SEMINARIO DE DOCTORADO

TÉCNICAS AVANZADAS DE INVESTIGACIÓN SOCIAL

Módulo 3 A
ANÁLISIS DE CORRELACIÓN
RECTA DE REGRESIÓN LINEAL
MODELO DE REGRESIÓN LINEAL

Agustín Salvia
Santiago Poy

¿Hay relación entre las variables años de estudio e ingresos laborales?

Años de estudio (años)	Ingresos ($)
5	1.700
6	2.000
7	2.300
8	2.600
9	2.900
10	3.200
11	3.500
12	3.800
13	4.100
14	4.400
16	5.000
17	5.300

CORRELACIÓN ENTRE VARIABLES CUANTITATIVAS

Se considera que dos variables cuantitativas están relacionadas entre sí cuando los valores de una de ellas varían de forma sistemática con respecto a los valores homónimos de la otra. Dicho de otro modo, si tenemos dos variables, *A* y *B*, existe relación entre ellas si al aumentar los valores de *B* también lo hacen los de *A*, o por el contrario si al aumentar los valores de *B* disminuyen los de *A*.

- **Para variables métricas, el gráfico de dispersión es la manera más intuitiva de evaluar la relación entre las dos variables, pudiendo esta adoptar diferentes formas.**

- **El método más usual para medir la fuerza de la relación lineal entre dos variables métricas es la correlación momento-producto o correlación de Pearson.**

CORRELACIÓN ENTRE VARIABLES CUANTITATIVAS

Los atributos fundamentales de una relación entre dos variables cuantitativas son:

- **Ajuste**
- **Sentido**
- **Forma**
- **Fuerza**

CORRELACIÓN ENTRE VARIABLES CUANTITATIVAS

- **El ajuste mide el nivel en que los pares de observaciones quedan representados en una línea en donde a cada valor de A le corresponde un valor en B. Si la nube de observaciones es estrecha y alargada, una línea recta representará a la nube de puntos y a la relación y por tanto ésta será fuerte.**

- **El sentido de la relación se refiere a cómo varían los valores de *B* con respecto a *A*. Si al crecer los valores de la variable *A* lo hacen los de *B*, será una relación positiva o directa. Si al aumentar *A*, disminuye *B*, será una relación negativa o inversa.**

- **La forma establece el tipo de línea a emplear para definir el mejor ajuste. Se pueden emplear tres tipos de líneas: una línea recta, una curva monotónica o una curva no monotónica.**

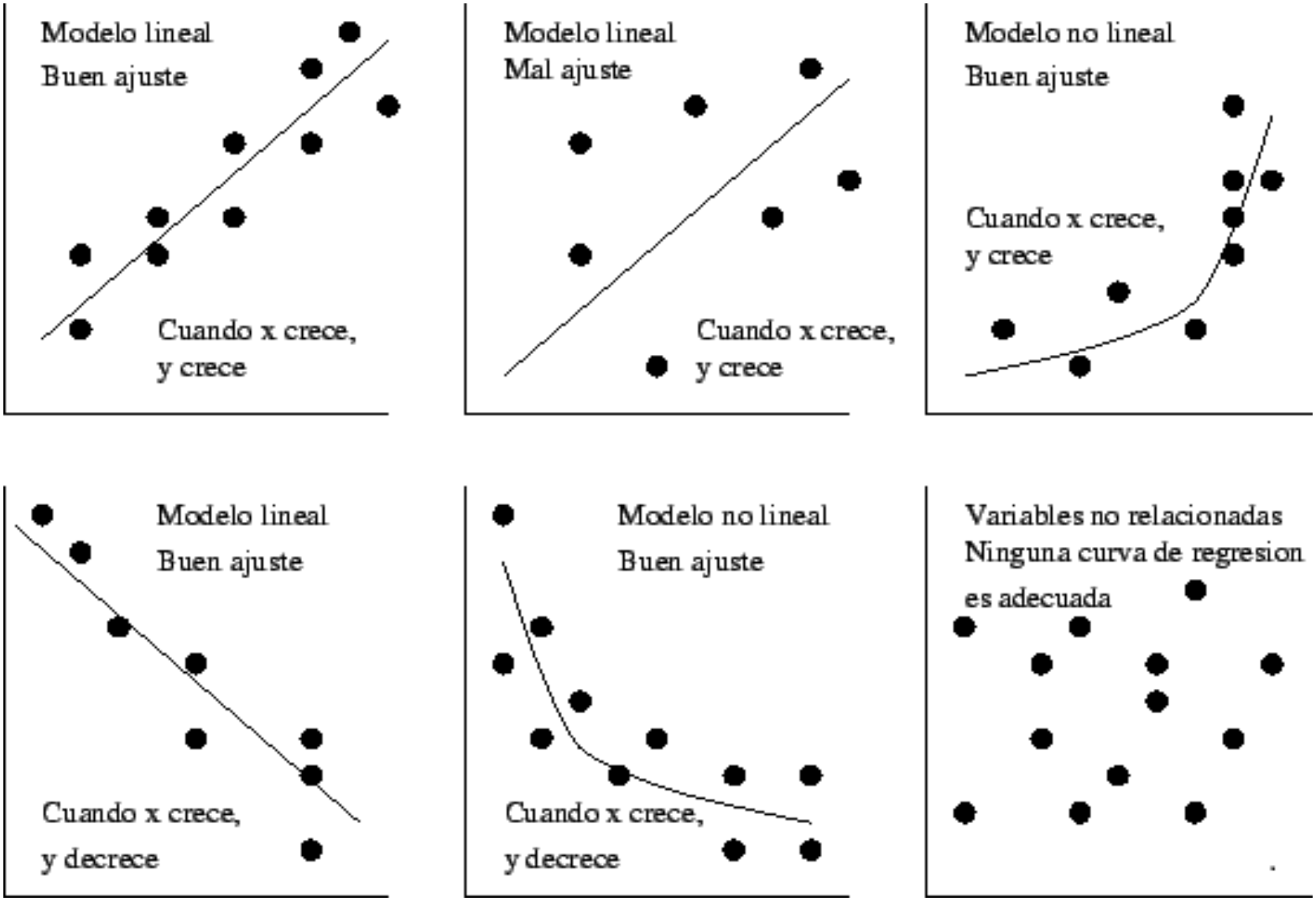
- **La fuerza es la pendiente de la recta. En cuantas unidades aumenta, o disminuye, la variable A al aumentar en una unidad la variable B.**

