# Correlación y Regresión Lineal

# Existe relación entre estas variables?

Un profesor de estadística realiza un estudio para investigar la relación que existe entre el rendimiento de sus estudiantes en los exámenes y su ansiedad. Elige a 10 estudiantes de su grupo para el experimento. Justo antes de asistir al examen final, los 10 estudiantes contestan un cuestionario de ansiedad. Los resultados obtenidos son los siguientes

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ansiedad | 28 | 41 | 35 | 39 | 31 | 42 | 50 | 46 | 45 | 37 |
| Examen final | 82 | 58 | 63 | 89 | 92 | 64 | 55 | 70 | 51 | 72 |

# Análisis de Correlación

# Correlación y Regresión Lineal

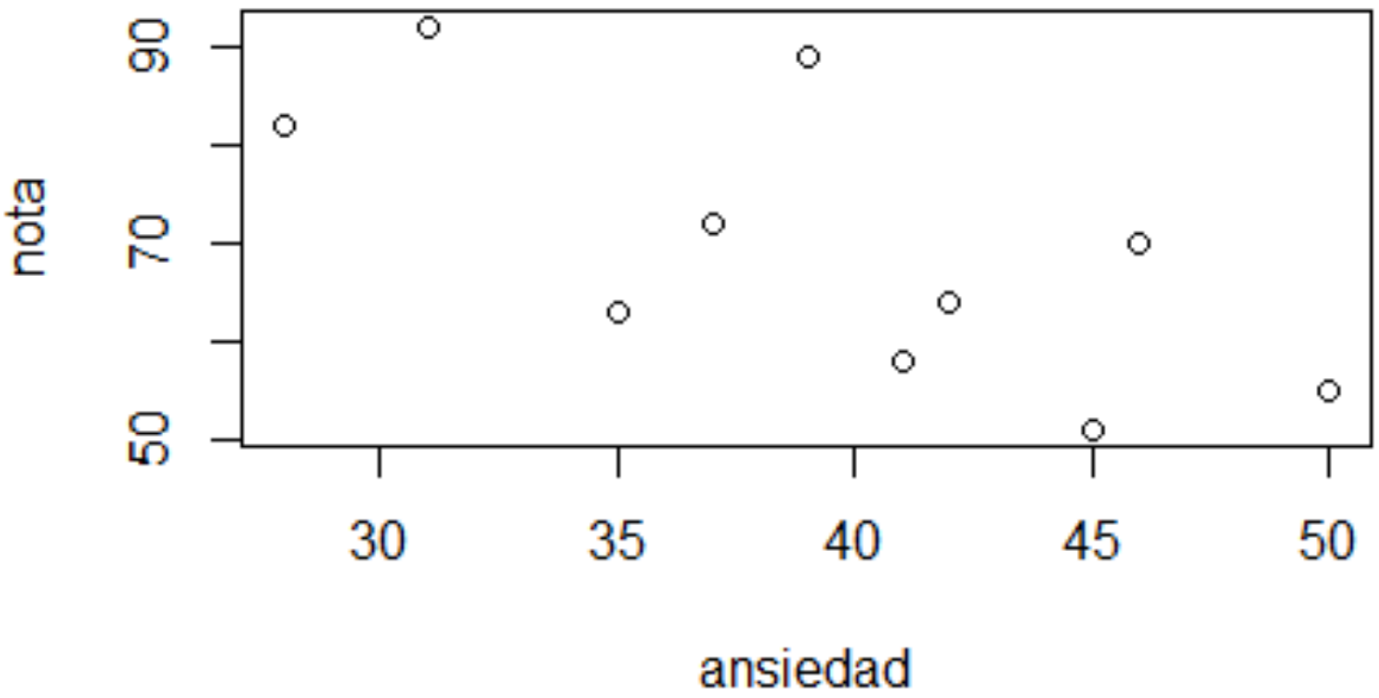
- El análisis de la asociación entre variables depende del nivel de medición de las mismas.
- Nos remitiremos en lo que sigue a analizar el caso en el que las variables analizadas son cuantitativas

- El análisis estadístico de la asociación entre variables se realiza mediante

  ❖Gráficos (Gráfico de dispersión)

  ❖Resúmenes numéricos (Covarianza/Coeficiente de correlación lineal)

