# Estadística I
# Tema 3: Análisis de datos bivariantes

## Tema 3: Análisis de datos bivariantes

### Contenidos

# Tema 3: Análisis de datos bivariantes

3.3 Recta de regresión.

- Definición de la recta.
- Estimación de los coeficientes.
- Interpretación de los coeficientes.
- Valores predichos, residuos y varianza residual.
- Bondad del ajuste.
- Análisis de residuos.
- Los datos de *Anscombe*.

# Tema 3: Análisis de datos bivariantes

## Lecturas recomendadas

- Peña, D. y Romo, J., *Introducción a la Estadística para las Ciencias Sociales*.
  - Capítulos 7, 8 y 9.
- Newbold, P. *Estadística para los Negocios y la Economía*.
  - Secciones 2.5 y 12.1–12.4.

# Datos bivariantes

**Ejemplo** Nivel educativo ($X$) y situación laboral ($Y$) de 10 Madrileños.

Nivel educativo (1=Primaria o menos, 2=Secundaria, 3=Post-secundaria)
Situación laboral (1=Empleado, 2=Desempleado, 3=Inactivo)

| Individuo | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Nivel educativo ($X$) | 2 | 3 | 2 | 3 | 2 | 2 | 1 | 1 | 3 | 2 |
| Situación laboral ($Y$) | 3 | 1 | 1 | 3 | 3 | 3 | 3 | 3 | 1 | 3 |

# Datos bivariantes

- Datos bivariantes: provienen de la observación simultánea de dos variables $(X, Y)$ en una muestra de $n$ individuos. Los datos bivariantes son parejas de valores, numéricos o no, de la forma:
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$
- Se usan para describir las dos variables conjuntamente o una variable en función de la otra.
- A menudo se intenta describir el comportamiento de una de las variables, que se llama la variable dependiente y se denota por $Y$, en función de la otra variable, que se llama la variable independiente o explicativa, y se denota por $X$.

## Estructura de la tabla de doble entrada

- Representamos los valores de una de las variables (p.ej. $X$) en las cabeceras de las filas de una tabla, y los valores de la otra variable (p.ej. $Y$) en las cabeceras de las columnas de la tabla.
- En la casilla correspondiente a cada par de valores de $X$ e $Y$, se escribe la frecuencia absoluta o número de individuos en cada combinación de valores.
- Cuando al menos alguna de las dos variables es cualitativa, la tabla de doble entrada también se denomina tabla de contingencia.

## Estructura de la tabla de doble entrada/tabla de contingencia

**Ejemplo** Datos de 1508 madrileños (Encuesta de Condiciones de Vida). $X$: Nivel educativo, $Y$: Situación laboral

| | | $Y$ | | |
|---|---|---|---|---|
| | | Empleado | Desempleado | Inactivo |
| $X$ | Primaria | 95 | 6 | 315 |
| | Secundaria | 393 | 28 | 257 |
| | Post-secundaria | 317 | 8 | 89 |

- Se denomina distribución conjunta de $(X, Y)$ al conjunto formado por los valores observados en forma de pares, junto con las frecuencias absolutas correspondientes a cada par.

## Estructura de la tabla de doble entrada

- Tabla de doble entrada con $k$ filas y $m$ columnas

| | | $Y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $y_1$ | $\cdots$ | $y_j$ | $\cdots$ | $y_m$ | Total |
| $X$ | $x_1$ | $n_{11}$ | $\cdots$ | $n_{1j}$ | $\cdots$ | $n_{1m}$ | $n_{1\cdot}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_i$ | $n_{i1}$ | $\cdots$ | $n_{ij}$ | $\cdots$ | $n_{im}$ | $n_{i\cdot}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_k$ | $n_{k1}$ | $\cdots$ | $n_{kj}$ | $\cdots$ | $n_{km}$ | $n_{k\cdot}$ |
| | Total | $n_{\cdot 1}$ | $\cdots$ | $n_{\cdot j}$ | $\cdots$ | $n_{\cdot m}$ | $n_{\cdot\cdot}$ |

- Notación:

  $n_{ij}$ frecuencia absoluta en la casilla $(i, j)$
  Total de fila $i$: $n_{i\cdot} = n_{i1} + n_{i2} + \cdots + n_{im}$
  Total de columna $j$: $n_{\cdot j} = n_{1j} + n_{2j} + \cdots + n_{kj}$
  $n_{\cdot\cdot}$ tamaño muestral $n_{\cdot\cdot} = n$

## Distribuciones de frecuencias marginales

- Los totales de las filas (columnas) se llaman frecuencias absolutas marginales de las filas (columnas).
- Éstas son las frecuencias univariantes de $X$ y de $Y$.

**Ejemplo** Distribuciones de frecuencias marginales:

| Nivel educativo ($X$) | Primaria | Secundaria | Post-secundaria | Total |
|---|---|---|---|---|
| $n_{i\cdot}$ | 416 | 678 | 414 | 1508 |

| Situación laboral ($Y$) | Empleado | Desempleado | Inactivo | Total |
|---|---|---|---|---|
| $n_{\cdot j}$ | 805 | 42 | 661 | 1508 |

## Distribuciones de frecuencias marginales

- Se denomina distribución marginal de $X$ al conjunto de valores que toma $X$ junto con sus frecuencias absolutas marginales.
- Análogamente se define la distribución marginal de $Y$.
- Observación: Si en lugar de tener dos variables $(X, Y)$ tuviéramos tres $(X, Y, Z)$ tendríamos tres distribuciones marginales.