DIAGRAMAS DE DISPERSIÓN ESTADÍSTICA

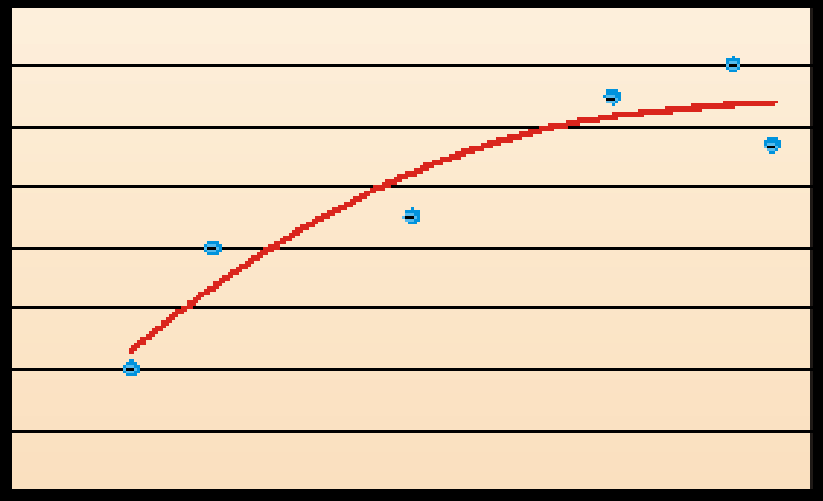
Dadas dos variables X y Y tomadas sobre el mismo elemento de la población, el diagrama de dispersión es simplemente un gráfico de dos dimensiones, donde en un eje (la abscisa) se grafica una variable (independiente), y en el otro eje (la ordenada) se grafica la otra variable (dependiente).

Si las variables están correlacionadas, el gráfico mostraría algún nivel de correlación (tendencia) entre las dos variables. Si no hay ninguna correlación, el gráfico presentará una figura sin forma, una nube de puntos dispersos en el gráfico.

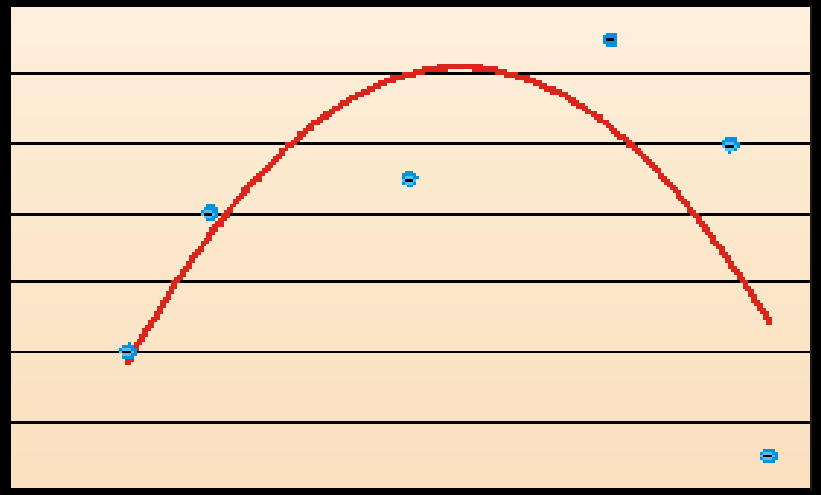
FORMAS TÍPICAS DE LOS DIAGRAMAS DE DISPERSIÓN ESTADÍSTICA



CURVA MONOTÓNICA



CURVA NO MONOTÓNICA



- **En el caso de una relación bajo la forma de una curva monotónica, la relación entre las dos variables no es constante a lo largo de toda la recta, y por lo tanto la pendiente de la misma es variable en su recorrido. Se dice que la línea de ajuste es no lineal puesto que es una curva.**

- **En el caso de una relación representada por una curva no monotónica varía tanto la pendiente de la curva como el sentido de la relación, que en unos sectores puede ser positiva (ascendente) y en otros negativa (descendente).**

EL COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

El Coeficiente de Correlación de Pearson es un índice estadístico que mide la fuerza de la relación lineal entre dos variables. El coeficiente representa cuánto de la varianza total presente entre variables se explica por la covarianza ENTRE ellas. Su resultado es un valor que fluctúa entre –1 (correlación perfecta de sentido negativo) y +1 (correlación perfecta de sentido positivo). Cuanto más cercanos al 0 sean los valores, mayor es la debilidad o ausencia de correlación.

Su cálculo se basa en la expresión:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

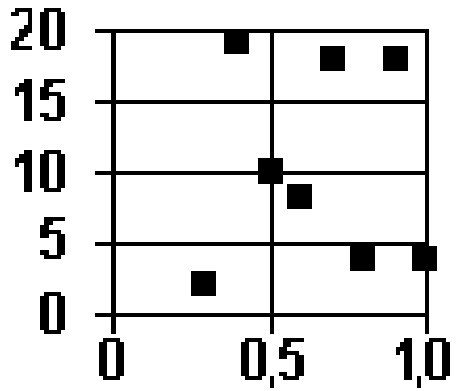
$\bar{x}, \bar{y}$ *son las medias aritméticas de x e y.*

Covarianza de las dos variables.

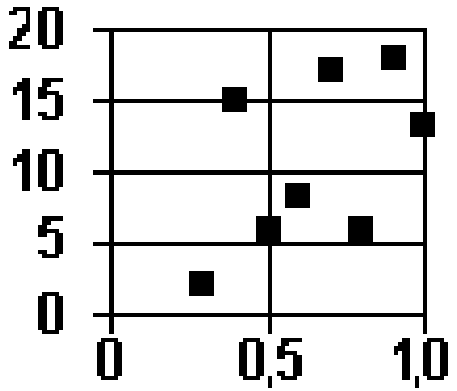
Producto de las desviaciones típicas de las dos variables.

EL COEFICIENTE DE CORRELACIÓN LINEAL DE PEARSON

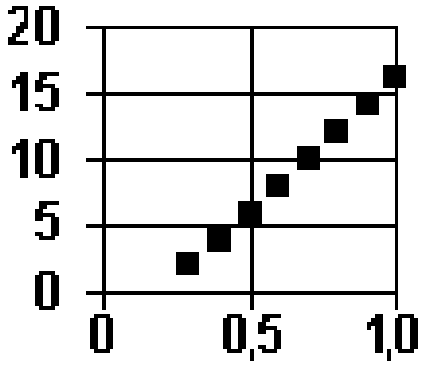
Si el coeficiente de correlación de Pearson (r) es cercano a 0, las dos variables no se relacionan de manera lineal entre sí (no tienen casi ninguna covariación *lineal*). Si su valor es cercano a +/-1, esto significa que la relación entre las dos variables es lineal y podrá ser bien representada por una línea recta.



r = 0,0



r = 0,5



r = 1,0

CORRELACIÓN LINEALES ENTRE VARIABLES CUANTITATIVAS

- **A pesar el coeficiente de Pearson sólo estima la relación entre dos variables, es fácil calcular una *matriz de correlación* entre todos los pares potenciales de variables, para luego evaluar aquellas relaciones relevantes.**