# Gráfico de dispersión



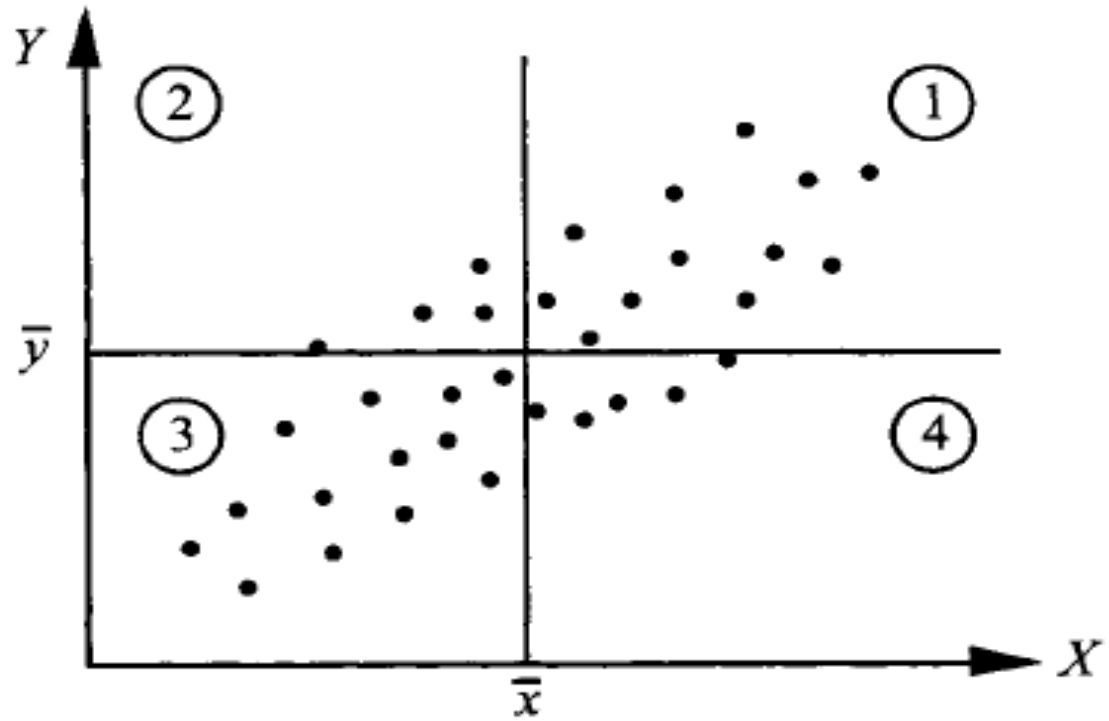
# Covarianza y Coeficiente de Correlación de Pearson

## Covarianza

$$Cov(X,Y) = \frac{\sum(x_i - \bar{X})(Y_i - \bar{Y})}{n}$$

- Permite evaluar asociaciones del tipo lineal
- Puede tomar cualquier valor
- Depende de las unidades de medida de las variables

