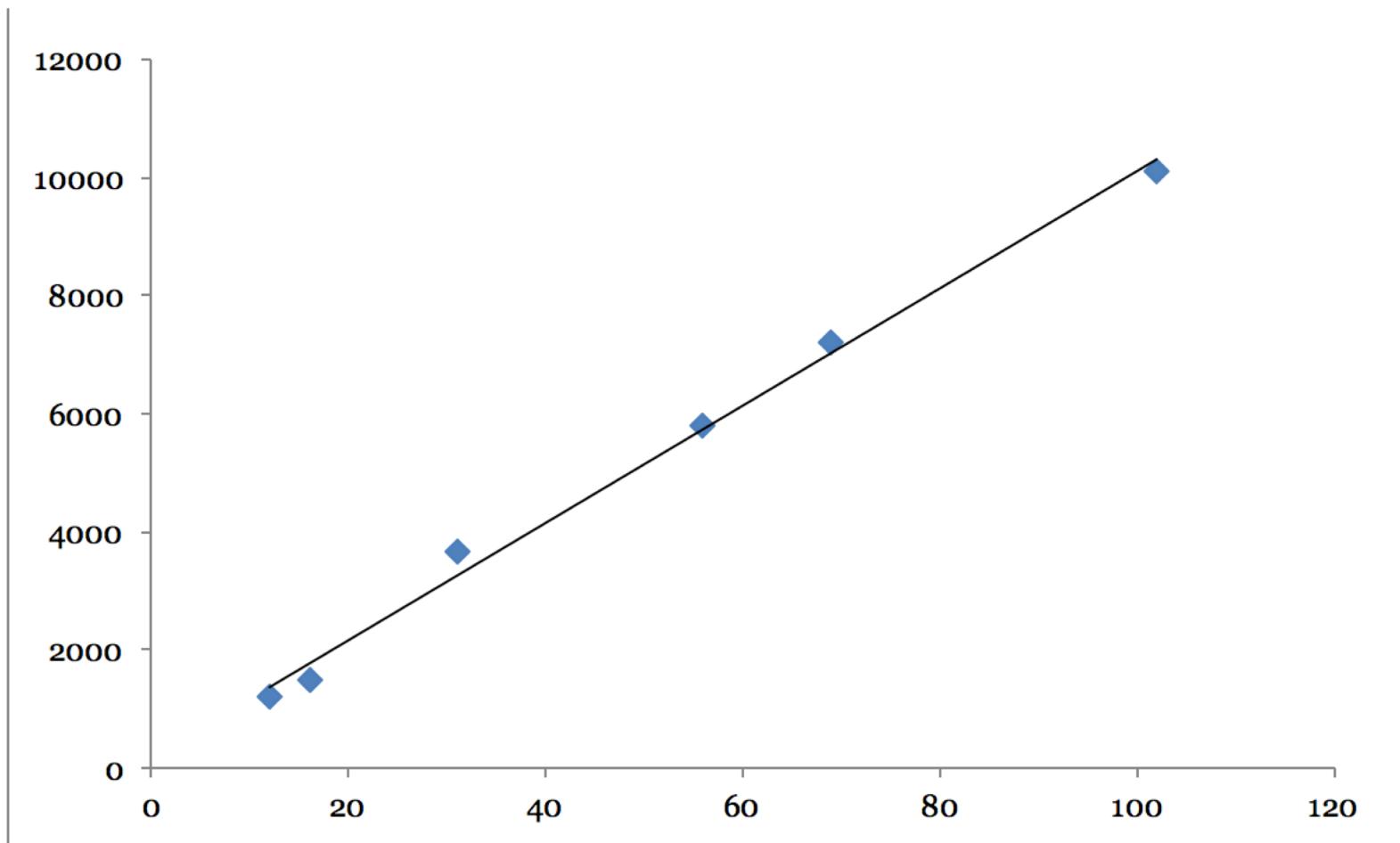
$$S_{xy} = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^h (x_i - \bar{x})(y_j - \bar{y}) = \frac{1}{6-1} \cdot (598690) = 119738$$

La varianza de x es:

$$S_x^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{6029,3}{6-1} = 1205,9$$