- **Un aspecto débil del análisis de correlación es que sólo detecta la parte lineal de las relaciones entre las variables. Por ejemplo, una relación que obedece a una ecuación curvilineal pasaría inadvertida.**

- **Sin embargo, las variables a evaluar pueden experimentar transformaciones que permite su "linealización", para cual resulta previamente necesario conocer la forma exacta esperada y/o observada de la relación.**

EJEMPLO CORRELACIÓN

Total Ocupados entre 25 y 45 años (con ingresos)

Correlations[a]

		Ingreso horario de la ocupación ppal	Años de estudio (aprox.)	Nivel de Instrucción	Cantidad de hijos menores de 12 años
Ingreso horario de la ocupación ppal	Pearson Correlation	1,000	,354**	,365**	-,072**
	Sig. (2-tailed)	.	,000	,000	,000
Años de estudio (aprox.)	Pearson Correlation	,354**	1,000	,945**	-,223**
	Sig. (2-tailed)	,000	.	,000	,000
Nivel de Instrucción	Pearson Correlation	,365**	,945**	1,000	-,217**
	Sig. (2-tailed)	,000	,000	.	,000
Cantidad de hijos menores de 12 años	Pearson Correlation	-,072**	-,223**	-,217**	1,000
	Sig. (2-tailed)	,000	,000	,000	.

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=10338

EJEMPLO CORRELACIÓN

Total Ocupados entre 25 y 45 años (con ingresos)
Varones

Correlations[a]

		Ingreso horario de la ocupación ppal	Años de estudio (aprox.)	Nivel de Instrucción	Cantidad de hijos menores de 12 años
Ingreso horario de la ocupación ppal	Pearson Correlation	1,000	,341**	,352**	-,071**
	Sig. (2-tailed)	,	,000	,000	,000
Años de estudio (aprox.)	Pearson Correlation	,341**	1,000	,940**	-,202**
	Sig. (2-tailed)	,000	,	,000	,000
Nivel de Instrucción	Pearson Correlation	,352**	,940**	1,000	-,191**
	Sig. (2-tailed)	,000	,000	,	,000
Cantidad de hijos menores de 12 años	Pearson Correlation	-,071**	-,202**	-,191**	1,000
	Sig. (2-tailed)	,000	,000	,000	,

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=5844

EJEMPLO CORRELACIÓN

Total Ocupados entre 25 y 45 años (con ingresos)

Mujeres

Correlations[a]

		Ingreso horario de la ocupación ppal	Años de estudio (aprox.)	Nivel de Instrucción	Cantidad de hijos menores de 12 años
Ingreso horario de la ocupación ppal	Pearson Correlation	1,000	,402**	,414**	-,075**
	Sig. (2-tailed)	,	,000	,000	,000
Años de estudio (aprox.)	Pearson Correlation	,402**	1,000	,949**	-,251**
	Sig. (2-tailed)	,000	,	,000	,000
Nivel de Instrucción	Pearson Correlation	,414**	,949**	1,000	-,251**
	Sig. (2-tailed)	,000	,000	,	,000
Cantidad de hijos menores de 12 años	Pearson Correlation	-,075**	-,251**	-,251**	1,000
	Sig. (2-tailed)	,000	,000	,000	,

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=4494

CORRELACIÓN PARCIAL

PARA IDENTIFICAR RELACIONES ESPURIAS

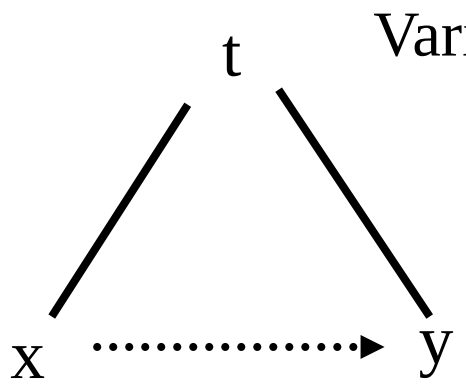
Modelo de correlación parcial

x ⟶ y

Relación original

r_{xy} = correlación simple

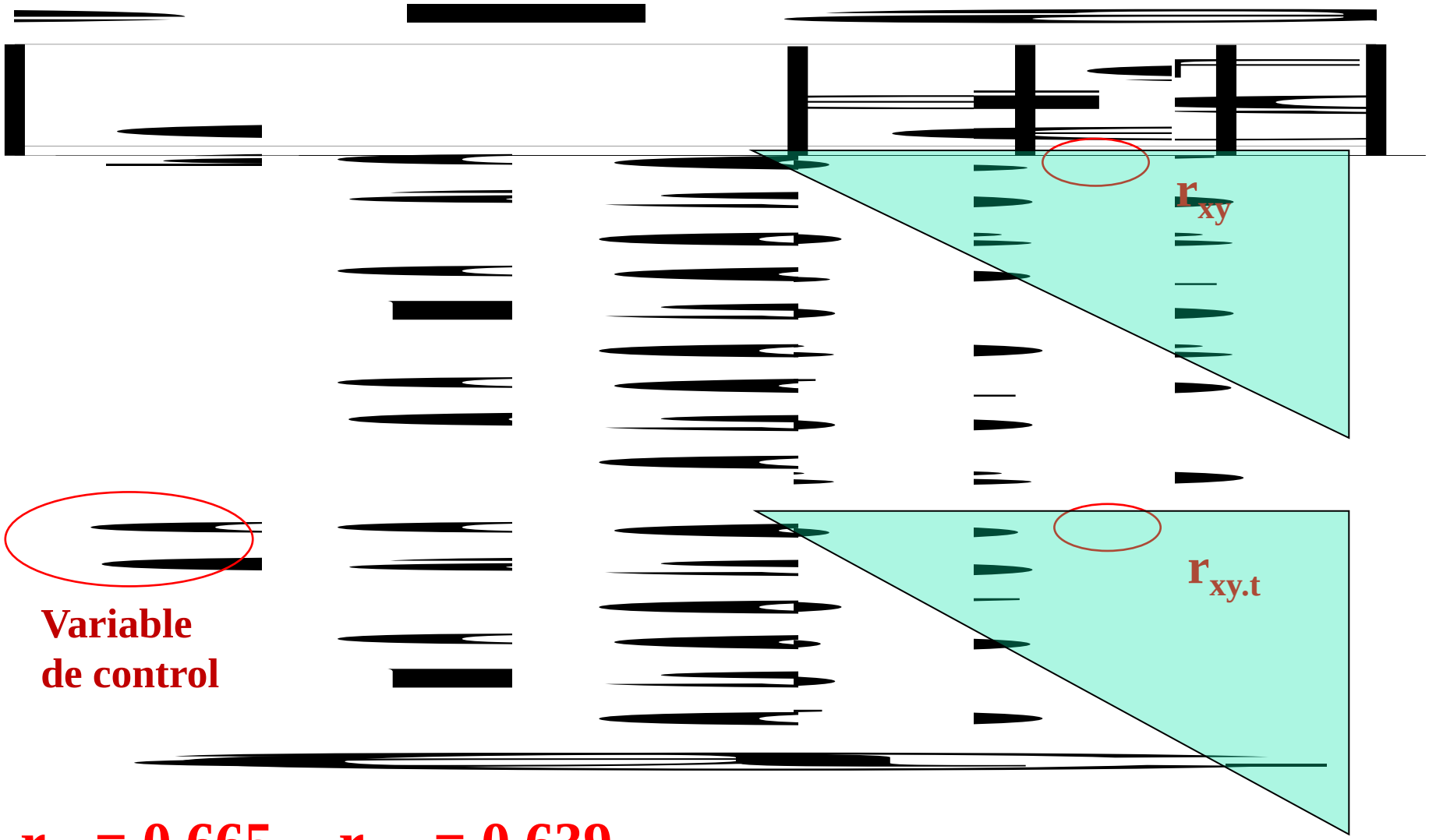
La <u>covariación</u> observada no necesariamente es <u>explicación</u>



