REGRESION Y CORRELACION LINEALES

Relaciones entre variables y regresión

- **El término regresión fue introducido por Galton (1889) refiriéndose a la "ley de la regresión universal":**
  - **"Cada peculiaridad en un hombre es compartida por sus descendientes, pero en media, en un grado menor."**
    - **Regresión a la media**
  - **Su trabajo se centraba en la descripción de los rasgos físicos de los descendientes (una variable) a partir de los de sus padres (otra variable).**
  - **Pearson realizó un estudio con más de 1000 registros de grupos familiares observando una relación del tipo:**
    - **Altura del hijo = 85cm + 0,5 altura del padre (aprox.)**
    - **Conclusión: los padres muy altos tienen tendencia a tener hijos que heredan parte de esta altura, aunque tienen tendencia a acercarse (*regresar*) a la media. Lo mismo puede decirse de los padres muy bajos.**

Francis Galton

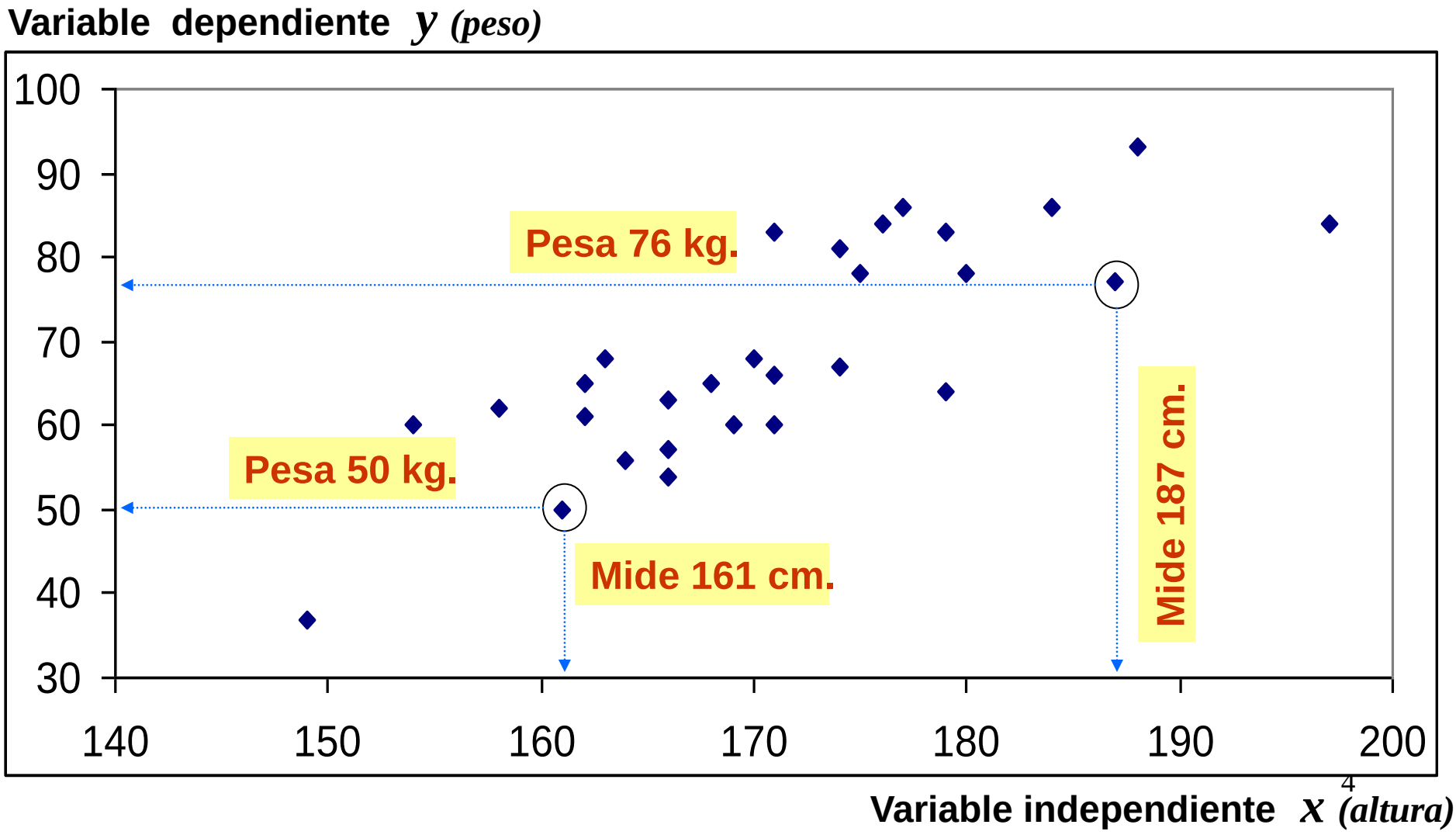
Karl Pearson

Regresión

- **Describir la relación entre dos variables numéricas**
- **El análisis de regresión sirve para predecir una medida en función de otra medida (o varias).**
  - **Y = Variable dependiente**
    - **predicha**
    - **explicada**
  - **X = Variable independiente**
    - **predictora**
    - **explicativa**
  - **¿Es posible descubrir una relación?**
    - **Y = f(X) + error**
      - **f es una función de un tipo determinado**
      - **el error es aleatorio, pequeño, y no depende de X**

Diagramas de dispersión , nube de puntos o “Scaterplot”

Tenemos las alturas y los pesos de 30 individuos representados en un diagrama de dispersión.



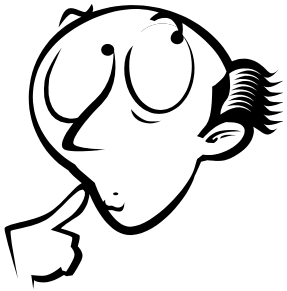
REGRESION LINEAL SIMPLE

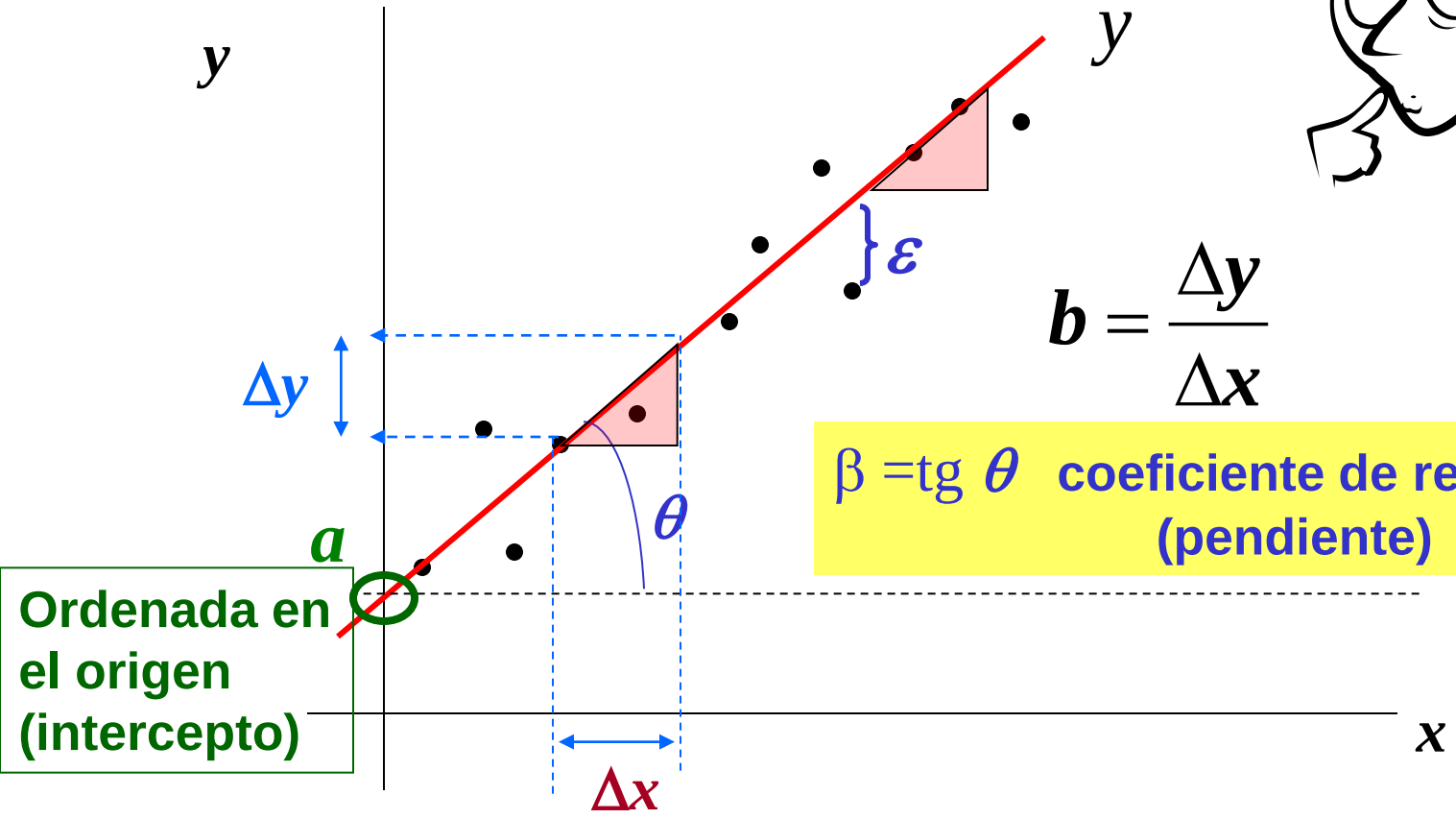
Finalidad

Estimar los valores de *y* (variable dependiente) a partir de los valores de *x* (variable independiente)