# Interpretación de la Covarianza



# Interpretación de la Covarianza

$cov(X,Y) > 0 \longrightarrow$ asociación lineal directa

$cov(X,Y) < 0 \longrightarrow$ asociación lineal inversa

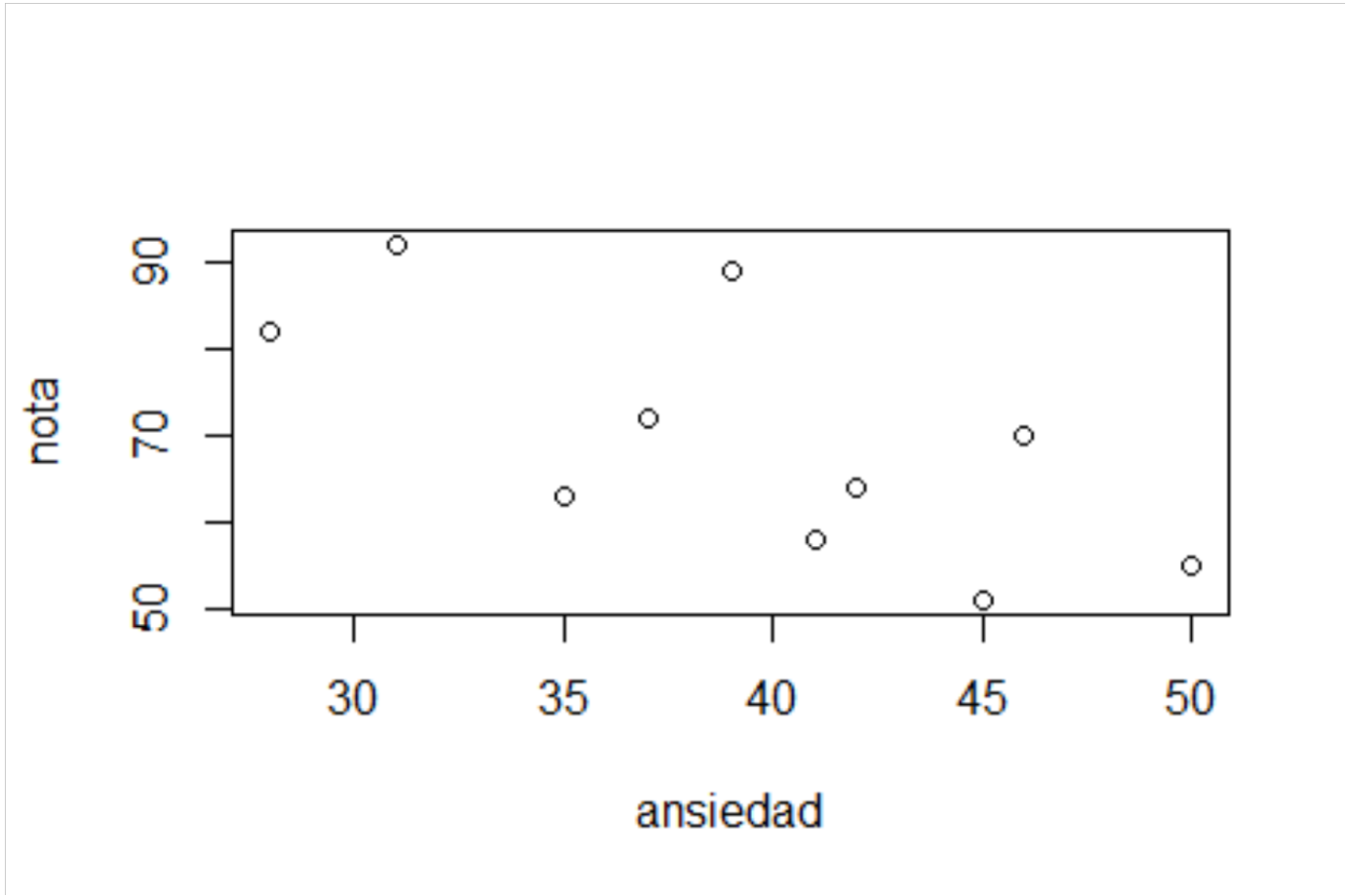
$cov(X,Y) = 0 \longrightarrow$ no están asociadas linealmente