$r_{xy.t}$ = correlación parcial

Correlación parcial es "lo que queda" de la relación original después de quitarle la incidencia de la variable de control.

Matriz de coeficientes de correlación parcial



$r_{xy} = 0{,}665$ $r_{xy.t} = 0{,}639$

Resultados de coeficientes de correlación parcial

$r_{xy} = 0{,}665$

$r_{xy.t} = 0{,}639$

$\Longrightarrow$

La variable de control presenta incidencia casi nula en la relación original: El nivel de asistencia escolar de los niños se relaciona con el nivel de privación material del hogar al que pertenecen, independientemente de la ocupación o desocupación del jefe de hogar.

MODELO ESTADÍSTICO DE REGRESIÓN LINEAL

¿Los años de estudio e ingresos determinan el valor de los ingresos laborales?

Años de estudio (años)	Ingresos ($)
5	1.700
6	2.000
7	2.300
8	2.600
9	2.900
10	3.200
11	3.500
12	3.800
13	4.100
14	4.400
16	5.000
17	5.300

Modelos de Regresión Lineal

Problemas de determinación

- **El investigador suele tener razones teóricas o prácticas para creer que una determinada variable es dependiente de una o más variables distintas.**
- **Si hay suficientes observaciones empíricas sobre estas variables, el análisis de regresión es un método apropiado para describir la estructura, fuerza y sentido exacto de esta asociación.**

Modelos de Regresión Lineal

Problemas de Causalidad

- **El modelo permite diferenciar variables explicativas, independientes o predictivas (métricas), variables a explicar o dependientes, y variables control o intervinientes (métricas o transformadas en variables categoriales).**
- **La distinción entre variables dependientes e independientes debe efectuarse con arreglo a fundamentos teóricos, por conocimiento o experiencia y estudios anteriores.**

$$Y : f(X, \epsilon) \;/\; Y = a + bX + e$$

Modelos de Regresión Lineal

Función Lineal de Regresión

El objetivo de la técnica de regresión es estimar la relación estadística que existe entre la variable *dependiente* (*Y*) y una o más variables *independientes* (X_1, X_2,... X_n). Para poder realizar esto, se postula una relación funcional entre las variables. Debido a su simplicidad analítica, la forma que más se utiliza en la práctica es la relación *lineal*:

$$\hat{y} = b_0 + b_1x_1 + \ldots b_nx_n$$

donde los coeficientes b_0 y b_1, ... b_n, son los factores que definen la variación promedio de y, para cada valor de x. Estimada esta función teórica a partir de los datos, cabe preguntarse qué tan bien se ajusta a la distribución real.