Modelo

$$y = \alpha + \beta x + \varepsilon$$



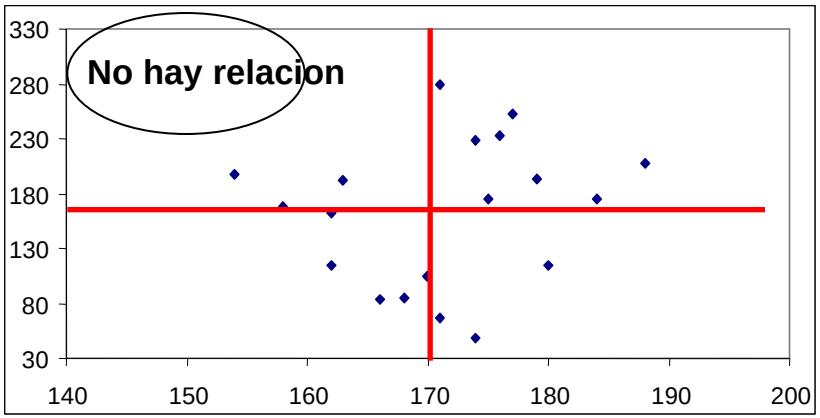


$$b = \frac{\Delta y}{\Delta x}$$

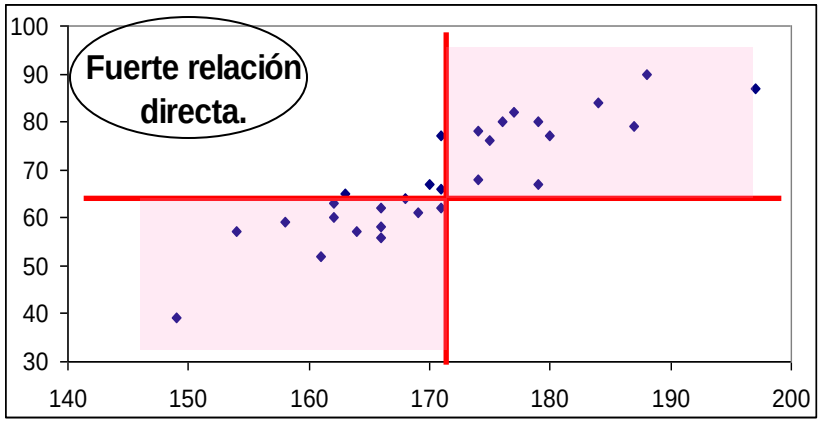
$\beta = \text{tg}\ \theta$ **coeficiente de regresión (pendiente)**

Ordenada en el origen (intercepto)

Relación directa e inversa

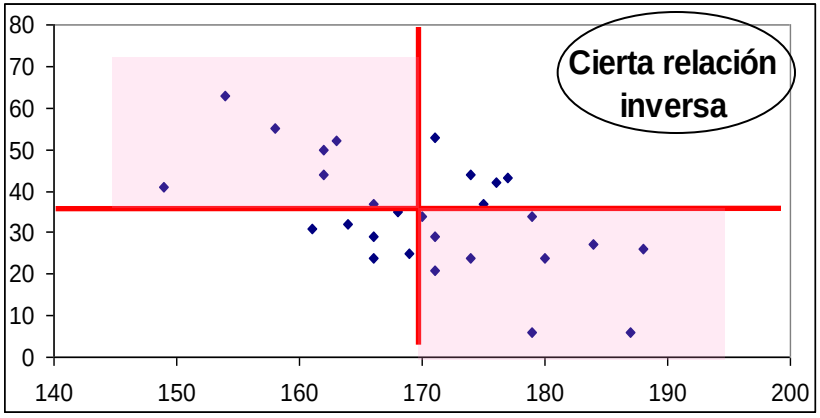


Para valores de X por encima de la media tenemos valores de Y por encima y por debajo en proporciones similares.



•Para los valores de X mayores que la media le corresponden valores de Y mayores también.

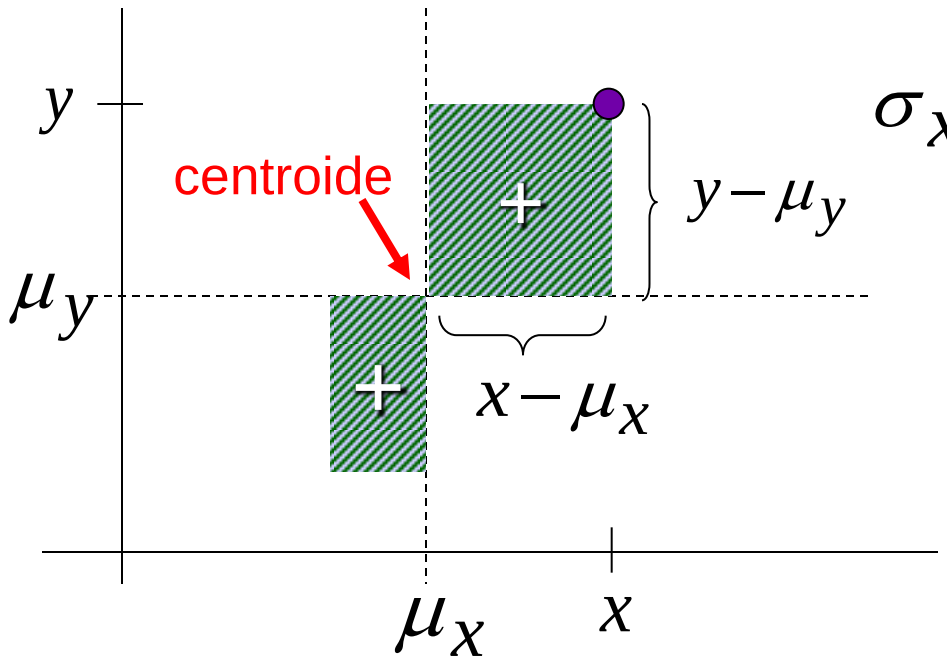
•Para los valores de X menores que la media le corresponden valores de Y menores también. : relación directa.



Para los valores de X mayores que la media le corresponden valores de Y menores. Esto es relación inversa o decreciente.

COVARIANZA

Es una medida de la variación lineal conjunta de dos variables



$$\sigma_{xy} = \frac{\sum (y-\mu_y)(x-\mu_x)}{N}$$

Estimación de σ_{xy}

$$\text{cov} = \frac{\sum (y-\bar{y})(x-\bar{x})}{n}$$

Es un estimador sesgado

- $\sigma_{xy} < 0$ asociación lineal con pendiente negativa
- $\sigma_{xy} = 0$ ausencia de asociación lineal
- $\sigma_{xy} > 0$ asociación lineal con pendiente positiva

- **El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables.**

Coef. de correlación lineal de Pearson

r	Valor en la muestra
ρ	(Rho) en la poblaciòn

El coeficiente de correlación lineal de Pearson de dos variables, r, indica si los puntos tienen una tendencia a disponerse alineadamente (excluyendo rectas horizontales y verticales).