# Covarianza y Coeficiente de Correlación de Pearson

**Coeficiente de Correlación lineal de Pearson**

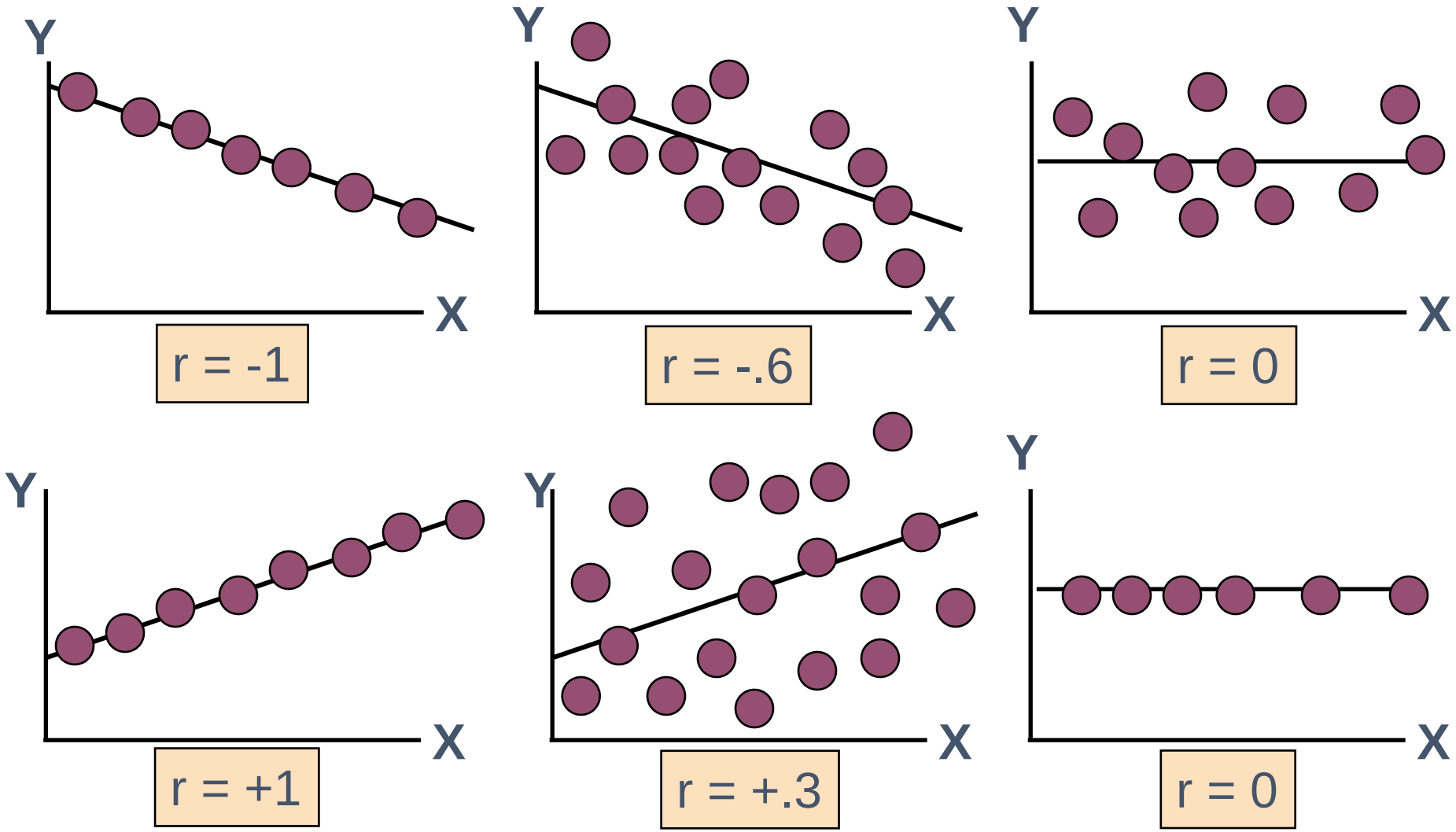
$$r = \frac{Cov(X,Y)}{S_X * S_Y}$$

- Permite evaluar asociaciones del tipo lineal
- $-1 \leq r \leq 1$
- Es independiente de las unidades de medida de las variables