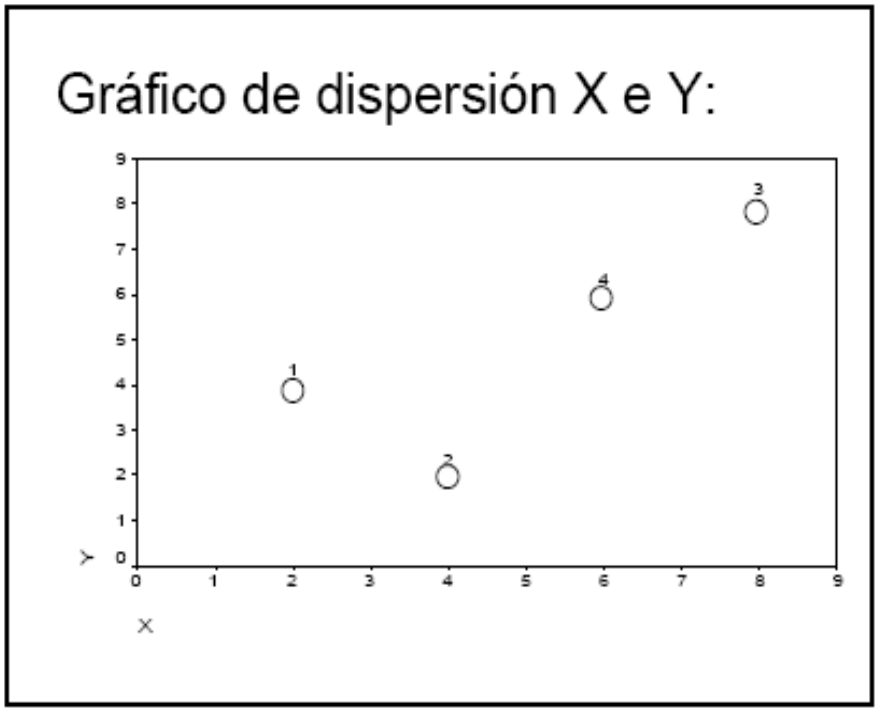
Modelos de Regresión Lineal

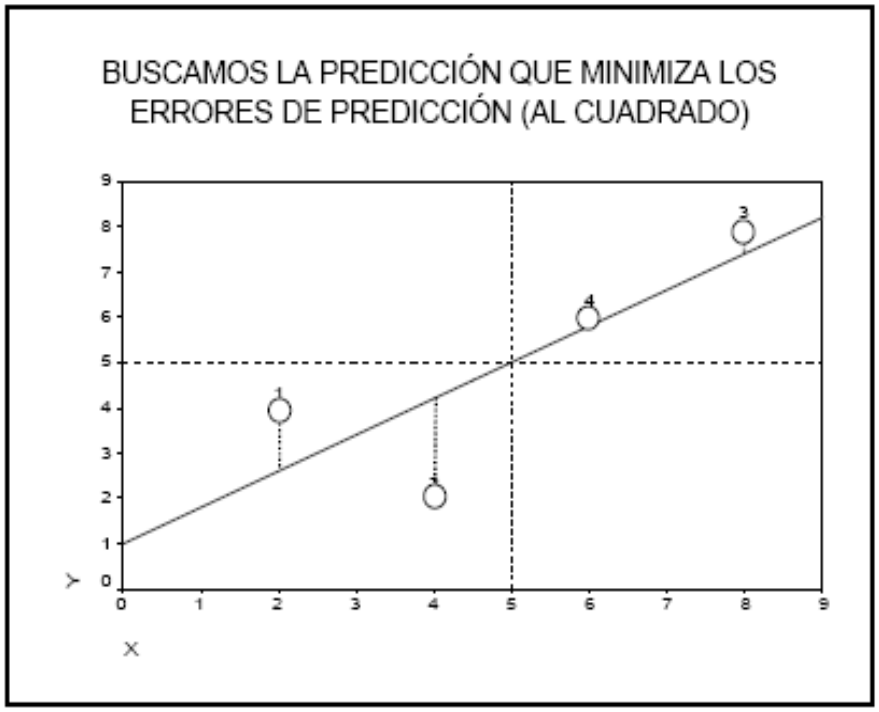
Respuestas Metodológicas

- **Estima el grado de ajuste o de bondad explicativa del modelo teórico independientemente del nivel de covarianza entre las variables introducidas**
- **Predice el valor medio que puede asumir la variable Y dado un valor de X (regresión a la media) bajo un intervalo de confianza**
- **Estima el efecto neto / fuerza / sentido de cada una de las variables predictoras de la variable dependiente (control sobre los demás efectos suponiendo independencia entre ellas).**

GRÁFICOS DE DISPERSIÓN / PENDIENTE DE LA RECTA

• En el caso de asumir una recta, se admite que variando en una unidad la variable X se registra un cambio constante en los valores de Y. A ese factor de ajuste entre ambas series se le llama pendiente de la recta, y se asume que es constante a lo largo de toda la recta.





GRÁFICOS DE DISPERSIÓN / RECTA DE REGRESIÓN

Para el cálculo de la recta de regresión se aplica el método de mínimos cuadrados entre dos variables. Esta línea es la que hace mínima la suma de los cuadrados de los residuos, es decir, es aquella recta en la que las diferencias elevadas al cuadrado entre los valores calculados por la ecuación de la recta y los valores reales de la serie, son las mínimas posibles.