CORRELACION LINEAL

Finalidad

Medir la intensidad de la asociación lineal entre dos variables aleatorias

coeficiente de correlación

$$\rho = \sigma_{xy} / \sigma_x \sigma_y$$

covarianza poblacional

$$r = s_{xy} / s_x s_y$$

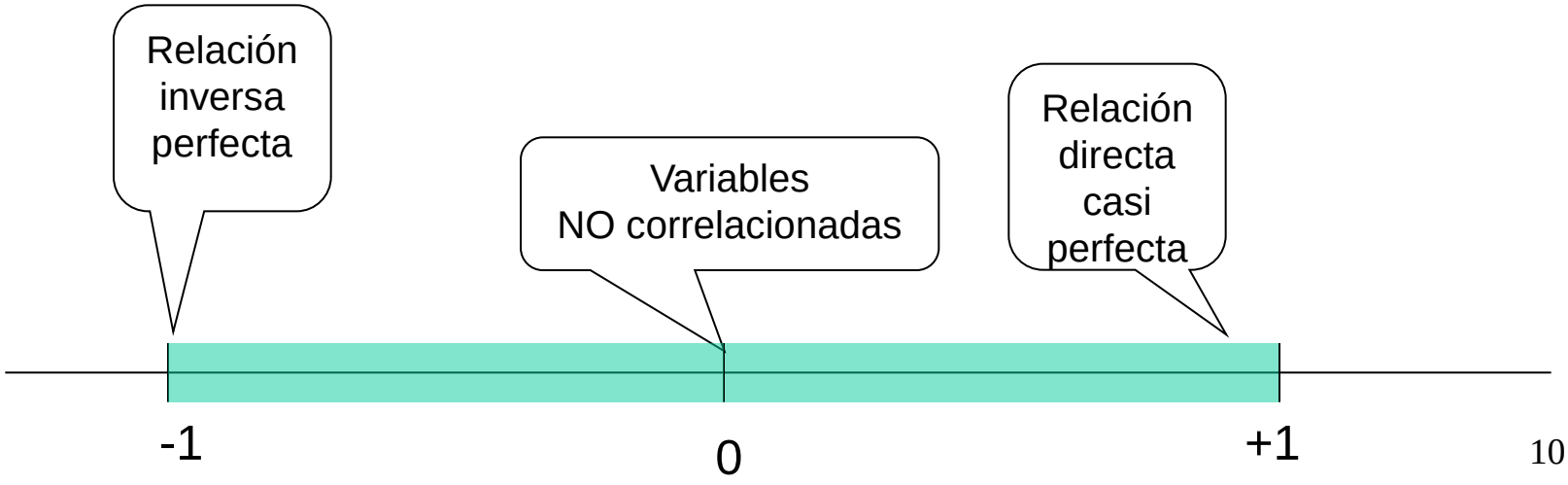
coeficiente de determinación

ρ^2 r^2

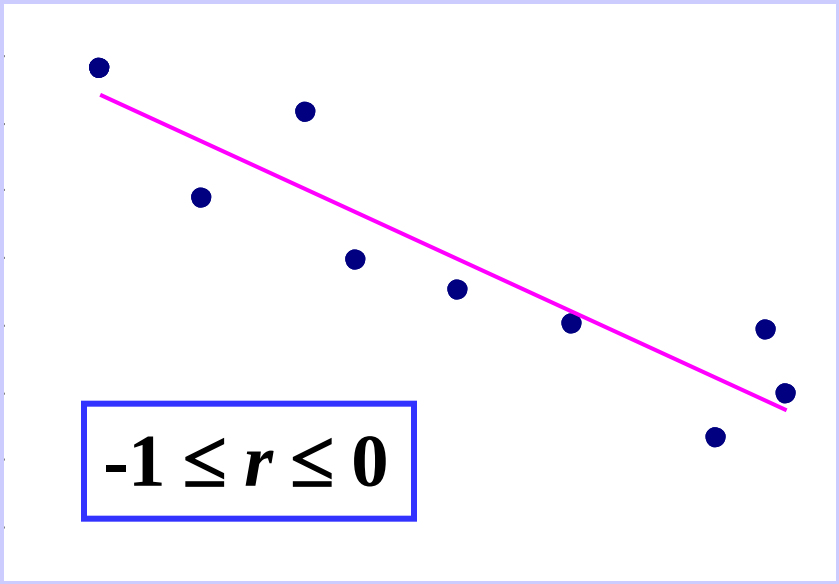
Proporción de varianza compartida por las dos variables

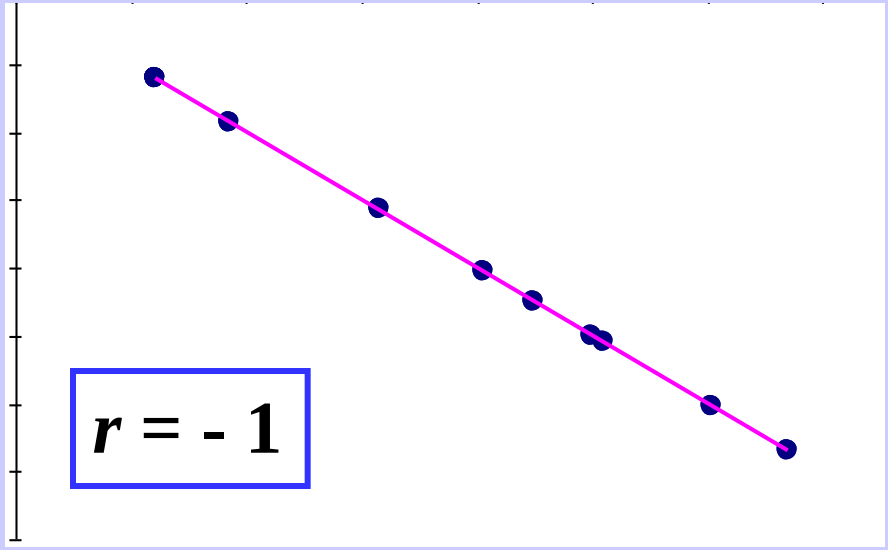
Propiedades de r

- **Es adimensional**
- **Sólo toma valores entre -1y +1**
- **Las variables NO estàn correlacionadas ⇔ r=0**
- **Relación lineal perfecta entre dos variables ⇔ r = +1 o r=-1**
  - **Excluimos los casos de puntos alineados horiz. o verticalmente.**
- **Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.**
  - **Siempre que no existan observaciones anómalas.**