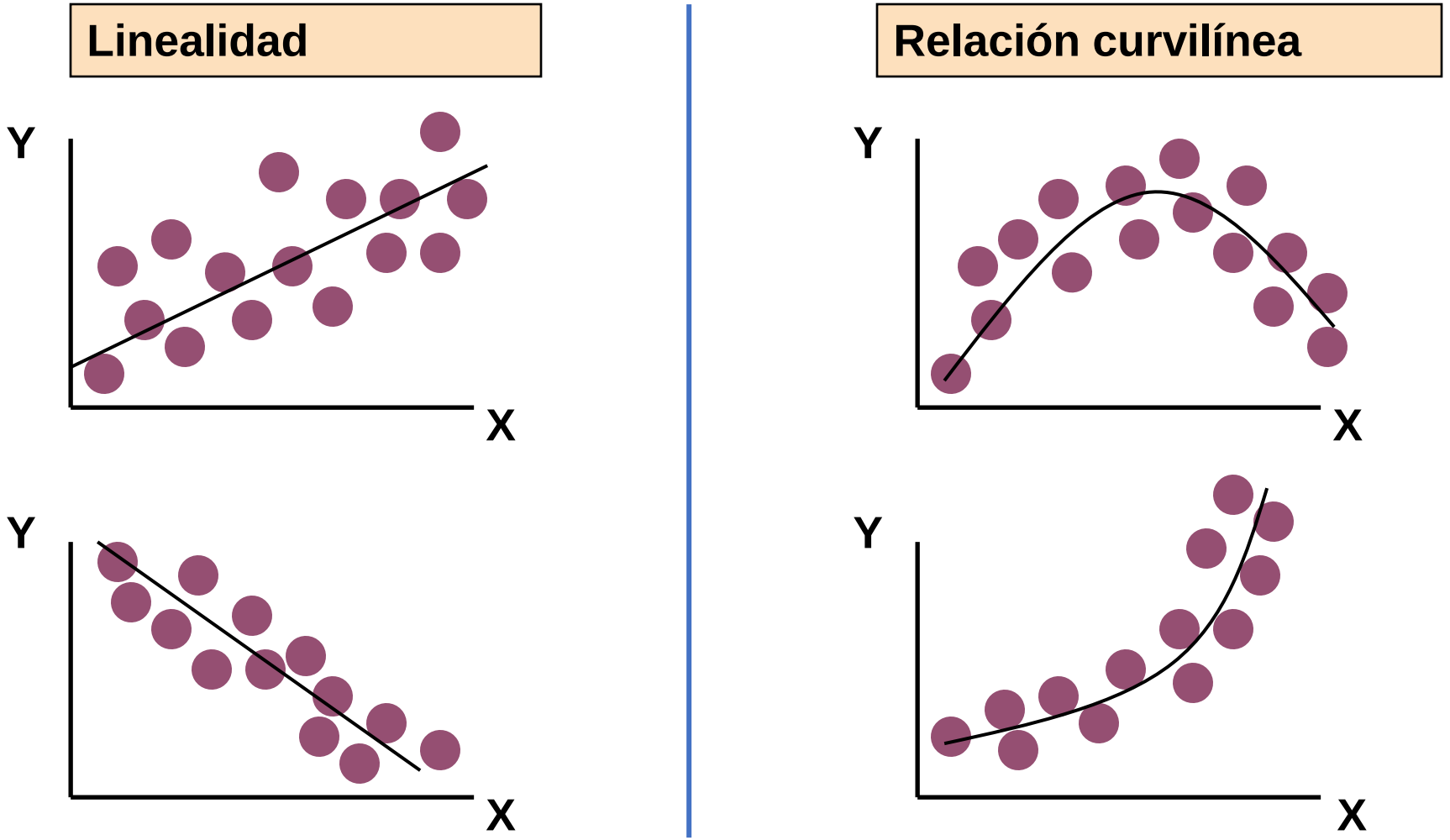



r = -0.6907746

# Distintos tipos de asociación

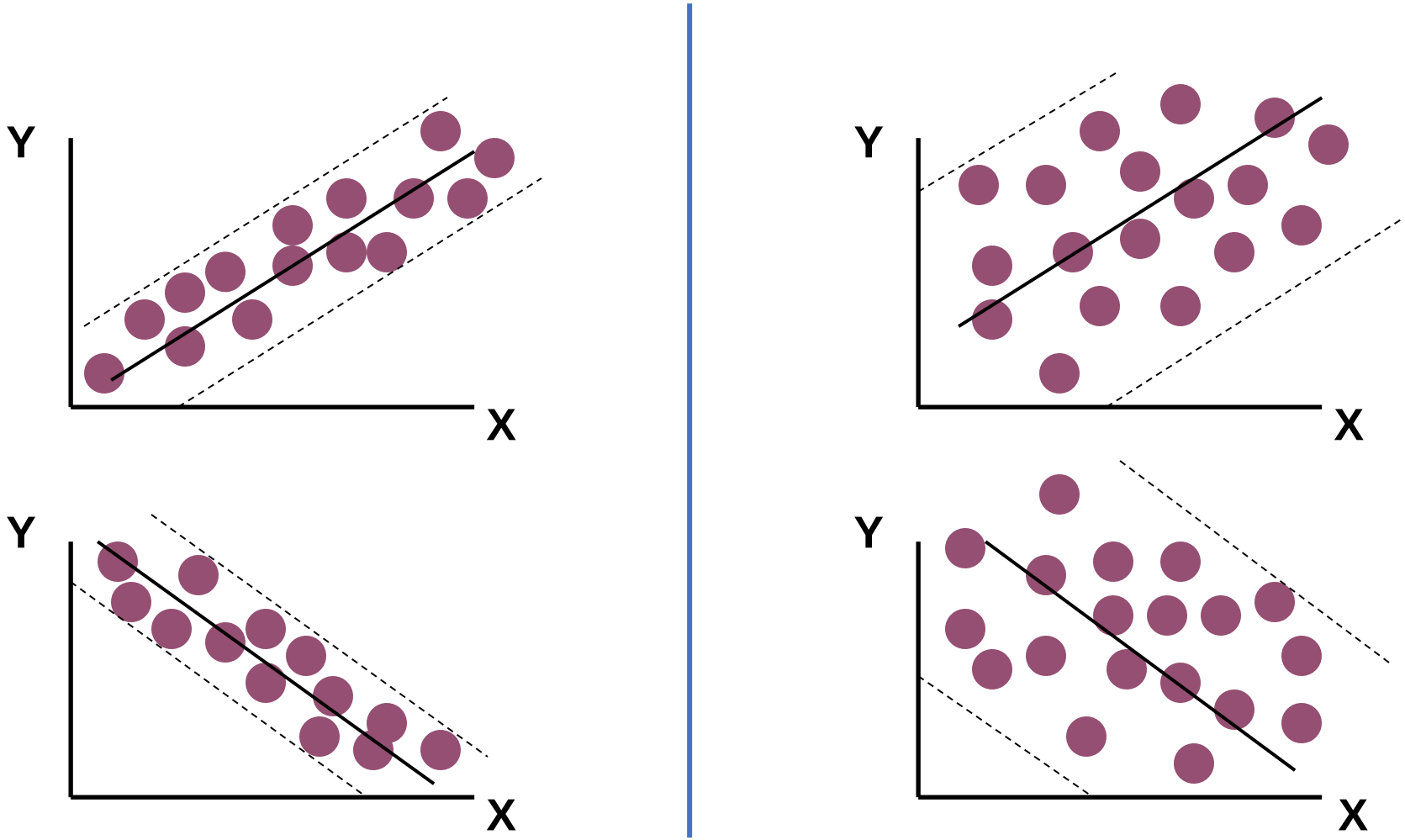


# Distintos tipos de asociación



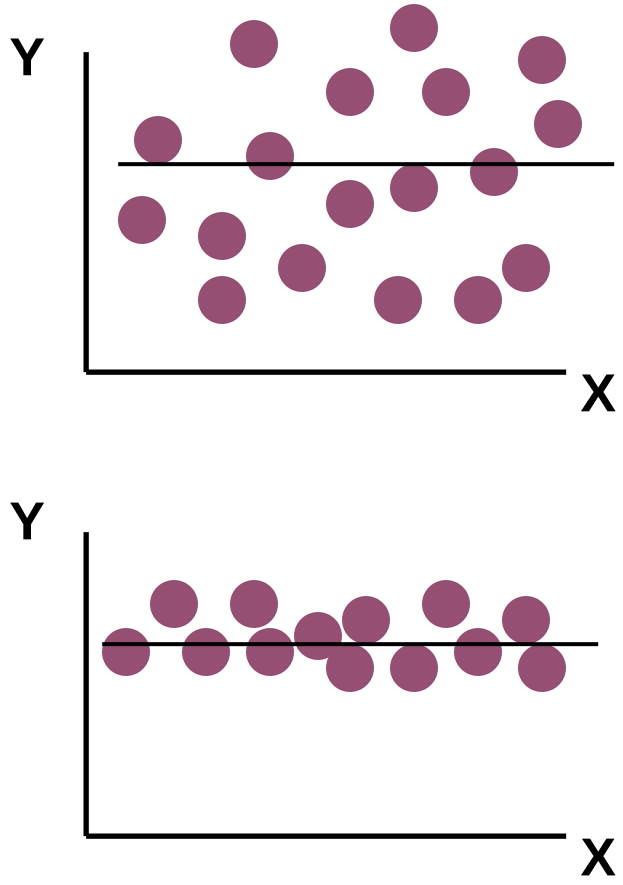

- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

# Asociación lineal




- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

# Asociación lineal




- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

# Análisis de Regresión lineal

- Es posible predecir el rendimiento de los alumnos a partir de su nivel de ansiedad?