## Distribución conjunta de frecuencias relativas

- $f_{ij} = n_{ij}/n_{\cdot\cdot}$ frecuencia relativa en la casilla $(i, j)$

| | | $Y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $y_1$ | $\cdots$ | $y_j$ | $\cdots$ | $y_m$ | Total |
| | $x_1$ | $f_{11}$ | $\cdots$ | $f_{1j}$ | $\cdots$ | $f_{1m}$ | $f_{1\cdot}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| $X$ | $x_i$ | $f_{i1}$ | $\cdots$ | $f_{ij}$ | $\cdots$ | $f_{im}$ | $f_{i\cdot}$ |
| | $\vdots$ | $\vdots$ | | $\vdots$ | | $\vdots$ | $\vdots$ |
| | $x_k$ | $f_{k1}$ | $\cdots$ | $f_{kj}$ | $\cdots$ | $f_{km}$ | $f_{k\cdot}$ |
| | Total | $f_{\cdot 1}$ | $\cdots$ | $f_{\cdot j}$ | $\cdots$ | $f_{\cdot m}$ | 1 |

- Frecuencia relativa marginal de la fila $i$:

$$f_{i\cdot} = f_{i1} + \cdots + f_{ij} + \cdots + f_{im}$$

- Frecuencia relativa marginal de la columna $j$:

$$f_{\cdot j} = f_{1j} + \cdots + f_{ij} + \cdots + f_{kj}$$

## Distribución de frecuencias condicionadas

- Dada la distribución conjunta de $(X, Y)$, se denomina distribución condicionada a la distribución de frecuencias absolutas de una de las variables, suponiendo conocido y fijado el valor de la otra variable.
  **Ejemplo** Distribución de frecuencias de la situación laboral $(Y)$ para personas con un nivel educativo $(X)$ de Primaria o menos.

| $Y\|X =$ Primaria | Empleado | Desempleado | Inactivo | Total |
|---|---|---|---|---|
| $n_{1j}$ | 95 | 6 | 315 | 416 |

- Distribución de frecuencias de $Y$ condicionado a $X = x_i$: Es la distribución de frecuencias de la variable $Y$, de entre los individuos que han tomado el valor $x_i$ de $X$.
- observa que se restringe al conjunto total de individuos que toman el valor $x_i$ de $X$.
- Notación: $Y|X = x_i$.

## Distribución de frecuencias condicionadas

**Ejemplo** Distribución de frecuencias del nivel educativo para inactivos.

| $X\|Y =$ Inactivo | Primaria | Secundaria | Post-secundaria | Total |
|---|---|---|---|---|
| $n_{i3}$ | 315 | 257 | 89 | 661 |

Distribución de frecuencias del nivel educativo para desempleados.

| $X\|Y =$ Desemp. | Primaria | Secundaria | Post-secundaria | Total |
|---|---|---|---|---|
| $n_{i2}$ | 6 | 28 | 8 | 42 |

Se ha visto, así, la definición más sencilla de distribución condicionada. Puede condicionarse también al hecho de que la variable tome varios valores, por ejemplo:
$X|(Y = \text{Inactivo}) \cup (Y = \text{Desempleado})$.

# Tabla de doble entrada para variables cuantitativas

**Ejemplo** 43 alumnos encuestados
$X$: Núm. de veces que han ido al teatro en el último mes
$Y$: Núm. de veces que han ido al cine en el último mes

| | | $Y$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | **0** | **1** | **2** | **3** | **4** | Total |
| | **0** | 12 | 5 | 4 | 2 | 1 | 24 |
| | **1** | 4 | 3 | 2 | 1 | 0 | 10 |
| $X$ | **2** | 3 | 3 | 2 | 0 | 0 | 8 |
| | **3** | 1 | 0 | 0 | 0 | 0 | 1 |
| | Total | 20 | 11 | 8 | 3 | 1 | 43 |

- Si $X$ e $Y$ son cuantitativas discretas tomando un número pequeño de valores, la tabla se construye de la misma forma que para el caso de variables cualitativas.

# Tabla de doble entrada para variables cuantitativas

**Ejemplo** Empresas americanas
$X$: Núm. de trabajadores, $Y$: Volumen de ventas

| | | $X$ | | | | |
|---|---|---|---|---|---|---|
| | | **[1,25)** | **[25,50)** | **[50,75)** | **[75,99]** | Total |
| | **[1,100)** | 0.293 | 0.122 | 0.098 | 0.049 | 0.561 |
| $Y$ | **[100,200)** | 0.098 | 0.073 | 0.049 | 0.024 | 0.244 |
| | **[200,300]** | 0.073 | 0.073 | 0.049 | 0.000 | 0.195 |
| | Total | 0.463 | 0.268 | 0.195 | 0.073 | 1.000 |

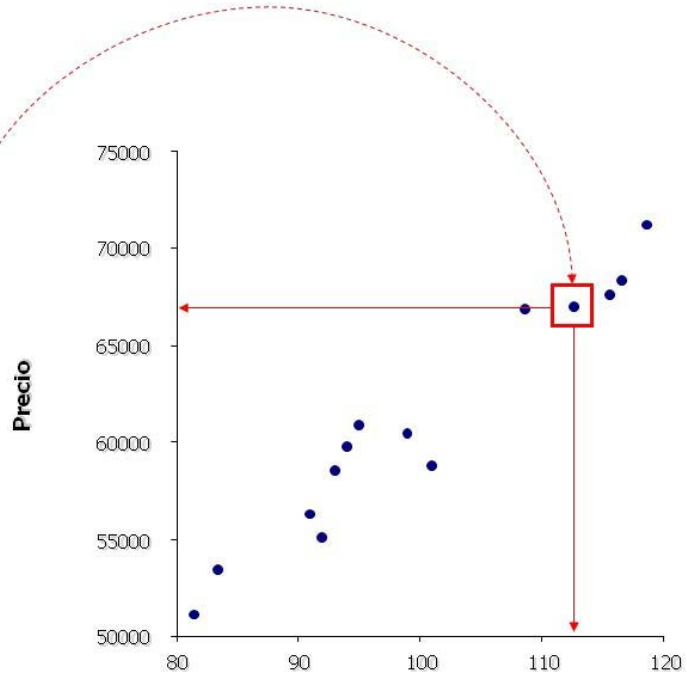
- Si $X$ e $Y$ son, bien cuantitativas discretas tomando un número grande de valores o bien continuas, es habitual agrupar los valores de las variables en intervalos.