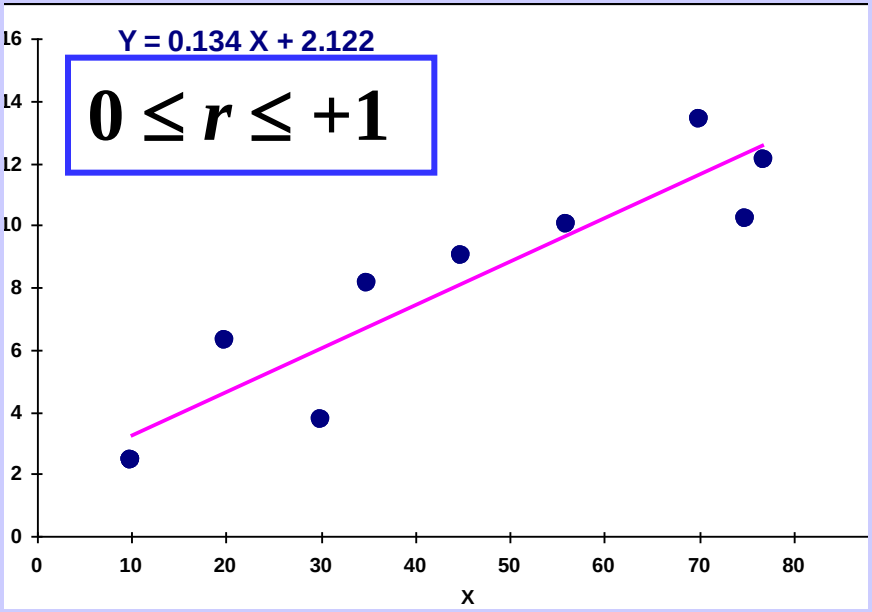


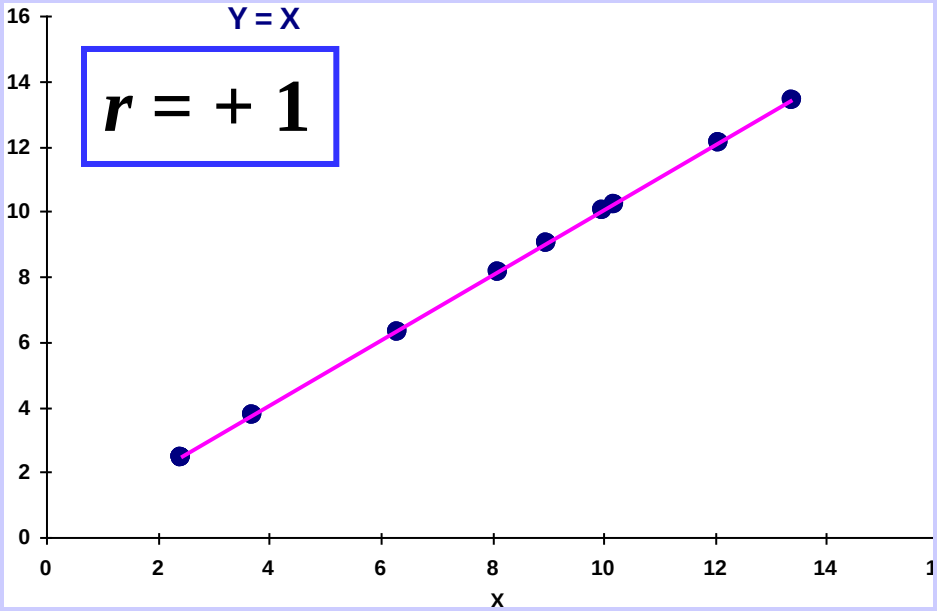
Correlación negativa

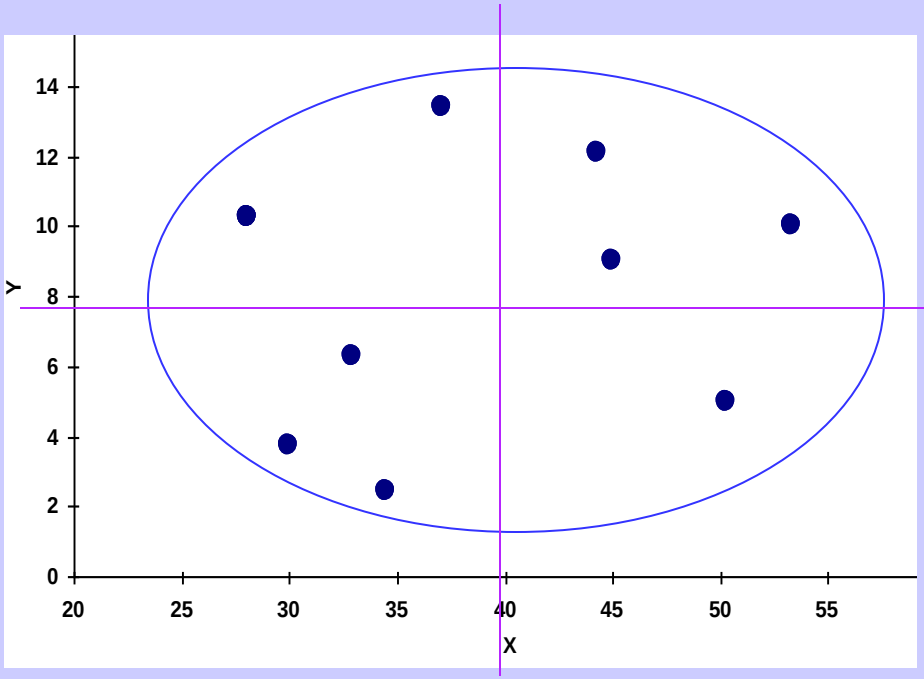


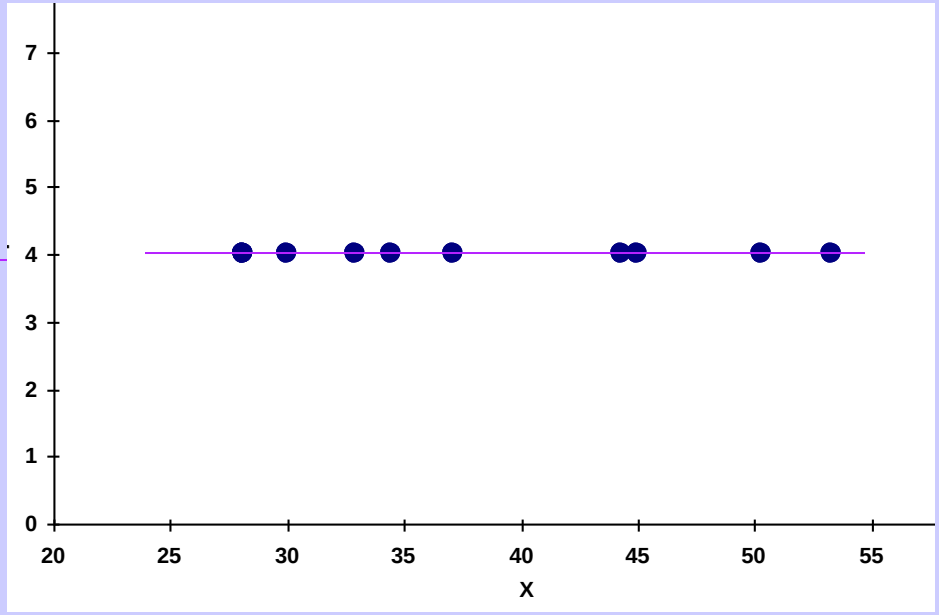


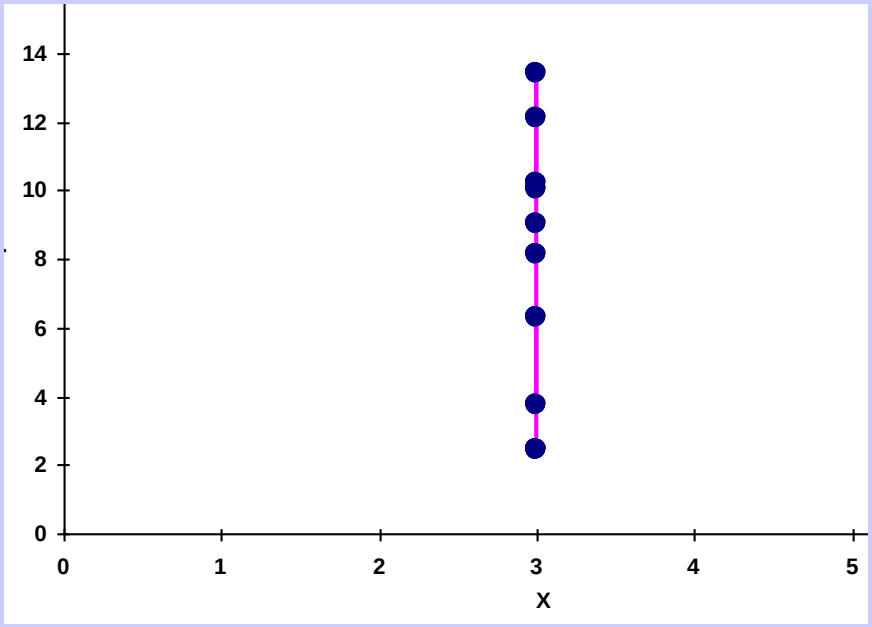
Correlación positiva







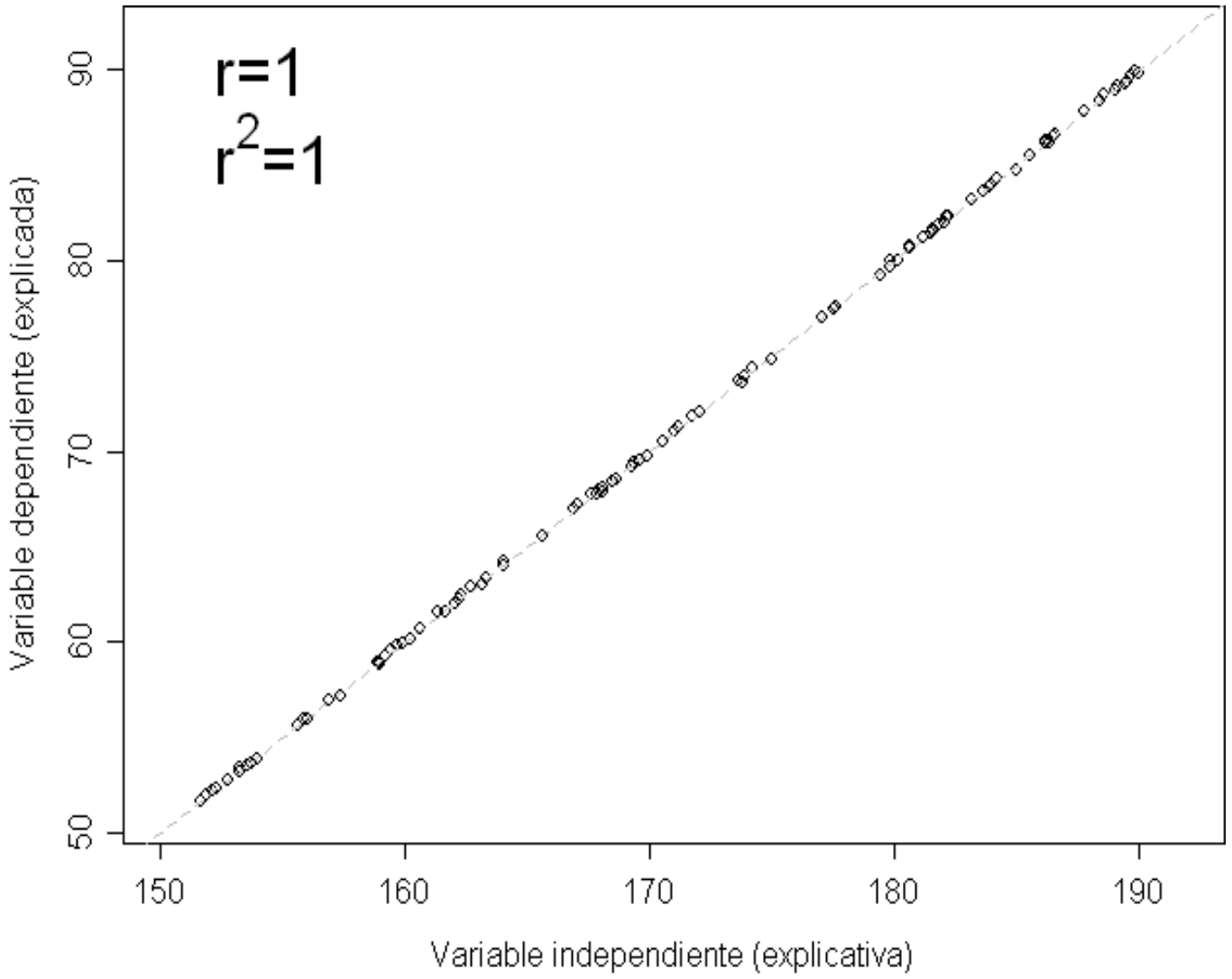




$$r = 0$$

Ausencia de correlación

Animación: Evolución de r y diagrama de dispersión



ESTIMACION DE ρ (rho)

$$r = Cov / s_x . s_y$$

PRUEBA DE

$$Ho: \rho = 0$$

$$t_{calc} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Se compara con el valor crítico (***t*** tabulado)

