

MODELOS LINEALES



Propósitos y contenidos del curso

Propósito del curso

Conocer y profundizar en los fundamentos básicos, propiedades y aplicaciones de los modelos lineales.

Contenidos

- Modelo general de regresión lineal
- Verificación de supuestos estadísticos para el modelo de regresión
- Análisis de Datos influyentes y datos atípicos.
- Estimación y Predicción
- Extensiones del modelo de regresión: Modelos polinomiales.

Fechas de las sesiones de clase

El curso de modelo lineales se desarrollará en 5 sesiones (10 horas cada sesión):

- Sesión 1 (septiembre 13 y 14)
- Sesión 2 (septiembre 20 y 21)
- Sesión 3 (septiembre 27 y 28)
- Sesión 4 (octubre 04 y 05)
- Sesión 5 (octubre 18 y 19)

Estrategias de evaluación

El curso de modelos lineales tendrá 3 componentes de evaluación: evaluación continua, donde se espera que el estudiante participe activamente en las actividades y talleres propuestos durante las sesiones de clase. Además, realizará trabajos prácticos durante la semana con el fin de repasar y reafirmar los conceptos vistos durante las sesiones y finalmente, la asistencia y participación en las clases.

La nota del curso estará compuesta de la siguiente manera:

- Evaluación continua (actividades durante las sesiones de clase): 60%
- Trabajo práctico (actividades que se desarrollarán desde casa): 30%
- Asistencia y participación en clase: 10%

La profesora ...

Licenciada en Matemáticas con Maestría y Doctorado en Estadística y una estancia posdoctoral en Estadística. Me he desempeñado como Profesora Universitaria en distintas Instituciones de Educación Superior (IES) del país y del extranjero. En la actualidad me encuentro adscrita a la Universidad Autónoma de Bucaramanga (UNAB) en el Departamento de Ciencias Básicas. He publicado artículos en revistas académicas indizadas en *Scopus* y en Publindex, tales como Formación Universitaria, Interciencia, Investigación Operacional y Cuadernos Latinoamericanos de Administración. Investigadora asociada de Minciencias según la última convocatoria de investigadores. Directora del Grupo de Investigación en Ciencias Aplicadas (GINCAP) categorizado como B. Líder del semillero de investigación en Aprendizaje Estadístico y Ciencia de datos (SINAPEC).

Google Scholar: <https://scholar.google.es/citations?hl=es&user=NYm-9GQAAAAJ>

ORCID: <https://orcid.org/0000-0002-4635-8003>

Scopus: <https://www.scopus.com/authid/detail.uri?authorId=57195715932>

La primera sesión está compuesta por:

- Importancia de los modelos lineales
- Correlación
- Modelo de regresión lineal simple