Un profesor de estadística realiza un estudio para investigar la relación que existe entre el rendimiento de sus estudiantes en los exámenes y su ansiedad. Elige a 10 estudiantes de su grupo para el experimento. Justo antes de asistir al examen final, los 10 estudiantes contestan un cuestionario de ansiedad. Los resultados obtenidos son los siguientes

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Ansiedad | 28 | 41 | 35 | 39 | 31 | 42 | 50 | 46 | 45 | 37 |
| Examen final | 82 | 58 | 63 | 89 | 92 | 64 | 55 | 70 | 51 | 72 |

# Análisis de Regresión Lineal

- En otras palabras: es posible definir un modelo estadístico que vincule estas variables?
- Es claro que el rol desempeñado por las variables es diferente. La variable cuyos valores quieren predecirse se denomina variable *dependiente*
- Las restantes se denominan *predictoras* o *independientes*

# Ecuación de Regresión Lineal para una variable dependiente

- *Si se tienen $(x_1,y_1),\ldots,(x_n,y_n)$, el modelo de regresión lineal es*

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- *Para i=1,2,..,n*

- *$\varepsilon_i$ es el error del modelo, que se supone aleatório y de media cero para todo i.*
- *Los parámetros α y β deben estimarse a partir de los datos.*

# Ecuación de Regresión Lineal para una variable dependiente

- *Si a y b son los estimadores de α y β respectivamente, se obtiene la recta de regresión estimada o ajustada*

$$y_i = a + bx_i + e_i$$

- *Para i=1,2,..,n*
- $e_i = y_i - \hat{y}_i$ *es el residuo y describe el error en el ajuste del modelo en cada punto.*

# Estimación de los parámetros de la ecuación de Regresión Lineal

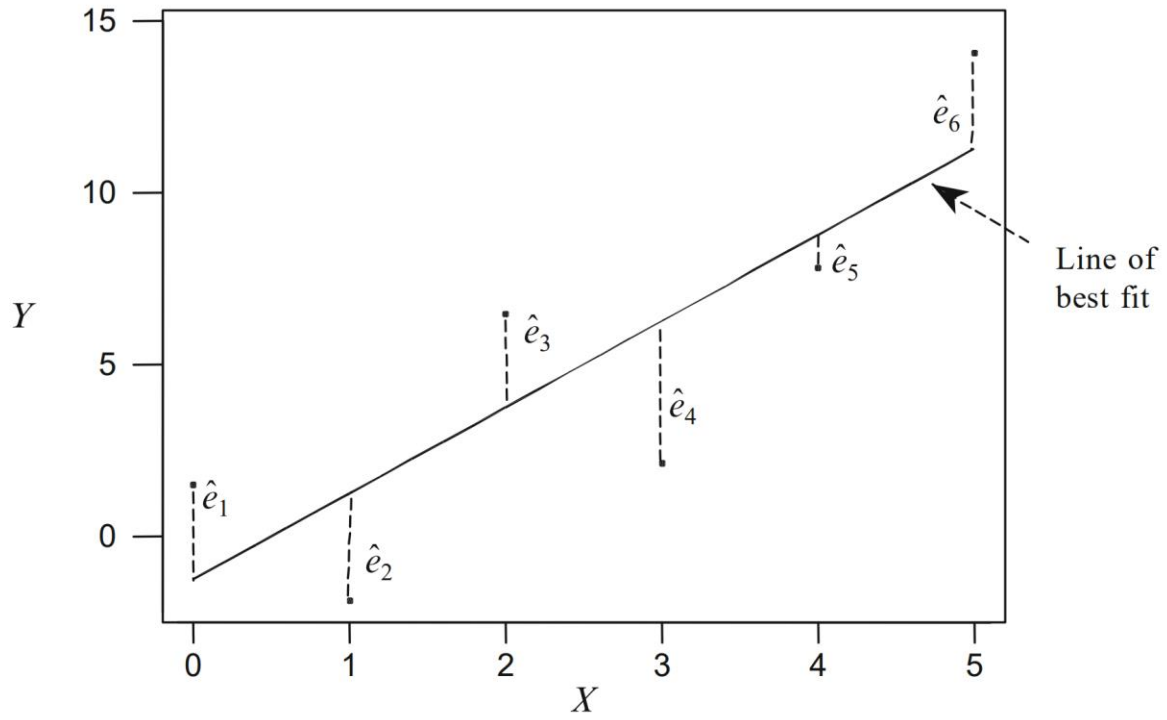
- *Los estimadores a y b de los parámetros de la regresión α y β, pueden hallarse mediante el método de mínimos cuadrados*
- *Se trata de encontrar los valores de a y b que minimicen*

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$

- *Es factible derivar fórmulas explícitas para estos estimadores mediante la resolución de este problema de minimización*
- *Observar que sólo se ha supuesto aleatoriedad de los errores y media cero*