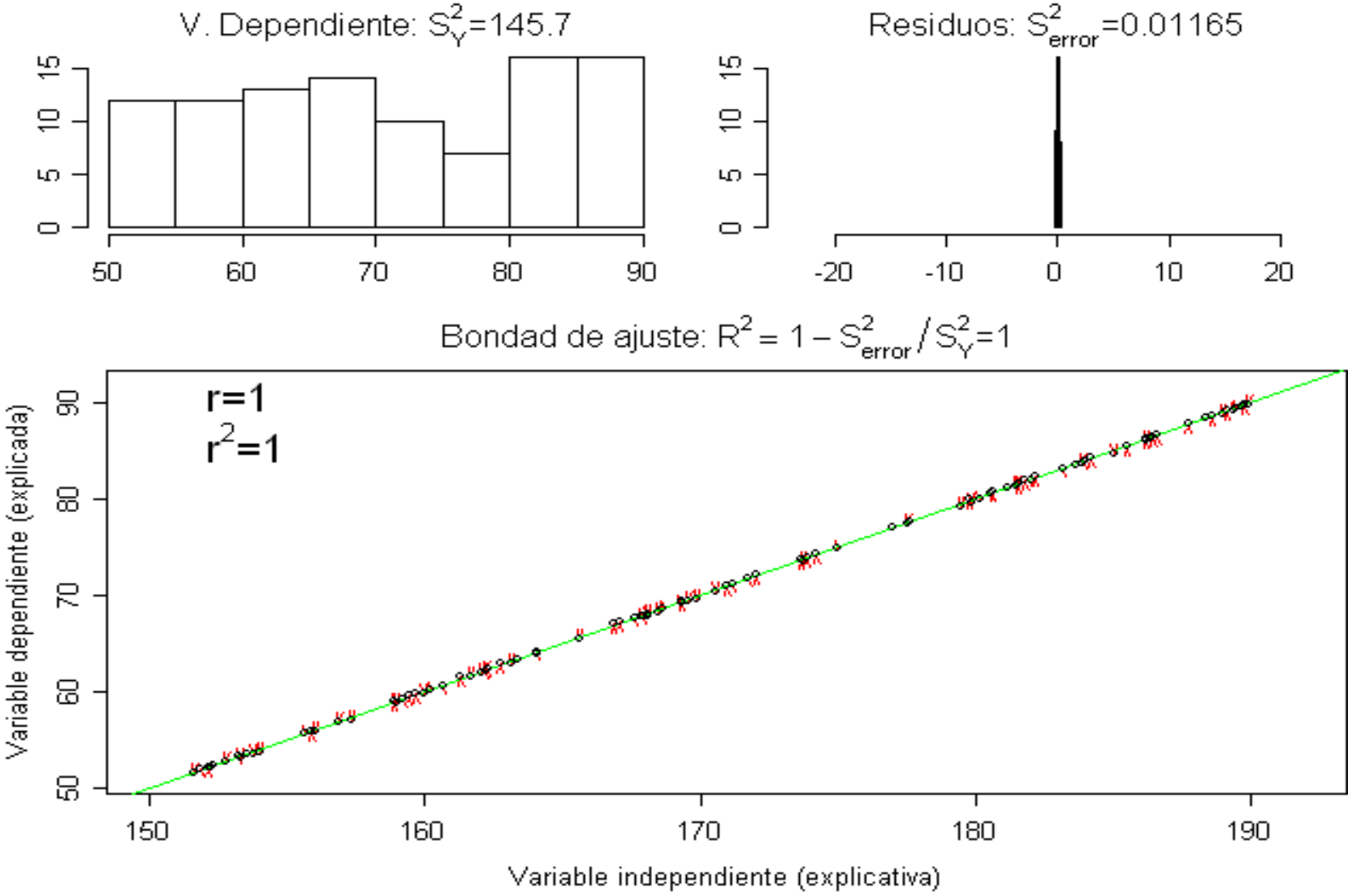
CONSIDERACIONES PARA LA VALIDEZ DEL TEST

Los residuos (e) deben ser :

- Normales
- Homocedasticos
- Independientes

Testar la Ho: $\rho = 0$ equivale a ensayar la Ho:$\beta = 0$

Animación: Residuos del modelo de regresión



ESTADISTICOS USUALES

Varianza residual (insesgada)

$$\hat{s}^2_{y.x} = \frac{\sum (y - \hat{y})^2}{n-2} = \frac{\sum \varepsilon^2}{n-2}$$

Error tipico de estimación de *y*

$$\hat{s}_{y.x} = \sqrt{\hat{s}^2_{y.x}}$$

Error tipico de estimación de *b*

$$\hat{s}_b = \hat{s}_{y.x} \Big/ \sqrt{SCX}$$

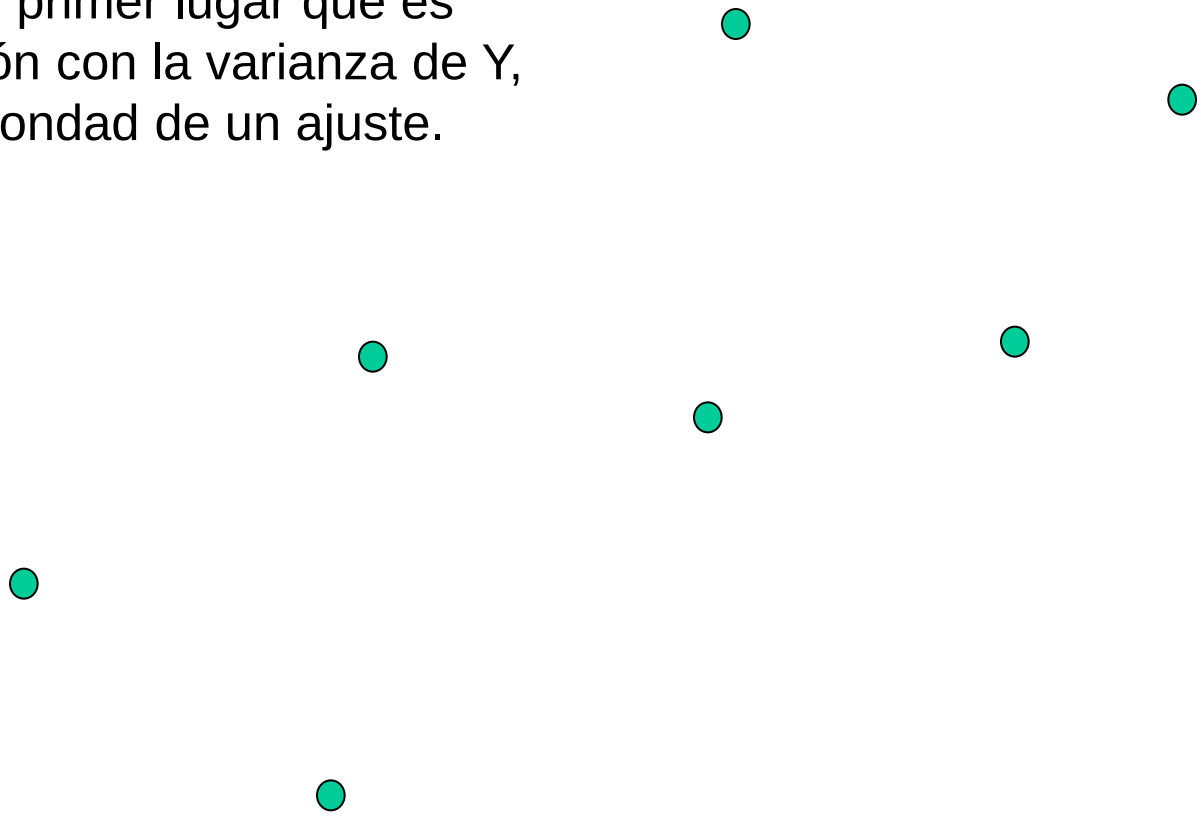
Coeficiente de Determinación R^2

$$R^2 = \frac{SCRegresión}{SCtotal} \quad (0 \le R^2 \le 1)$$

$$R^2 = 1 - \frac{S_e^2}{S_Y^2}$$

¿Cómo medir la bondad de una regresión?

Imaginemos un diagrama de dispersión, y vamos a tratar de comprender en primer lugar qué es el error residual, su relación con la varianza de Y, y de ahí, cómo medir la bondad de un ajuste.