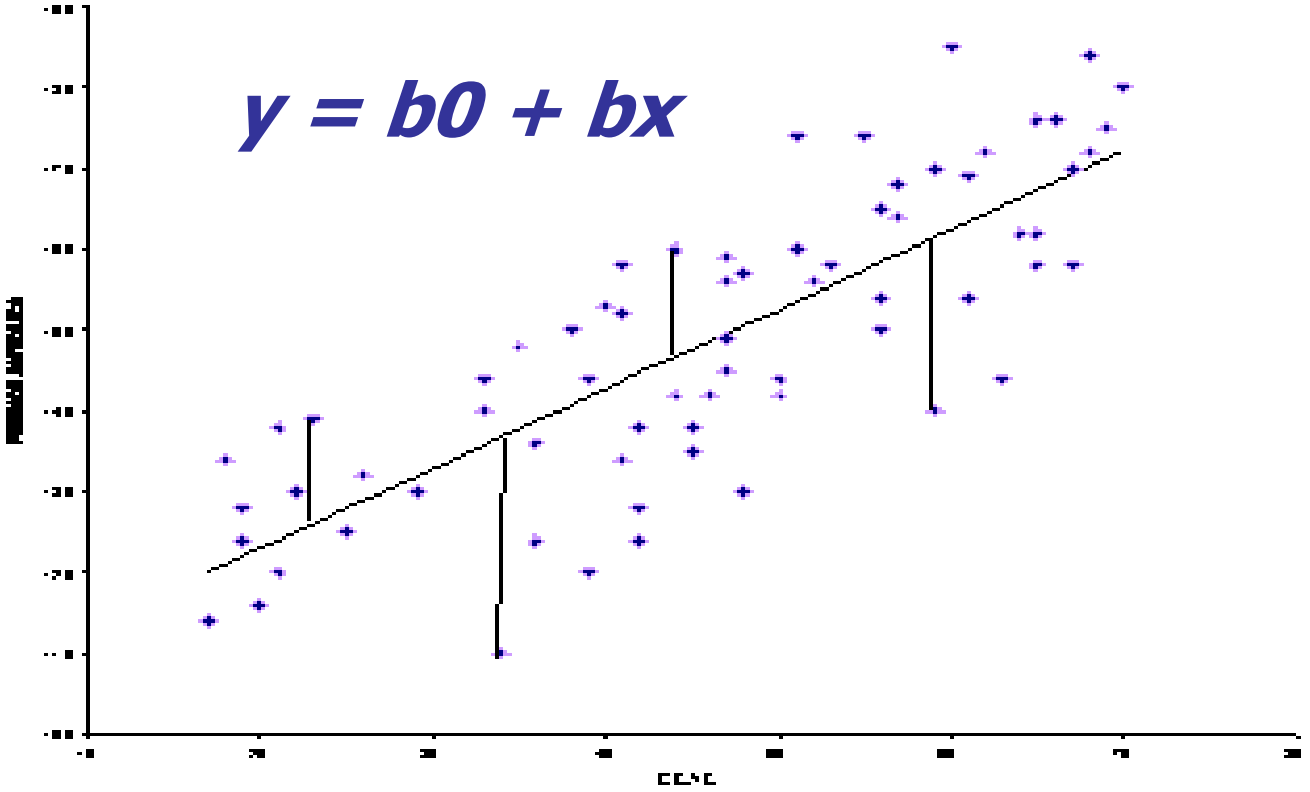


Diagrama de dispersión años de estudio e ingresos