# Estimación de los parámetros de la ecuación de Regresión Lineal

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - bx_i)^2$$



# Estimación de los parámetros de la ecuación de Regresión Lineal

- *También se obtiene un estimador para la varianza común del problema, σ*

$$S^2 = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum (y_i - a - bx_i)^2}{n-2}$$

# Inferencia sobre los parámetros de la ecuación de Regresión Lineal

- *Para testear la significación del estimador de $\beta$*

$$H_0: \beta = 0 \qquad H_1: \beta \neq 0$$

Estadístico de prueba: $t = \frac{b}{s/S_{xx}}$

*Siendo* $S_{xx} = \sum(x_i - \bar{x})^2$

*Bajo el supuesto de normalidad t sigue una distribución t de Student con n-2 grados de libertad*

# Inferencia sobre los parámetros de la ecuación de Regresión Lineal

*Para testear la significación del estimador de α se utiliza el estadístico*

$$H_0: \alpha = 0 \qquad H_1: \alpha \neq 0$$

*Estadístico de prueba* $t = \dfrac{a}{s\sqrt{\sum x_i^2 / n S_{xx}}}$

*Bajo el supuesto de normalidad, t sigue una distribución de Student con n-2 grados de libertad*

# Análisis de la Varianza en Regresión Lineal

- *La suma total de cuadrados SST de la variable dependiente puede escribirse como*

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

- *En otras palabras*

- *SST = SSR+SSE*

# Análisis de la Varianza en Regresión Lineal

- SST es la suma de cuadrados de *y* y refleja su variabilidad en torno a su media
- SSR se llama suma de cuadrados de la regresión y refleja la cantidad de variación de los valores de *y* explicados por el modelo
- SSE es la suma de cuadrados del error

# Análisis de la Varianza en Regresión Lineal

- Análisis de la varianza para testear la significatividad de la regresión

| Fuente de variación | Suma de cuadrados | Grados de libertad | Cuadrados medios | F |
|---|---|---|---|---|
| Regresión<br>Error | SSR<br>SSE | 1<br>n-2 | SSR<br>SSE/(n-2) | F |
| Total | SST | n-1 | | |

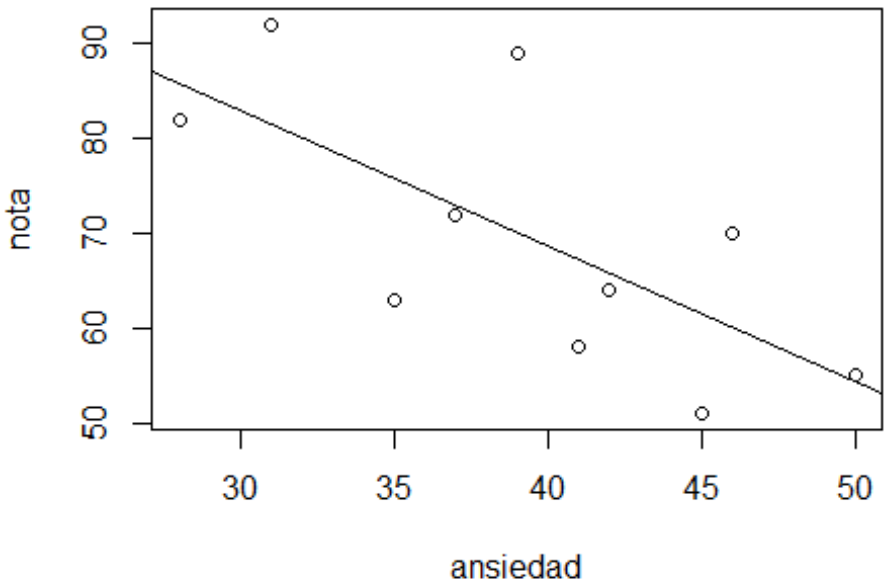
- Siendo

$$F = \frac{SSR/1}{SSE/(n-2)}$$

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 125.8830    21.1109   5.963 0.000337 ***
## ansiedad     -1.4285     0.5287  -2.702 0.026986 *
```

```
## Multiple R-squared:  0.4772, Adjusted R-squared:  0.4118
## F-statistic: 7.301 on 1 and 8 DF,  p-value: 0.02699
```

$$nota = 125.88 - 1.42 * ansiedad$$



# La función lm()

- `lm(formula, data, subset, weights, na.action, method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE, singular.ok = TRUE, contrasts = NULL, offset, …)`