Sustituyendo:  $y - 4910 = \frac{119738}{1205,9} (x - 47,7)$

Operando se saca la ecuación de la recta:  $y = 99,30x + 176,9$



Para calcular el coeficiente de determinación hacemos la siguiente tabla:

| x   | y     | $\hat{y} = 99,30x + 176,9$ | $y - \hat{y}$ | $(y - \hat{y})^2$ | $y - \bar{y}$ | $(y - \bar{y})^2$ |
|-----|-------|----------------------------|---------------|-------------------|---------------|-------------------|
| 12  | 1200  | 1365,5                     | -165,5        | 27390,2           | -3710         | 13764100          |
| 16  | 1500  | 1765,7                     | -265,7        | 70596,5           | -3410         | 11628100          |
| 31  | 3670  | 3255,2                     | 414,8         | 172059            | -1240         | 1537600           |
| 69  | 7200  | 7028,6                     | 171,4         | 29378             | 2290          | 5244100           |
| 102 | 10090 | 10305,5                    | -215,5        | 46440,2           | 5180          | 26832400          |
| 56  | 5800  | 5737,7                     | 62,3          | 3881,3            | 890           | 792100            |
|     |       |                            |               | 345864            |               | 59798400          |

$$r^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = 1 - \frac{345864}{59798400} = 0,9942$$

El coeficiente de correlación es:

$$r = \sqrt{r^2} = 0,997$$

**B.** Calcular el error estándar de la estimación.

$$S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{345864}{6 - 2}} = 294,05$$

**C.** ¿Entre qué valores hubiera oscilado la ganancia si el miércoles hubiese habido 25 huéspedes a un nivel de significancia del 99,7%?

Interpolando en la recta para  $x=25$  huéspedes se tiene un valor estimado de  $y$  de 2659 euros de ganancia. A un nivel de significancia del 99,7%:

$$\hat{y} - 3 \cdot S_e = 2659 - 3 \cdot 294,05 = 1776,85$$

$$\hat{y} + 3 \cdot S_e = 2659 + 3 \cdot 294,05 = 3541,15$$

El 99,7 % de las ganancias estarán entre 1776,85 y 3541,15 Euros.