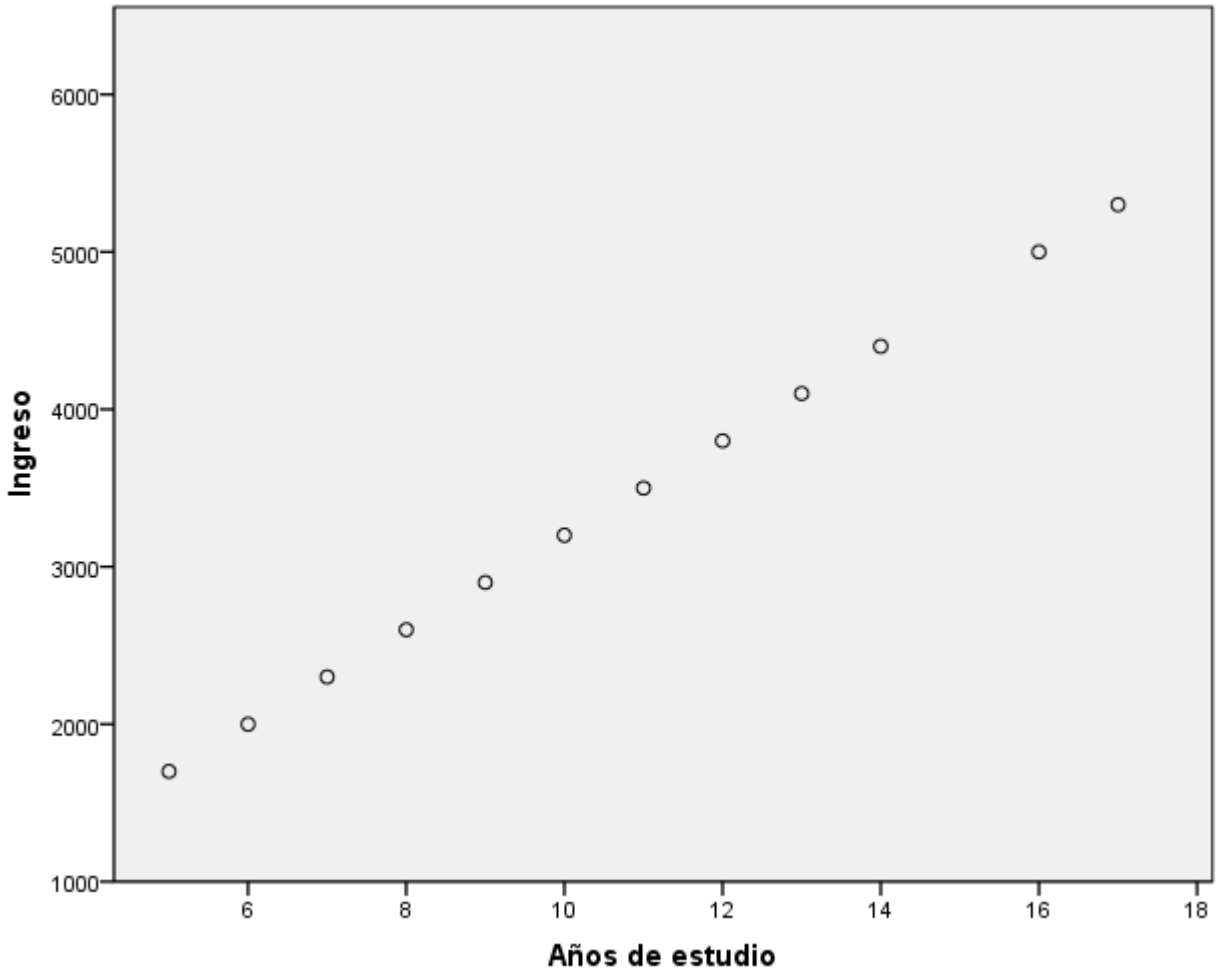
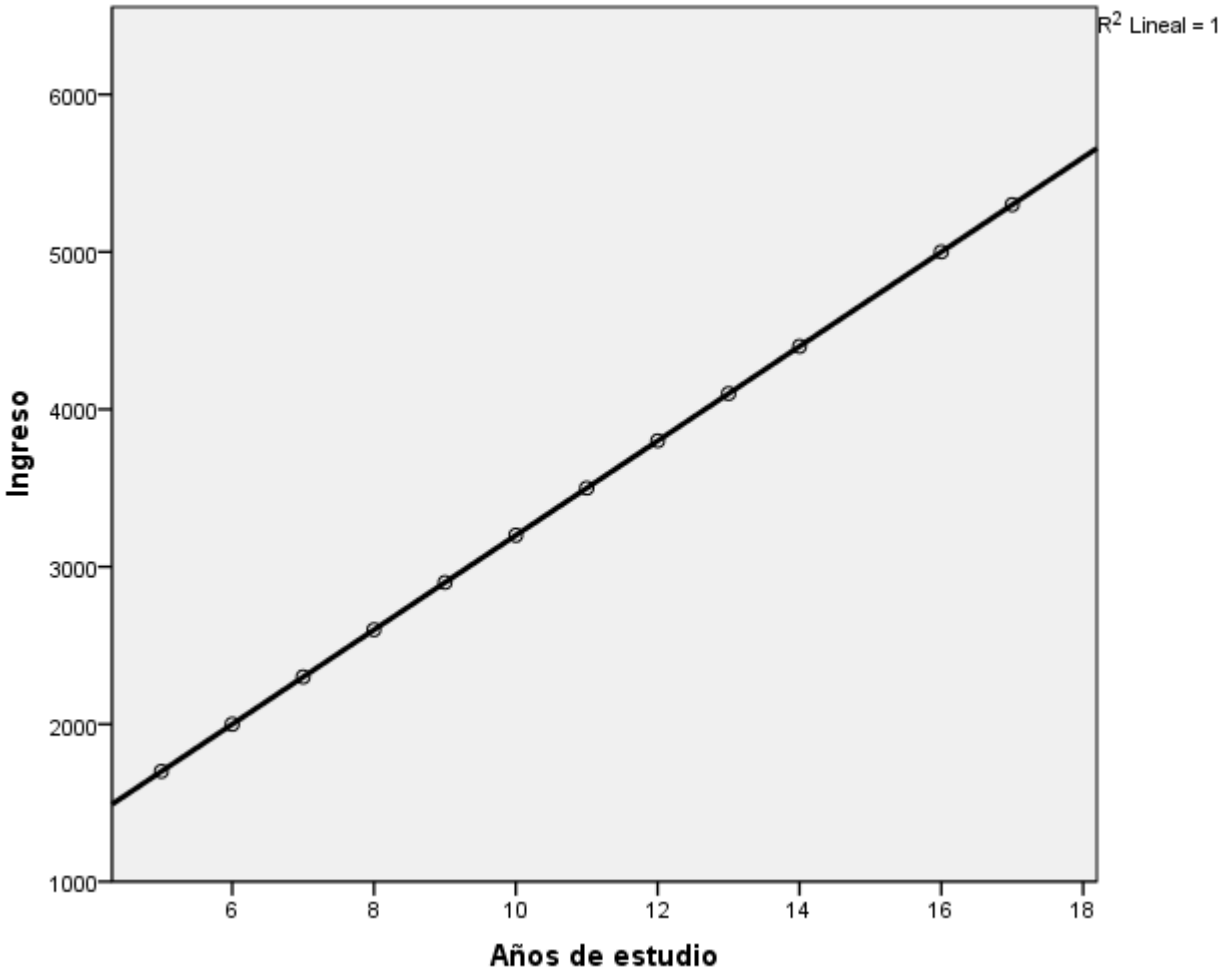
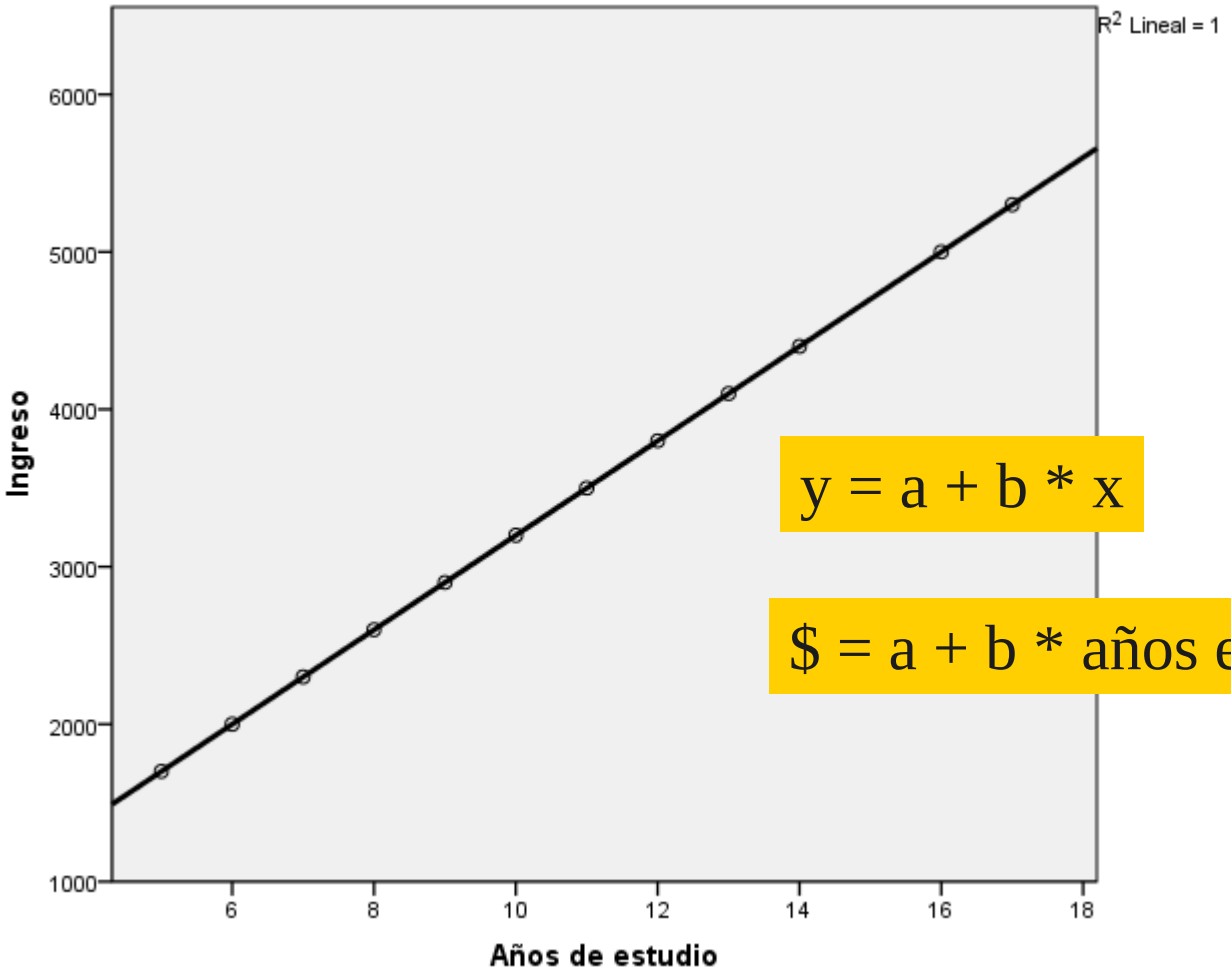