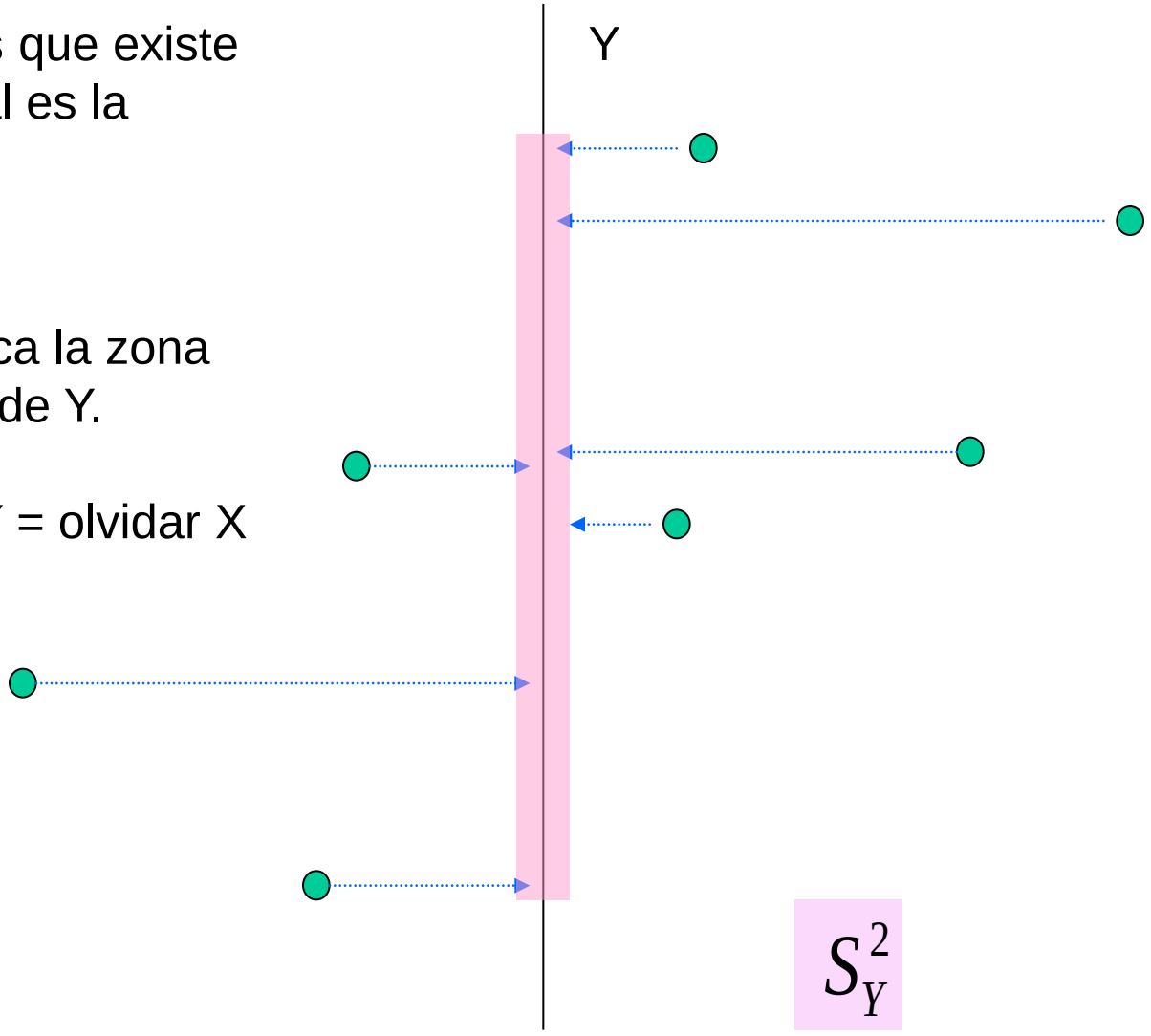


Interpretación de la variabilidad en Y

En primer lugar olvidemos que existe la variable X. Veamos cuál es la variabilidad en el eje Y.

La franja sombreada indica la zona donde varían los valores de Y.

Proyección sobre el eje Y = olvidar X

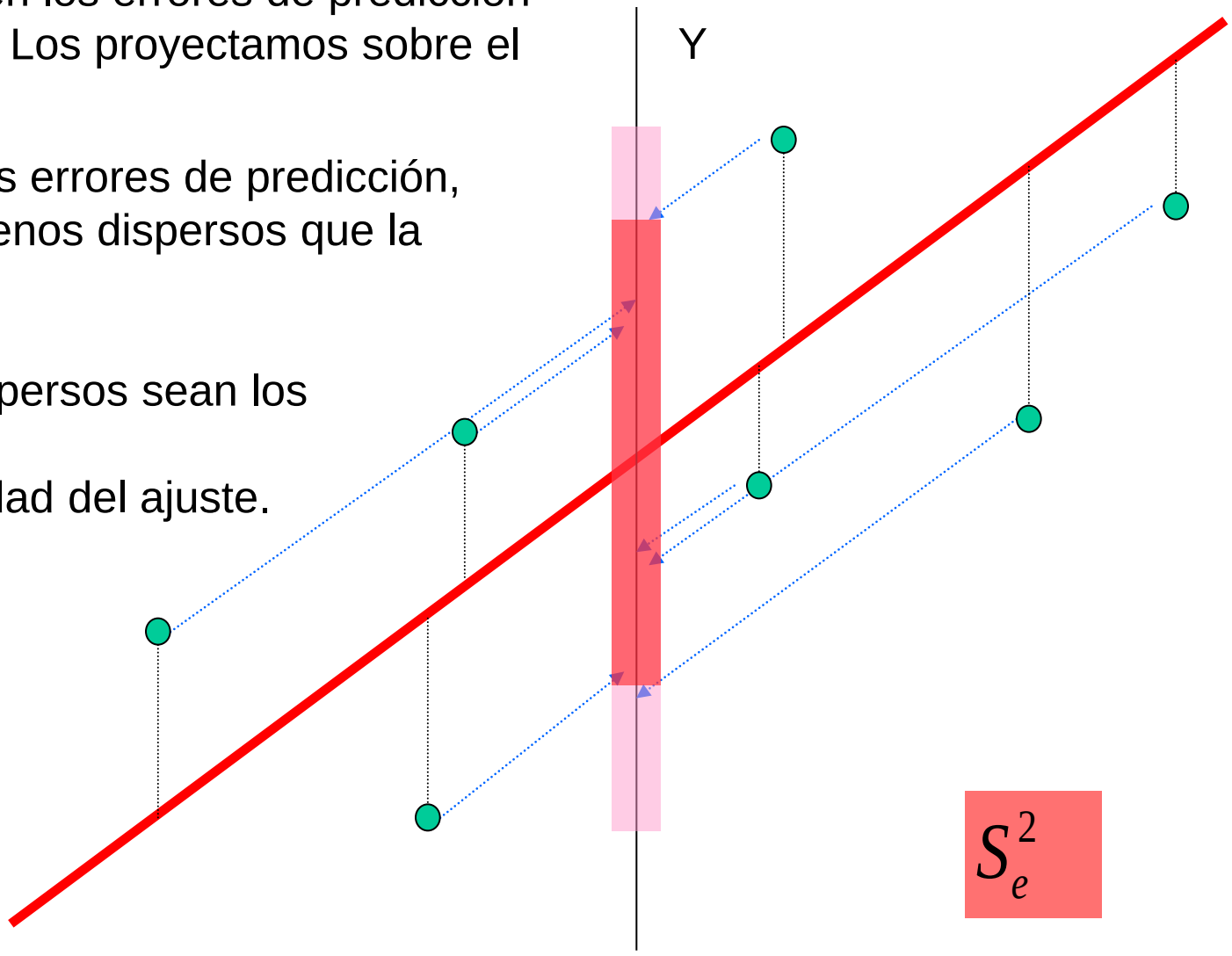


Interpretación del residuo ($y - \hat{y}$)

Fijémonos ahora en los errores de predicción (líneas verticales). Los proyectamos sobre el eje Y.

Se observa que los errores de predicción, residuos, están menos dispersos que la variable Y original.

Cuanto menos dispersos sean los residuos,
mejor será la bondad del ajuste.