## 8.4. Otros tipos de Regresión por mínimos cuadrados

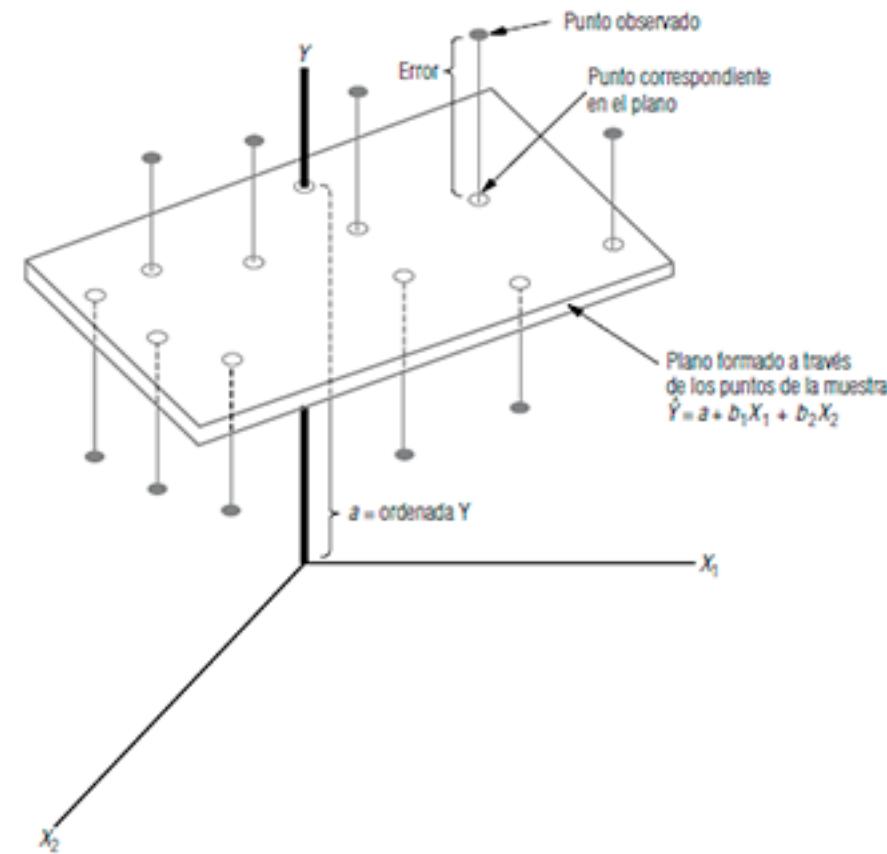
### Regresión múltiple

- cuando se emplea **más de una variable independiente** para estimar una variable dependiente
- **Ventaja:** permite usar más información disponible
- La ecuación de estimación con dos variables independientes:
- $\hat{y} = a + b_1x_1 + b_2x_2$  donde  $x_1$  y  $x_2$  son los valores de cada uno de las dos variables independientes y  $b_1$  y  $b_2$  los coeficientes de regresión.
- la ecuación de regresión múltiple es un **plano**

## 8.4. Otros tipos de Regresión por mínimos cuadrados

### Regresión múltiple

- cuando se tienen dos o más variables independientes y una variable dependiente
- Ventaja: se tienen más datos
- La ecuación general es la misma que para la regresión simple, pero las variables independientes son más de una
- $\hat{y} = a + b_1x_1 + b_2x_2 + \dots$  donde  $a$  es la ordenada Y y  $b_1$  y  $b_2$  los coeficientes de pendiente
- la ecuación general es:



describir la relación entre las variables  
disponibles  
los valores  
pendientes  
plano

## 8.4. Otros tipos de Regresión por mínimos cuadrados

### Regresión múltiple

- El error estándar de la estimación se puede calcular a partir de la siguiente fórmula:
  - $S_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n-k-1}}$  donde  $n$  es el número de puntos y  $k$  el número de variables independientes, en este caso dos.
- El coeficiente de determinación múltiple es la fracción de variación total de la variable y su fórmula es la misma que para la regresión lineal simple:
  - $R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2}$

## 8.4. Tipos de Residuos

---

Un residuo es la diferencia entre un valor observado y uno estimado:  $e = y - \hat{y}$

---

En el caso de los residuos ordinarios, se establece que la **media de los residuos es cero**:  $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i = 0$

---

La varianza se estima de manera aproximada:  $S^2 = \frac{1}{n-k-1} \sum_{i=1}^n (e_i - \bar{e})^2 = S_e^2$

---

Los **residuos se pueden estandarizar** de manera que tengan media cero y desviación estándar 1:  $d_i = \frac{e_i}{\sqrt{S_e^2}}$

---

Se puede comprobar que:  $S^2(e_i) = S^2(1 - h_{ii})$  donde  $h_{ii}$  es el i-ésimo elemento de la matriz proyección de los residuos

## 8.4. Tipos de Residuos

---

Se recomienda trabajar con residuos **estudentizados** de la siguiente manera:  $r_i = \frac{e_i}{\sqrt{Se^2(1-h_{ii})}}$

---

Además, en el caso de la regresión lineal simple se cumple que el valor  $h_{ii}$  es una medida para localizar el  $i$ -ésimo punto  $x_i$  respecto al punto medio:  $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$

---

En el caso de la regresión lineal múltiple el valor  $h_{ii}$  se calcula como:  $h_{ii} = \frac{1}{n} [1 + (x_i - \bar{x})S_{xx}(x_i - \bar{x})] = \frac{1}{n} (1 - D_i^2)$   $D_i$  es la distancia de Mahalanobis.

---

De esta manera, la varianza de un error  $e_i$  depende de la posición a la que se halle un valor  $x_i$ . La varianza de los puntos cercanos al valor medio es mayor que la de los alejados y por lo tanto supone un pobre ajuste por mínimos cuadrados.

UNIVERSIDAD  
INTERNACIONAL  
DE LA RIOJA

unir

[www.unir.net](http://www.unir.net)