# Diagrama de dispersión

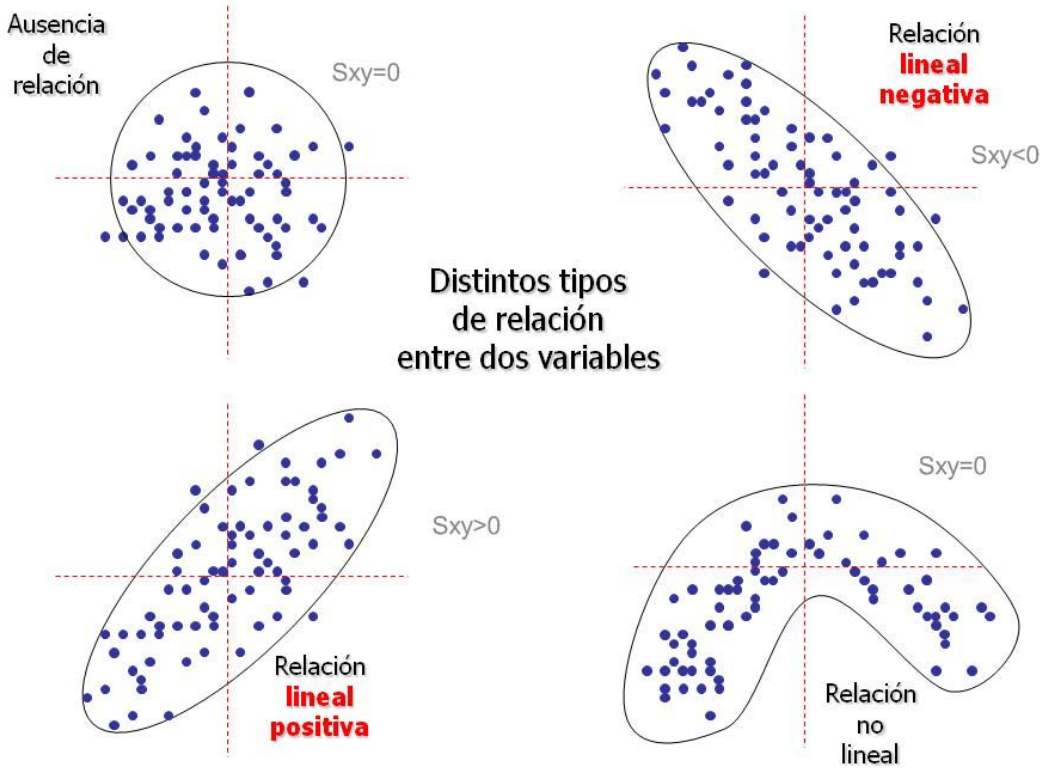
- La representación gráfica más común para dos variables cuantitativas es el diagrama de dispersión

**Ejemplo** $m^2$ habitables y Precio de 15 viviendas.

| m2 | Precio |
|---|---|
| 107 | 62657 |
| 114 | 65554 |
| 91 | 54506 |
| 100 | 62103 |
| 96 | 58271 |
| 107 | 66925 |
| 104 | 61917 |
| 100 | 61149 |
| 80 | 52263 |
| 81 | 51878 |
| 105 | 65678 |
| 111 | 66696 |
| 108 | 65387 |
| 97 | 61806 |
| 106 | 63824 |

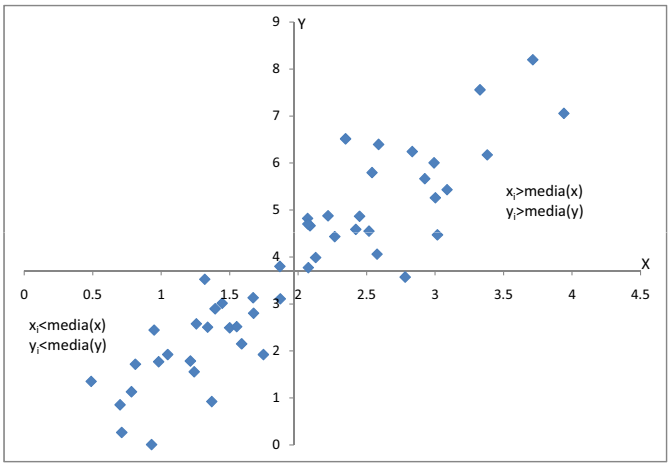



# Tipos de relación entre variables cuantitativas



# Medidas de dependencia lineal para variables cuantitativas

- La covarianza es una medida de variación conjunta de las dos variables. Cuantifica la información existente en un gráfico de dispersión sobre la asociación **lineal** entre dos variables.