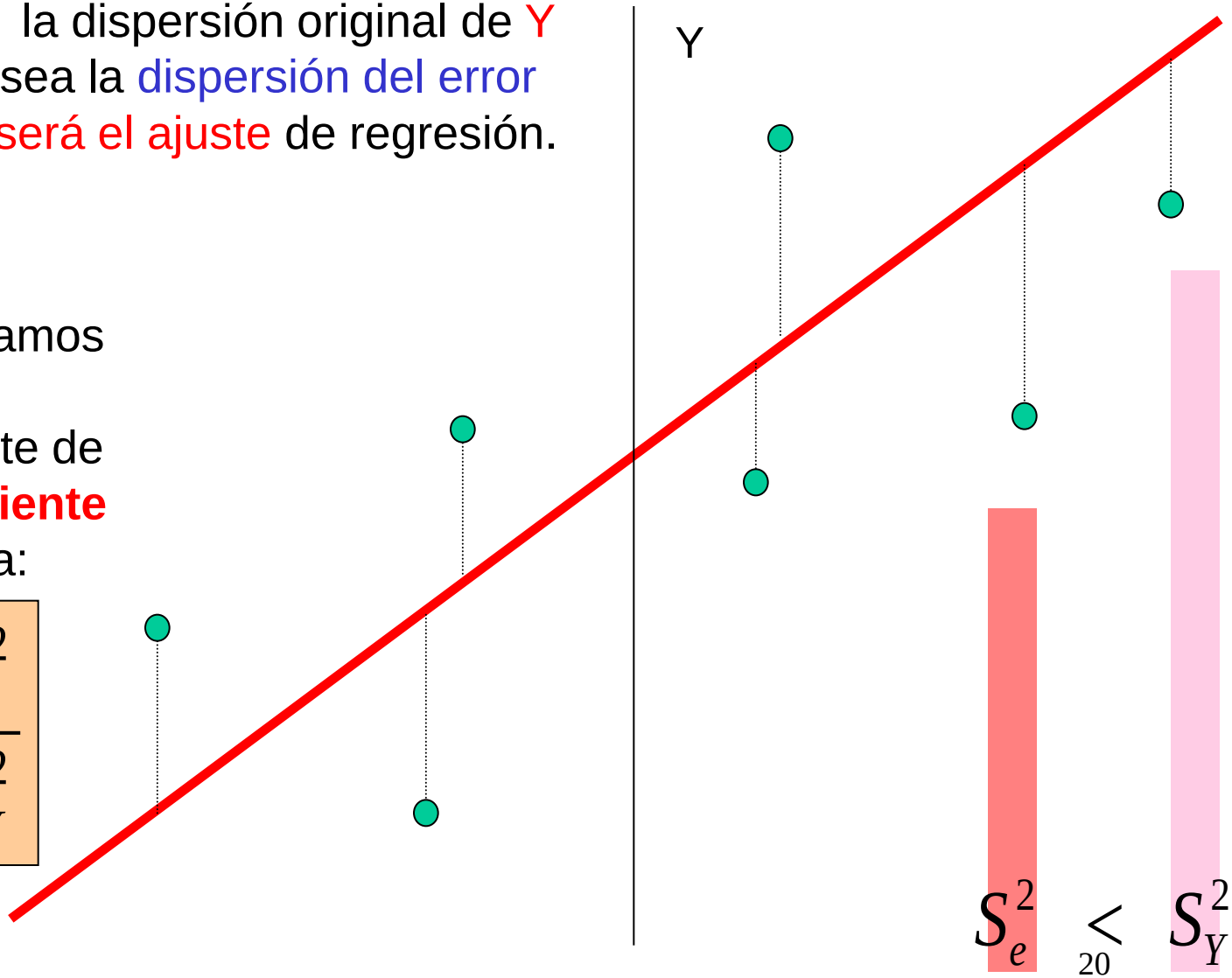
Bondad de un ajuste

Resumiendo:

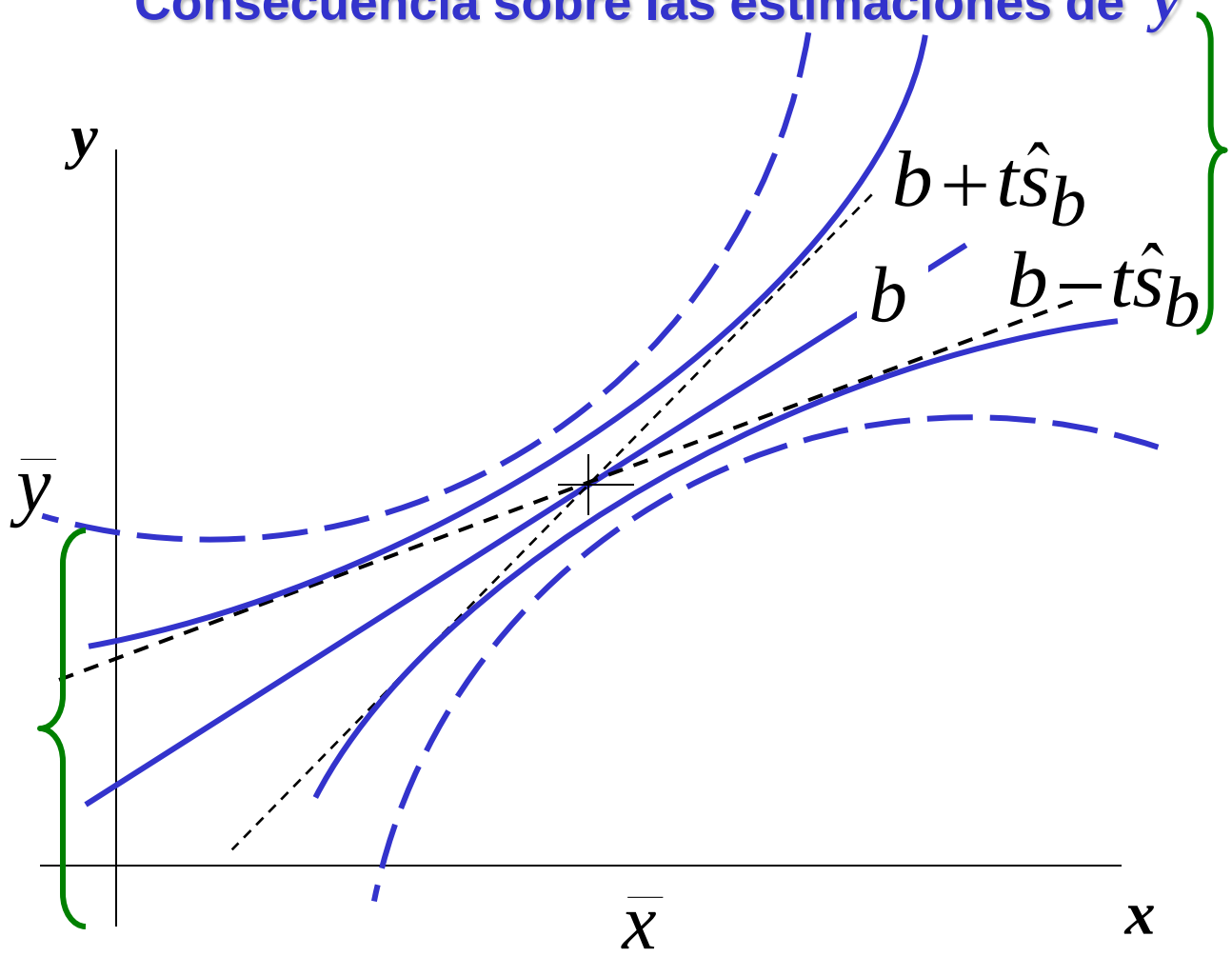
- La dispersión del error residual será una fracción de la dispersión original de Y
- Cuanto menor sea la dispersión del error residual mejor será el ajuste de regresión.

Eso hace que definamos como medida de bondad de un ajuste de regresión, o **coeficiente de determinación** a:

$$R^2 = 1 - \frac{S_e^2}{S_Y^2}$$



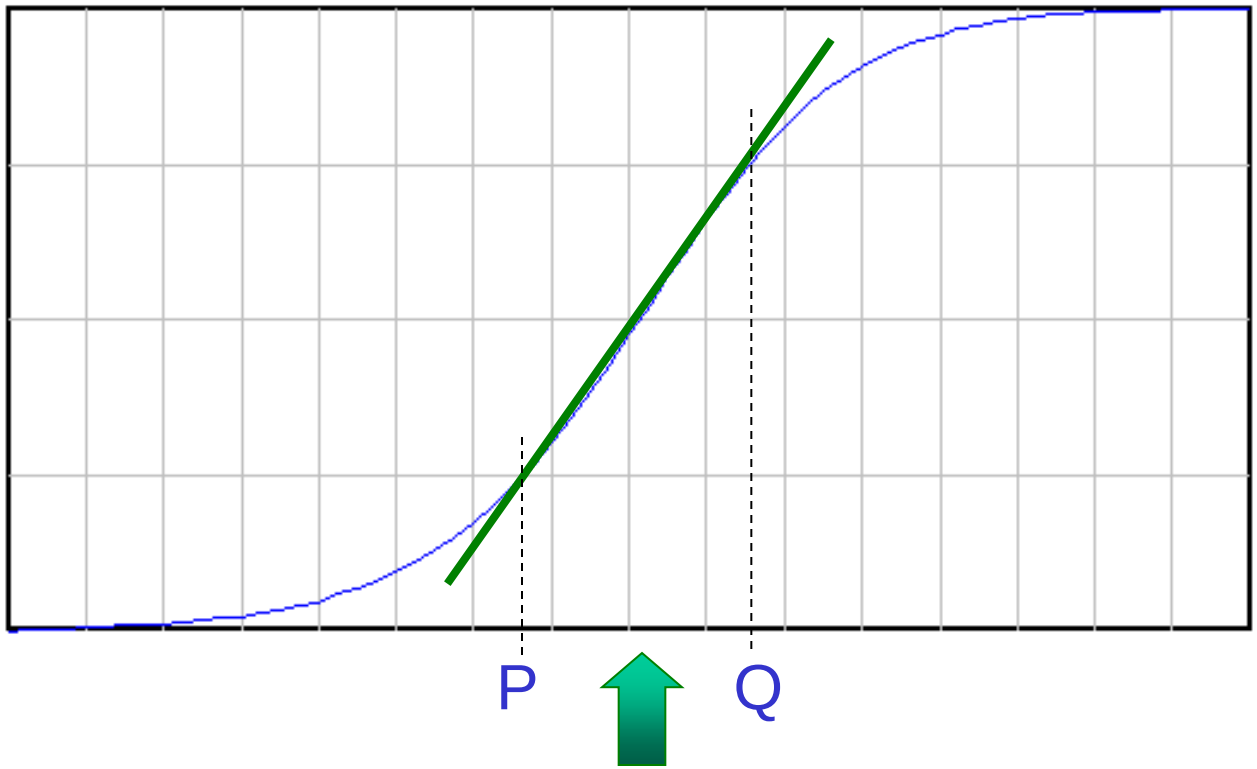
Consecuencia sobre las estimaciones de y