Diagrama de dispersión años de estudio e ingresos



Recta de regresión

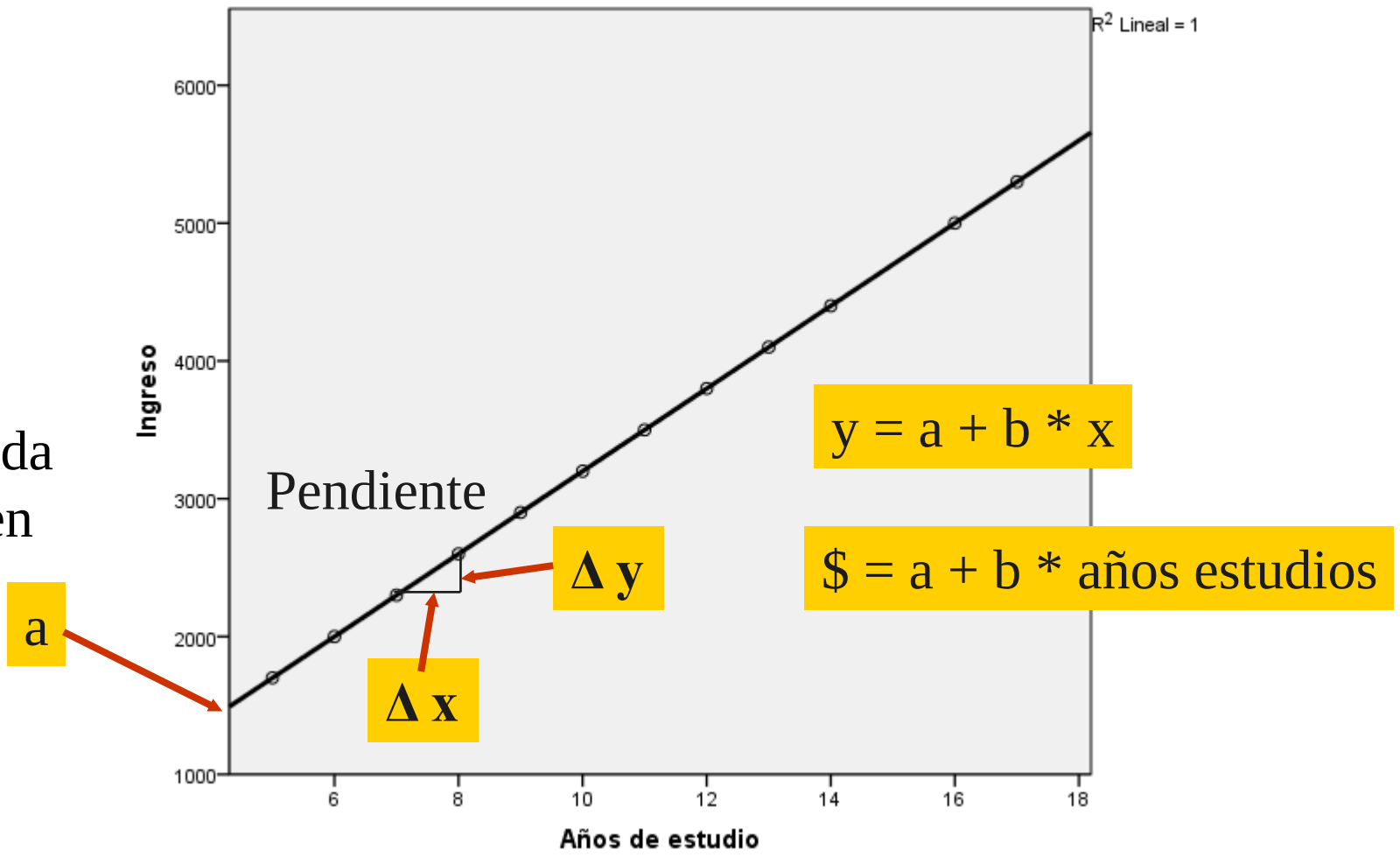


$y = a + b * x$

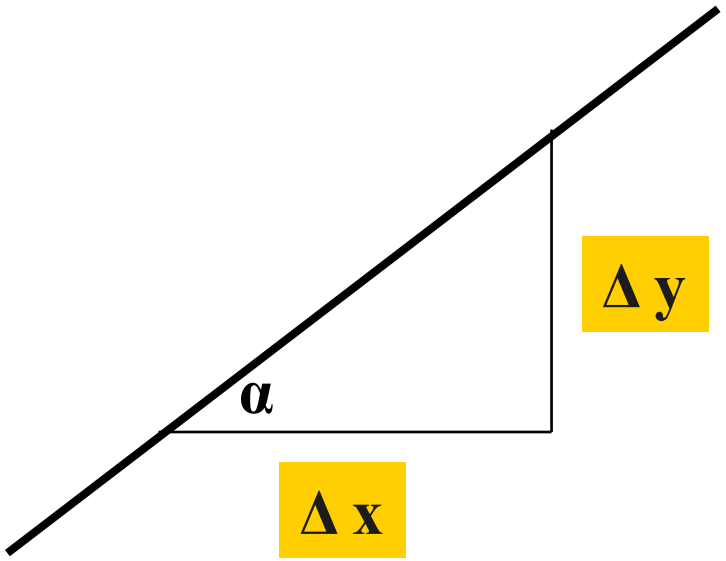
$\$ = a + b * \text{años estudios}$

Particularidades de recta de regresión

Ordenada al origen



Pendiente de recta de regresión



$$b = \operatorname{tg} \alpha = \frac{\Delta y}{\Delta x}$$

Modelos de Regresión Lineal

Función Lineal de Regresión

- El parámetro b_0, conocido como la "ordenada en el origen," nos indica cuánto vale *Y* cuando *X* = 0. El parámetro b_1, conocido como la "pendiente," nos indica cuánto aumenta *Y* por cada aumento en *X*.

- La técnica consiste en obtener estimaciones de estos coeficientes a partir de una muestra de observaciones sobre las variables *Y* y *X*.

- En el análisis de regresión, estas estimaciones se obtienen por medio del método de *mínimos cuadrados*. Logradas estas estimaciones se puede evaluar la bondad de ajuste y significancia estadística.

Modelos de Regresión Lineal

Supuestos Estadísticos del Método

- **Se asume que la forma funcional que relaciona la variable DEPENDIENTE con la/las variables explicativas es de tipo LINEAL.**

- **Las variables explicativas deben ser entre sí INDEPENDIENTES, la varianza de los errores constante, con distribución normal y los errores no deben estar correlacionados.**

- **La CONSTANTE (b_0) no sólo expresa el valor estimado de *y* en la ordenada al origen, sino también el conjunto de los errores no lineales y desconocidos del modelo.**

Modelos de Regresión Lineal

Requisitos del Método de Regresión

- **Linealidad: La relación debe ser lineal directa o inversa, y los valores observados deben quedar claramente ajustados sobre una recta.**
- **Distribución normal de errores: La variable aleatoria ϵ (error) entre los valores Y observados y los Y esperados debe ser independiente de los valores de X, y tales errores deben tener una distribución normal.**
- **No correlación de errores: Cualquier par de errores, ϵ_i y ϵ_j deben ser estadísticamente independientes entre sí, es decir que su covarianza sea igual a 0.**
- **Homocedasticidad: Las variables aleatorias ϵ_j deben tener una varianza finita σ^2 que sea constante para todos los valores de x_j .**
- **En un modelo de regresión múltiple se agrega el supuesto de variables independientes no correlacionadas.**

ESTADÍSTICOS DEL MODELO DE REGRESIÓN AJUSTADO

Modelos de Regresión Lineal

Función Lineal de Regresión

Una pregunta importante que se plantea en el análisis de regresión es la siguiente: ¿Qué parte de la variación total en *Y* se debe a la variación en *X*? ¿Cuánto de la variación de Y no se explica por X?

El estadístico que mide esta proporción o porcentaje se denomina coeficiente de determinación (R^2). Si, por ejemplo, al hacer los cálculos respectivos (elevar al cuadro el R de correlación de Pearson) se obtiene un valor de 0.846, esto significa que el modelo explica el 84.6 % de la variación de la variable dependiente.

Modelos de Regresión Lineal

Salidas Estadísticas del Método

- **Se evalúa la fuerza y la bondad de ajuste del modelo teórico a a través del coeficiente de determinación R^2**
- **La capacidad explicativa del modelo se hace a partir del método de mínimos cuadrados (ANOVA), cuyo resultado es testeado a través de F de Fisher**
- **Predice los valores de la variable dependiente a partir de estimar el valor del coeficiente (B), el error estándar (S) y el coeficiente R parcial (BETA) de cada una de las variables y de la Constante**
- **Mide la fuerza, sentido y significancia estadística de las variables del modelo sobre la variable dependiente a través de la prueba t de Student**

Modelo ajustado y recta de regresión

r de Pearson

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado corregida	Error típ. de la estimación
1	,965[a]	,931	,930	59,48101

a. Variables predictoras: (Constante), VAR00001

ANOVA[b]

Significancia

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	5035914,380	1	5035914,380	1423,382	,000[a]
	Residual	375027,055	106	3537,991		
	Total	5410941,435	107			

a. Variables predictoras: (Constante), VAR00001

b. Variable dependiente: VAR00002

"a" Ordenada al origen

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	186,277	11,346		16,418	,000
	VAR00001	28,411	,753	,965	37,728	,000

"b" Pendiente

a. Variable dependiente: VAR00002