Covarianza:

$$s_{xy} = \frac{1}{n-1}\left(\overbrace{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}^{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}\right) \qquad -\infty < s_{xy} < \infty$$

# Medidas de dependencia lineal

- $s_{xy} >> 0 \Rightarrow$ Relación lineal positiva.
- $s_{xy} << 0 \Rightarrow$ Relación lineal negativa.
- $s_{xy} \approx 0 \Rightarrow$ No existe relación lineal o existe relación no lineal.
- Inconvenientes de la covarianza:
  - No está acotada ni superior ni inferiormente. Por lo tanto no se sabe cuándo es $s_{xy}$ suficientemente grande o pequeña.
  - Depende de las unidades de medida de las variables:
    Si $s_{xy}$ es la covarianza de $X$ e $Y$, y $a, b \in \mathbb{R}$, $b \neq 0$ y $T = a + bY$, entonces:
    $s_{xt} = bs_{xy}$

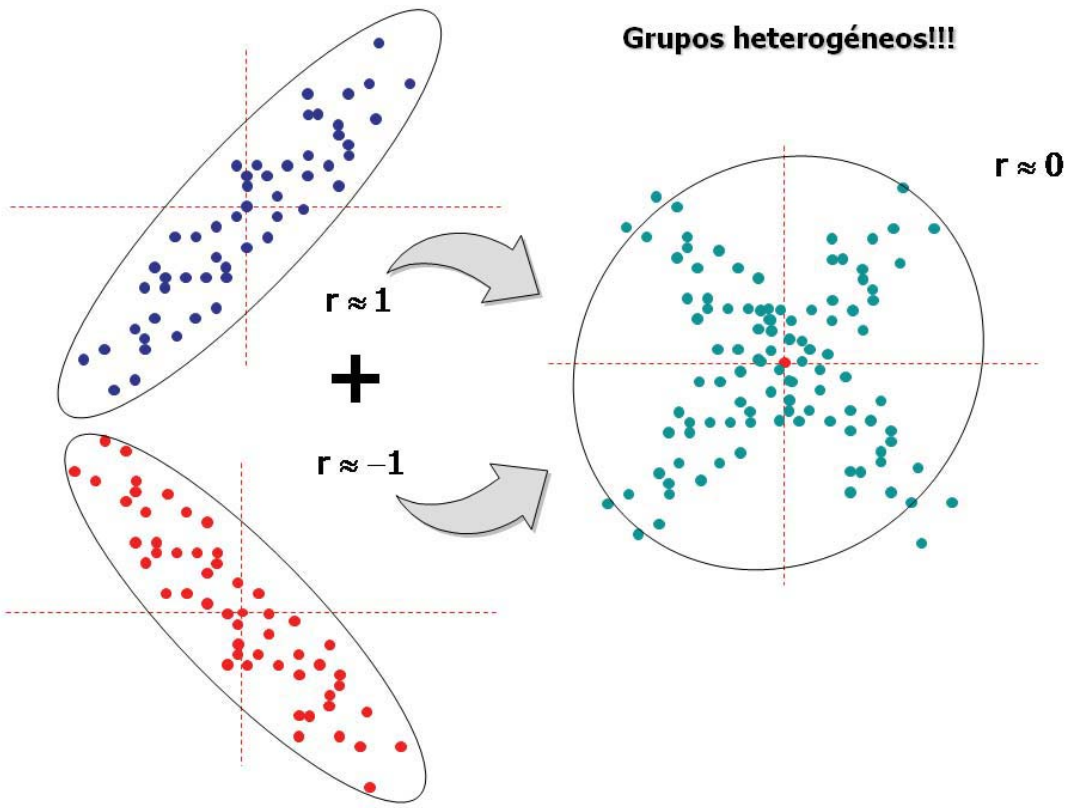
# Medidas de dependencia lineal

- Coeficiente de correlación lineal de Pearson:

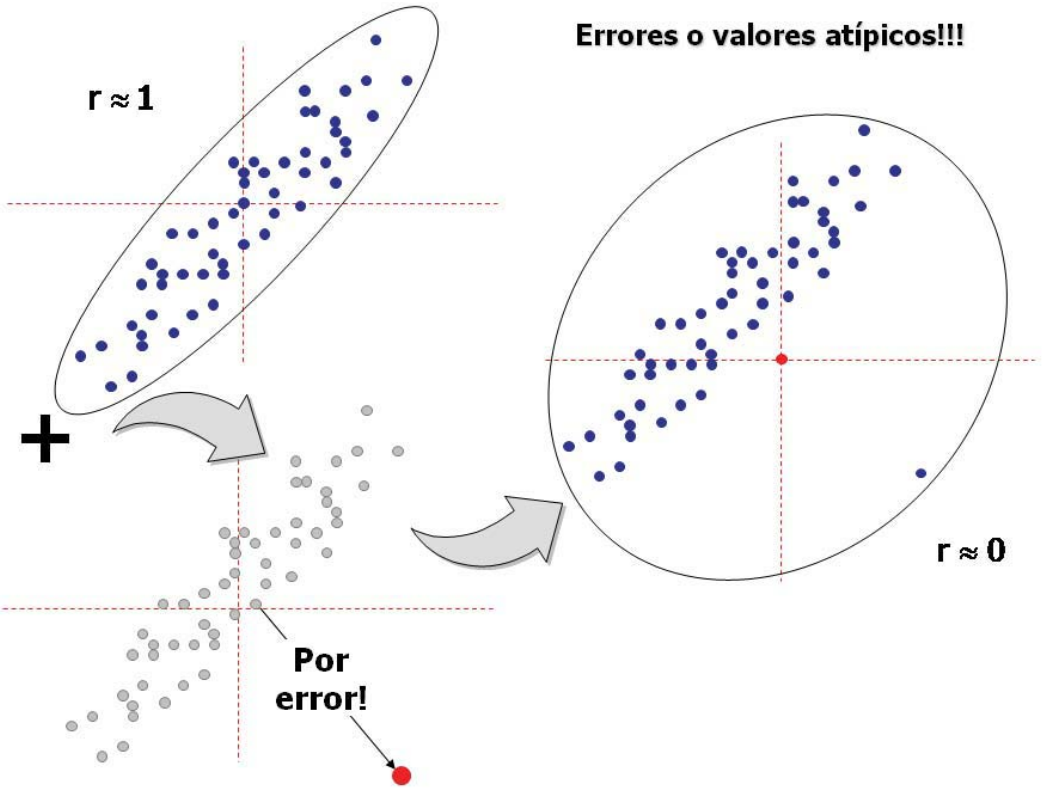
$$r_{(x,y)} = \frac{s_{xy}}{s_x s_y}$$

- ¿Ventajas?
  - Está acotado: $-1 \leq r_{(x,y)} \leq 1$
  - Es adimensional.
  - Interpretación del coeficiente de correlación de Pearson:
    - $r_{(x,y)} > 0$ Dependencia Directa.
    - $r_{(x,y)} < 0$ Dependencia Indirecta.
    - $|r_{(x,y)}| = 1$ Relación Lineal Perfecta.
    - $r_{(x,y)} = 0$ $X$ e $Y$ están Incorreladas (ausencia de relación **lineal**).

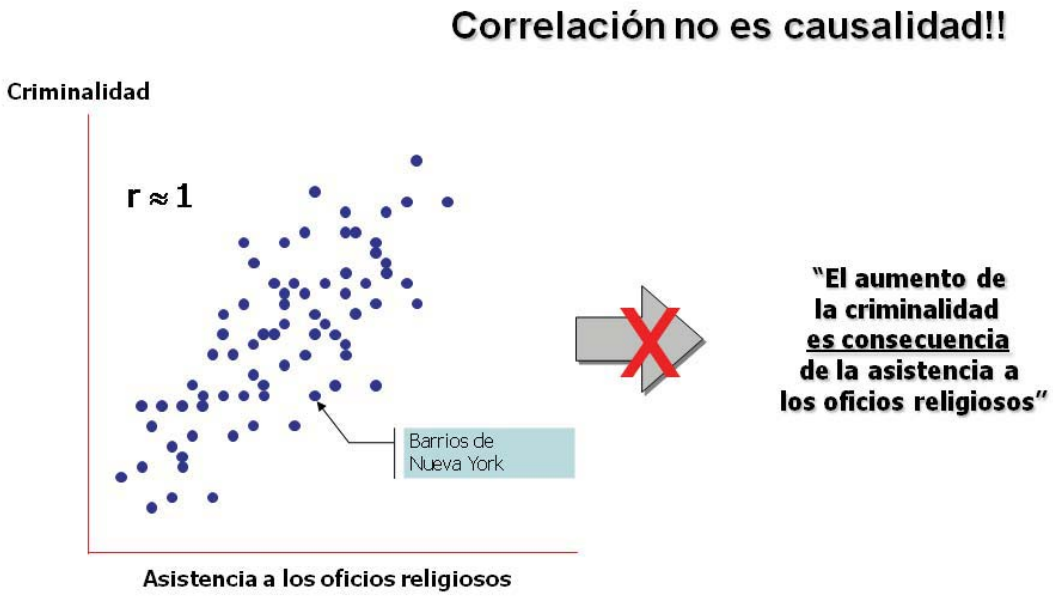
# Correlación y heterogeneidad



# Correlación y datos atípicos

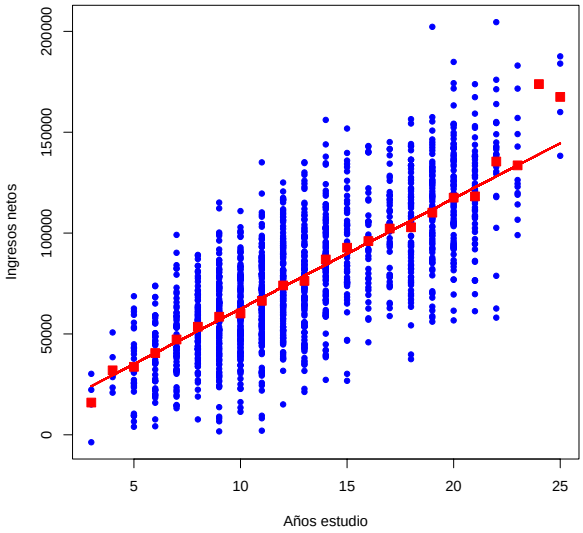



# Correlación ≠ Causalidad



# Recta de regresión

**Ejemplo** Diagrama de dispersión de $Y$: Ingresos netos frente a $X$: Años de estudio para 1508 madrileños.




- Los ingresos netos medios para cada valor de años de estudio (medias de $Y$ condicionadas a cada valor de $X$) forman aproximadamente una línea recta.