faja de confianza para $\mu_{y|x}$

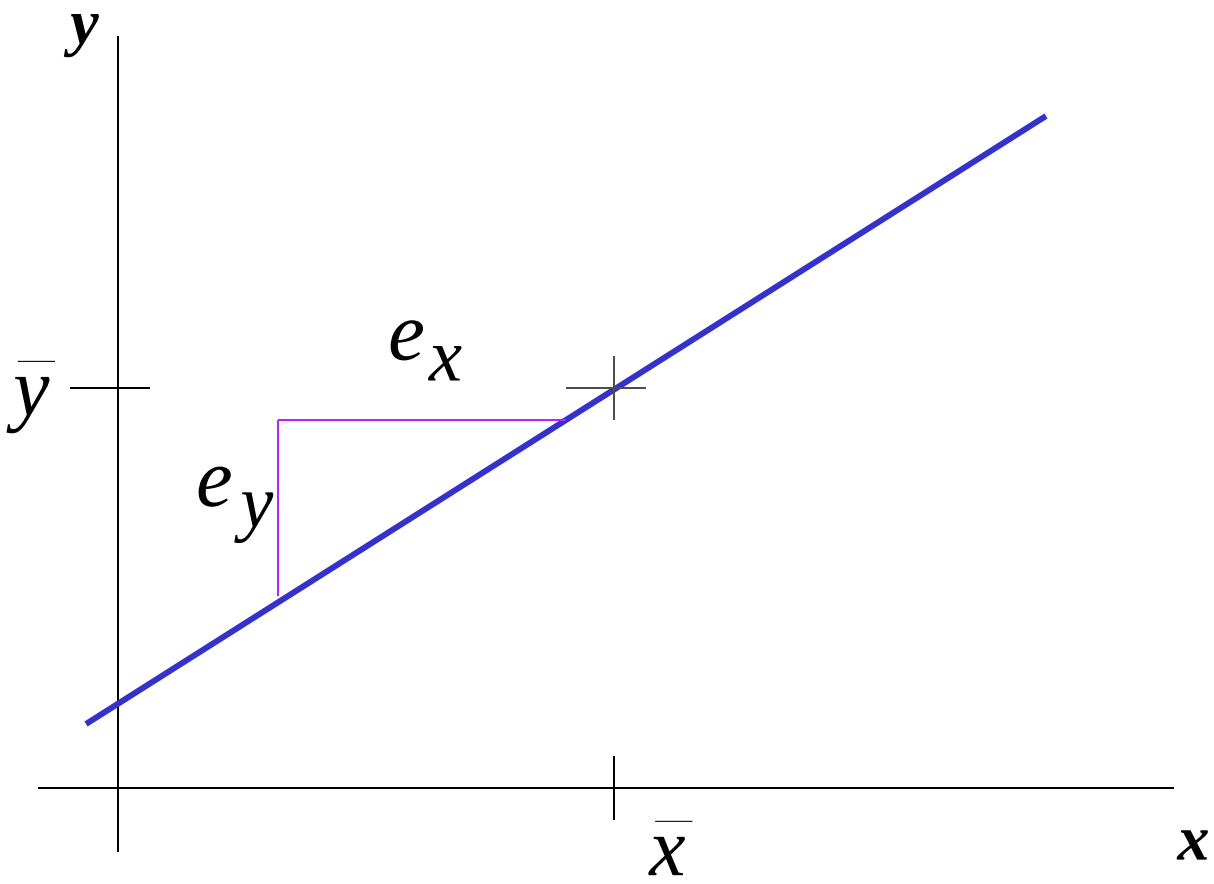
faja de confianza para $\hat{y}|x$

A medida que los valores se alejan del centroide ($\bar{x}$, $\bar{y}$)
las estimaciones de ***y*** son más imprecisas



Buen ajuste a la recta en el intervalo PQ

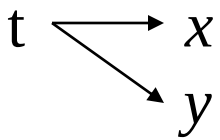
NO implica que la relación sea lineal fuera del mismo



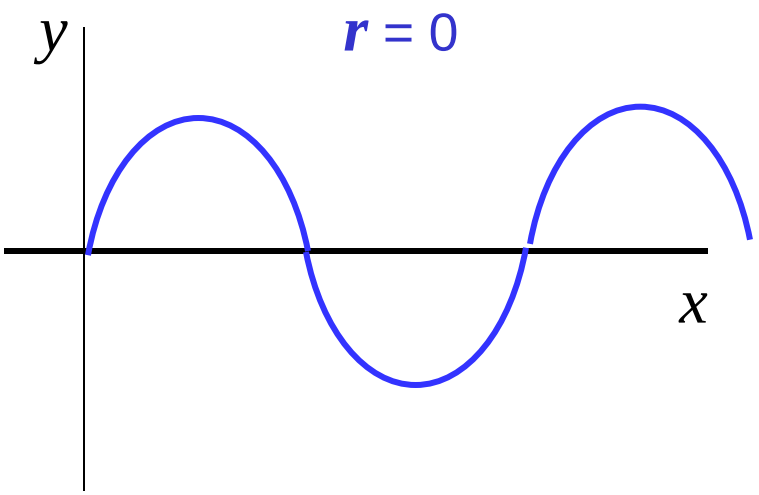
La recta de regresión de ***y*** sobre ***x*** no es la misma que la de ***x*** sobre ***y*** , salvo que todos los puntos estén sobre la recta

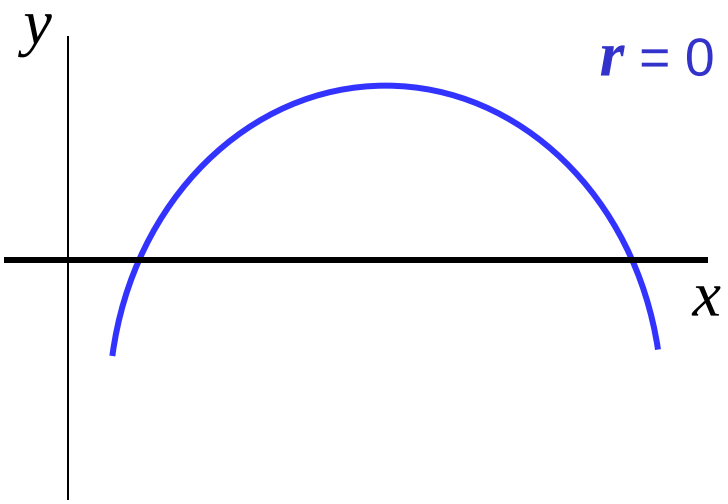
Precauciones en la interpretación de *r*

- ***r*** significativo NO implica relación de causalidad entre las variables



- ***r*** = 0 NO implica ausencia de asociación entre las variables





Los problemas de regresión y de correlación lineales se parecen pero difieren

- En la finalidad
- En las variables

REGRESION	CORRELACION
x variable independiente fija	**NO** hay distinción entre variable dependiente e independiente
y variable dependiente aleatoria	x e y son variables aleatorias

Cálculos en correlación y regresión

- Entrar ***x*** → Hallar $\bar{x}$ y s_x → Borrar la memoria estadística
- Entrar ***y*** → Hallar $\bar{y}$ y s_y → Borrar la memoria estadística
- Entrar los productos (***x y***) → Hallar $\overline{xy}$
- Calcular: $Cov = \overline{xy} - \bar{x}.\bar{y}$

$$r = Cov / s_x.s_y$$ ⇒ Testar: $Ho: \rho = 0$

$$b = rs_y / s_x$$

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

$$y = a - bx$$