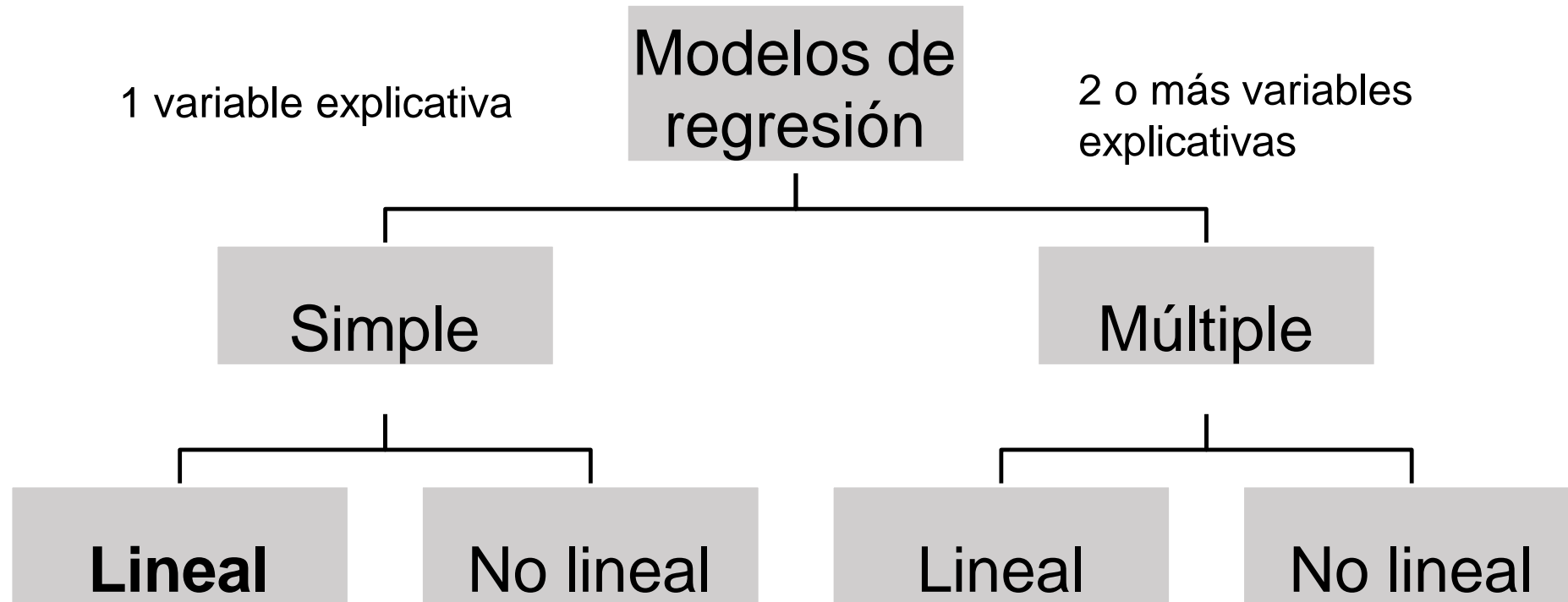
Algunas preguntas de investigación

- ¿Qué efecto tiene una reducción del número de estudiantes en clase sobre sus notas?
- ¿Y un año de educación sobre el salario futuro?
- ¿Cuál es el efecto del impacto de cambios en la política fiscal sobre los indicadores económicos del país, la demanda interna, exportaciones e importaciones, desempleo,...?
- ¿la demanda de cigarrillos se ve afectada por las campañas antitabaco?
- ¿cómo afecta el tabaquismo al peso de nacimiento y posterior crecimiento de un bebe?

MODELOS LINEALES...¿POR QUÉ?

- Existen muchas situaciones donde la variabilidad de una cierta magnitud de interés se puede explicar en términos de los valores de otras magnitudes. Cuando esta explicación se puede expresar a través de una relación funcional más un componente aleatorio, se pueden predecir valores de la variable de interés cuando se conocen los valores de las variables que la explican. Si este modelo es, además, lineal se gana en sencillez y en facilidad en los necesarios procesos de estimación.
- En este curso se estudian las técnicas estadísticas que permiten ajustar los modelos lineales y conocer su potencialidad al aplicarlos a numerosos contextos.

Modelos de análisis de regresión



No trabajaremos modelos a mano. Usaremos para ello RStudio.

Actividades a realizar

A lo largo del curso practicaremos la correcta aplicación de los métodos estadísticos, que podemos establecer de modo general en las siguientes fases:

Establecer claramente cuál es el problema a resolver, así como los objetivos a alcanzar en su ámbito profesional concreto.

Traducir el problema a términos estadísticos: identificar qué métodos y modelos habrán de aplicarse. ¿Es un problema de simple descripción de variables? ¿Se deben establecer relaciones entre las mismas estimando los modelos adecuados? ¿Se va a realizar algún tipo de inferencia o predicción?

Explorar los datos, describiendo sus características mediante medidas descriptivas, tablas y gráficos.

Ajustar modelos, estimando sus parámetros y su nivel de incertidumbre. Validar las hipótesis en que se sustentan los modelos

Extraer las conclusiones que procedan.

Actualmente los procesos de exploración de datos, ajuste de modelos, validación y, en su caso, predicción tienen una componente computacional importante.



El software que utilizaremos en este curso será R/Rstudio



Manejaremos librerías modernas, con especial hincapié en los paradigmas del tidyverse (datos bien ordenados y estructurados) y el uso de pipelines (tuberías) para enlazar tareas.