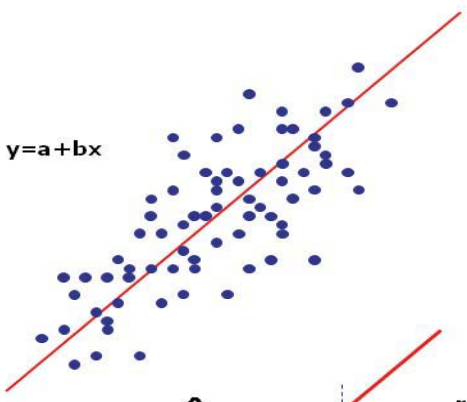
# Definición de la recta de regresión

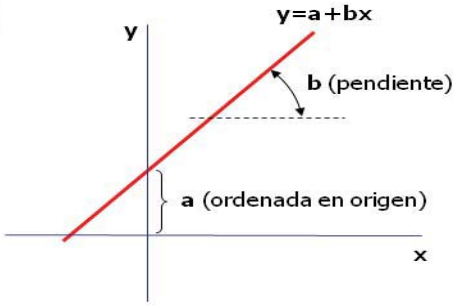
- Dada una muestra de $n$ puntos $(x_i, y_i)$ queremos encontrar la recta $y = a + bx$ que mejor se ajuste a la nube de puntos.
- Ecuación de la recta de regresión:

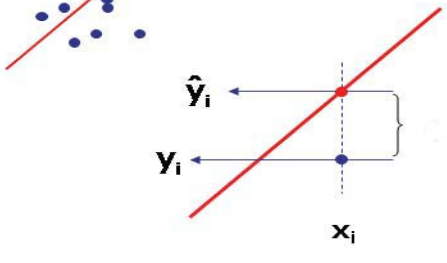
$$\hat{y} = a + bx$$

**Recta de regresión**

Cuando existe una relación lineal, la forma natural de expresar esta relación es a través de la recta que describe la evolución conjunta de ambas variables




residuo $(y_i - \hat{y}_i)$

Generalizando el concepto de la media de una variable.....
.... se llega a la **recta de medias**
→ **método de mínimos cuadrados**
minimizando los residuos cuadráticos

## Estimación de los coeficientes

- Calculamos la ordenada en origen $a$ y la pendiente $b$ minimizando los residuos al cuadrado (método de mínimos cuadrados):

$$\min_{a,b} \sum_{i=1}^{n} (\overbrace{y_i - \hat{y}_i}^{residuo})^2 = \min_{a,b} \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

- Estimadores de $b$ y $a$:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} \\ a &= \bar{y} - b\bar{x} \end{aligned}$$

- Propiedad: $b = r_{(x,y)} \frac{s_y}{s_x}$

porque $b = \frac{s_{xy}}{s_x^2} \frac{s_y}{s_y} = r_{(x,y)} \frac{s_y}{s_x}$

## Interpretación de los coeficientes

- Ecuación de la recta de regresión:

$$\hat{y} = a + bx$$

- $a$ ordenada en el origen: valor medio de $y$ cuando $x = 0$:

$$x = 0 \longrightarrow \hat{y} = a + b(0) = a$$

- $b$ pendiente de la recta: incremento medio en $y$ cuando se aumenta $x$ una unidad:

$$x_2 = x_1 + 1 \longrightarrow \hat{y}_2 - \hat{y}_1 = a + b(x_1 + 1) - (a + bx_1) = b$$

Algunas veces el parámetro $a$ carece de interpretación.

# Interpretación de los coeficientes

**Ejemplo:** Recta de regresión ajustada para Ingresos netos ($Y$) en función de Años de estudio ($X$):

$$\hat{y} = 7667{,}9 + 5474{,}9x$$

- La renta neta media (anual) para individuos con 0 años de estudios es 7667.9 euros. En este caso, NO tiene sentido interpretar este parámetro.
- La renta neta media (anual) aumenta 5474.9 euros por cada año extra de estudios.

# Recta de Regresión con R Commander

**Ejemplo cont.**

Ventana de resultados

Ejecutar

```
Call:
lm(formula = precio ~ tamaño)

Residuals:
     Min       1Q   Median       3Q      Max
-2.39432 -1.05972 -0.00472  0.75528  2.46308

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 113.89371    4.25732   26.75 9.42e-13 ***
tamaño        0.47260    0.04218   11.21 4.73e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.589 on 13 degrees of freedom
Multiple R-squared: 0.9062, Adjusted R-squared: 0.899
F-statistic: 125.5 on 1 and 13 DF,  p-value: 4.735e-08
```

## Valores predichos, residuos y varianza residual

- Valores predichos (siempre que $R^2$ sea elevado y no se esté *extrapolando*:

$$\hat{y}_i = a + bx_i, \quad i = 1, \ldots, n$$

- Residuos:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n$$

  $\overline{e} = 0$ porque $\overline{y} = \overline{\hat{y}}$.
- Varianza residual: Es una medida del error cometido en la predicción de los datos. Divides entre $n-2$ porque se han estimado $a$ y $b$ (menos dos *grados de libertad*).

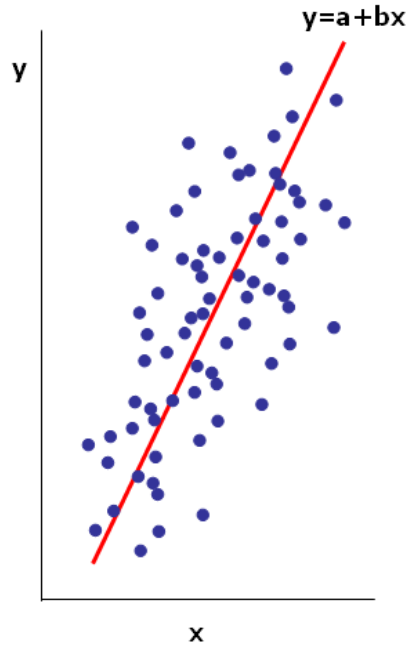
$$s_e^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

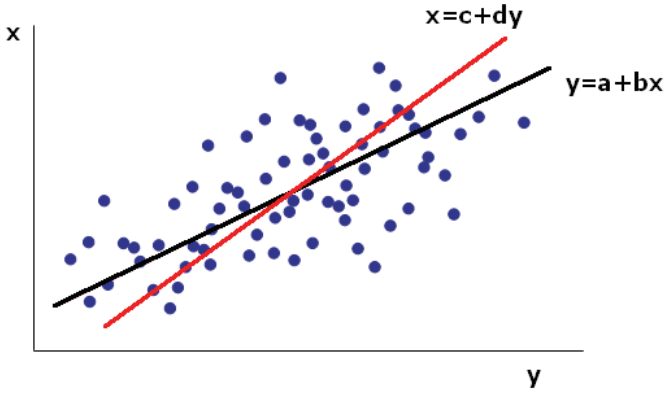
- Relación fundamental de la Regresión: La variabilidad total de la variable dependiente se puede explicar como la suma de la variabilidad del modelo más la variabilidad residual (como sumas de cuadrados): $\mathbf{SCT} = \mathbf{SCR} + \mathbf{SCE}$.

## Propiedades de la recta de regresión

- Existen dos rectas distintas porque provienen de dos problemas de optimización diferentes:

$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$ $\qquad$ $\sum_{i=1}^{n}(x_i - \widehat{x}_i)^2$




$b = \frac{s_{xy}}{s_x^2}$ $\quad a = \overline{y} - b\overline{x}$ $\qquad$ $d = \frac{s_{xy}}{s_y^2}$ $\quad c = \overline{x} - d\overline{y}$

# Bondad del ajuste

- El **coeficiente de determinación** $R^2$ mide el porcentaje de la varianza total (o *información*) recogido por el modelo:

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

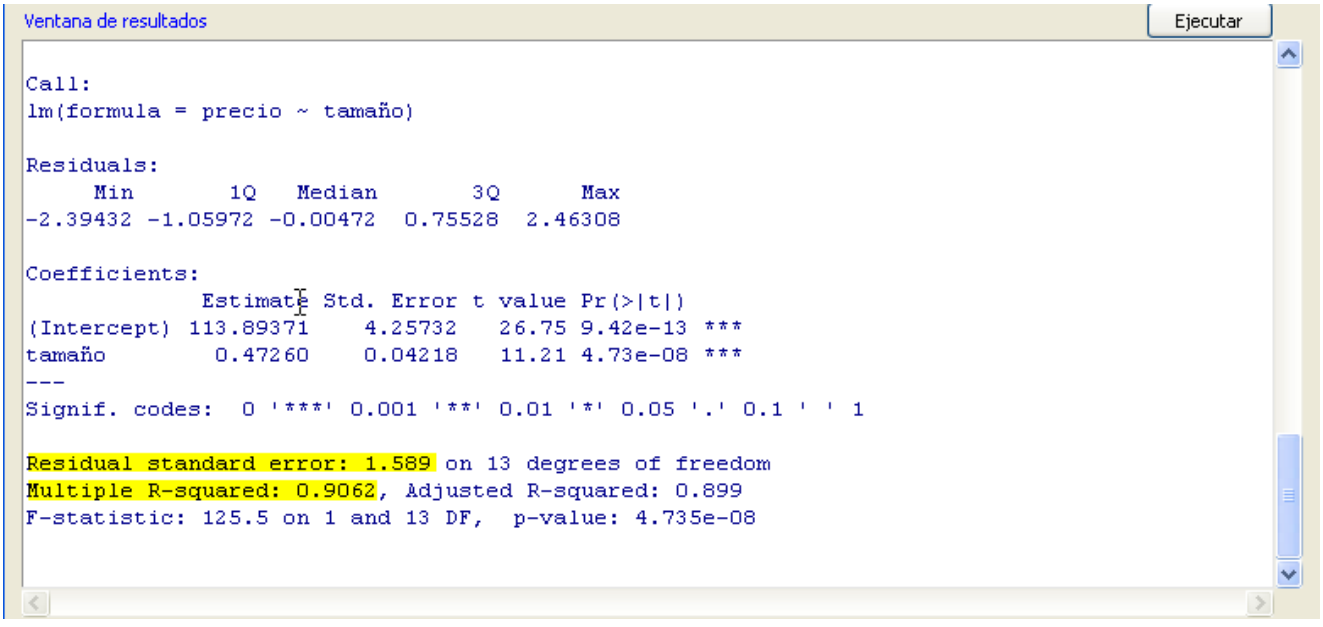
- En el modelo de regresión lineal simple, el Coeficiente de determinación es igual al coeficiente de correlación al cuadrado $R^2 = r^2_{(x,y)}$
- $0 \leq R^2 \leq 1$: Mayor $R^2$ indica un mejor ajuste.
- Interpretación:
  - $R^2 = 1 \Rightarrow$ Ajuste perfecto ($SCE = 0$) todos los residuos son 0
  - $R^2 = 0 \Rightarrow$ Ajuste muy malo ($SCE = SCT$)
  - $R^2 < 0{,}6 \Rightarrow$ Modelo con escasa fiabilidad: Buscar otro modelo mejor...

# Recta de Regresión: $R^2$ en RCommander

**Ejemplo cont.** Calcular e interpretar $R^2$.

$$R^2 = (0{,}9519)^2 = 0{,}9061 = 90{,}61\,\%$$

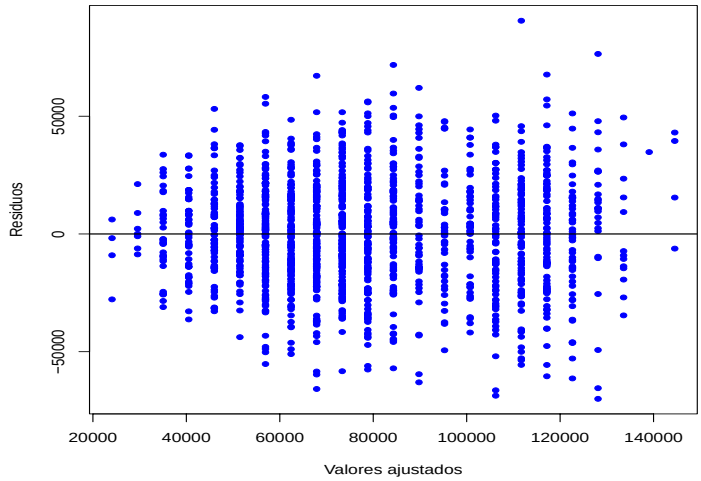
$R^2$ es 90.61 %, así el 90.61 % de la variabilidad muestral en el precio de las casa se explica por el tamaño.



# Análisis de los residuos

- Para ver si el modelo de regresión es adecuado, se hace un diagrama de dispersión de residuos frente a valores predichos.
- Un gráfico de residuos en el que los puntos parecen aleatorios indica que la recta de regresión se ajusta correctamente.

**Ejemplo** Gráfico de residuos para la recta de ingresos netos:



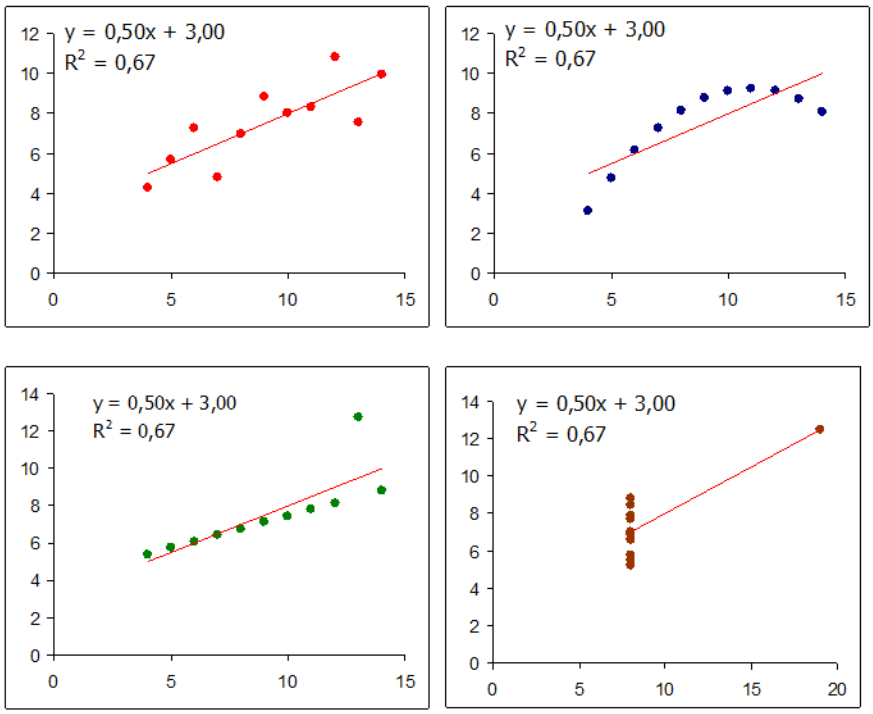

# Los datos de *Anscombe*

**Ejemplo** Los datos de *Anscombe*

| | Datos 1 | | Datos 2 | | Datos 3 | | Datos 4 | |
|---|---|---|---|---|---|---|---|---|
| i | *x* | *y* | *x* | *y* | *x* | *y* | *x* | *y* |
| 1 | 10 | 8,04 | 10 | 9,14 | 10 | 7,46 | 8 | 6,58 |
| 2 | 8 | 6,95 | 8 | 8,14 | 8 | 6,77 | 8 | 5,76 |
| 3 | 13 | 7,58 | 13 | 8,74 | 13 | 12,74 | 8 | 7,71 |
| 4 | 9 | 8,81 | 9 | 8,77 | 9 | 7,11 | 8 | 8,84 |
| 5 | 11 | 8,33 | 11 | 9,26 | 11 | 7,81 | 8 | 8,47 |
| 6 | 14 | 9,96 | 14 | 8,1 | 14 | 8,84 | 8 | 7,04 |
| 7 | 6 | 7,24 | 6 | 6,13 | 6 | 6,08 | 8 | 5,25 |
| 8 | 4 | 4,26 | 4 | 3,1 | 4 | 5,39 | 19 | 12,5 |
| 9 | 12 | 10,84 | 12 | 9,13 | 12 | 8,15 | 8 | 5,56 |
| 10 | 7 | 4,82 | 7 | 7,26 | 7 | 6,42 | 8 | 7,91 |
| 11 | 5 | 5,68 | 5 | 4,74 | 5 | 5,73 | 8 | 6,89 |

- Los datos de *Anscombe* son cuatro conjuntos de datos de dos variables cada uno. Estos conjuntos de datos son distintos pero la recta de regresión que determinan es la misma.
- Nos muestran la necesidad de observar los datos antes de interpretar la recta de regresión.

# Los datos de *Anscombe*




# Los datos de *Anscombe*

- Recta ajustada:

$$\hat{y} = 3 + 0{,}5x$$

- Interpretación de la pendiente: un incremento de una unidad en la variable $X$ produce, en promedio, un incremento de 0.5 unidades en la variable $Y$.
- Predicción para $x = 7{,}5$ (dentro del rango de $X$)

$$\hat{y} = 3 + 0{,}5(7{,}5) = 6{,}75$$