La única forma de aprender/comprender los distintos temas de este módulo y de todo el curso es practicar, practicar y practicar.

Manejo de Software estadístico (R y Rstudio)



BIBLIOGRAFÍA

- Existen numerosos tutoriales, libros, canales de youtube, ... en la web dedicados a los métodos estadísticos que veremos este módulo, y en particular a su aplicación con R.
- [Linear Models with R](#) *Julian J. Faraway*, Chapman & Hall/CRC, 2005
- *Analysing Ecological Data*. *Zuur, A.F., Ieno, E., Smith, G.M.* Springer (2007)
- [ggplot2: elegant graphics for data analysis](#). *Wickham, H.* Springer (2009). [\[Página web del programa\]](#)
- [R para Ciencia de Datos](#) *Grolemund, G, Wickham H.* Thomson (2006) [\[Página web de Tidyverse\]](#)

VARIABLES CORRELACIONADAS

CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE

Estudio de la relación entre dos variables numéricas de forma gráfica

Para estudiar la relación entre dos variables numéricas de forma gráfica se utiliza el *diagrama de dispersión*.

Por ejemplo, para evaluar políticas públicas, es preciso identificar y estudiar las relaciones entre variables, ya que puede proporcionar información sobre las formas de mejorar las condiciones de vida, el bienestar social y el desarrollo económico en general.

Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.

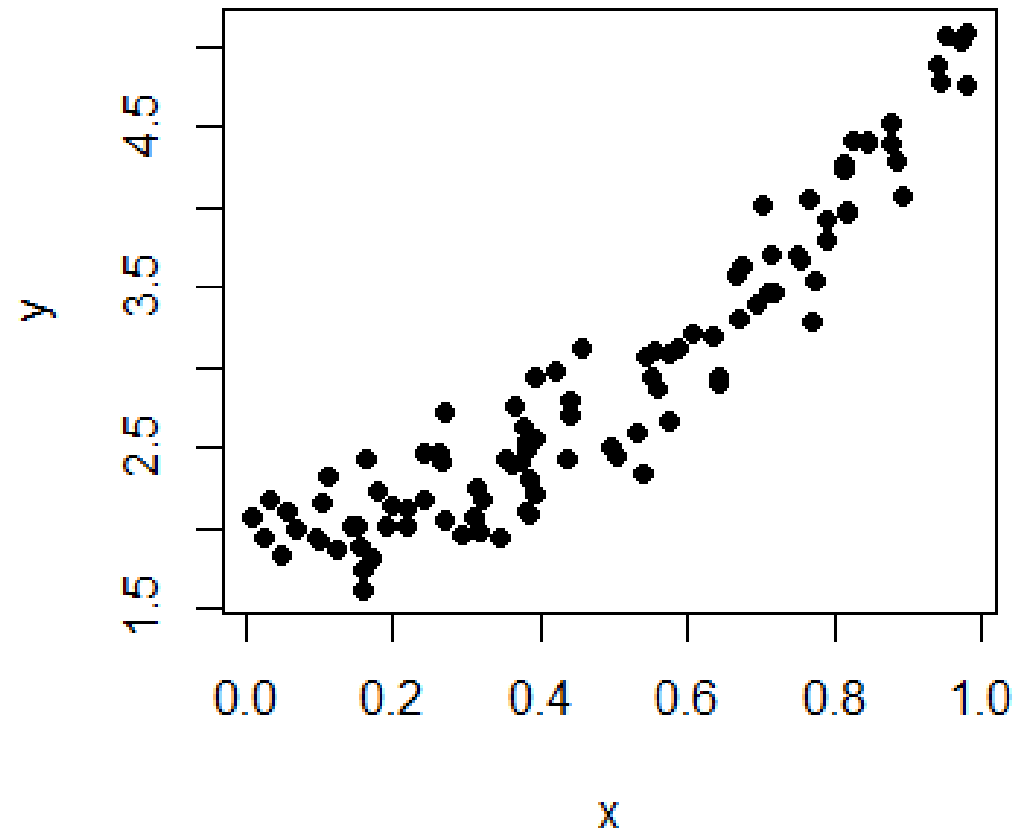
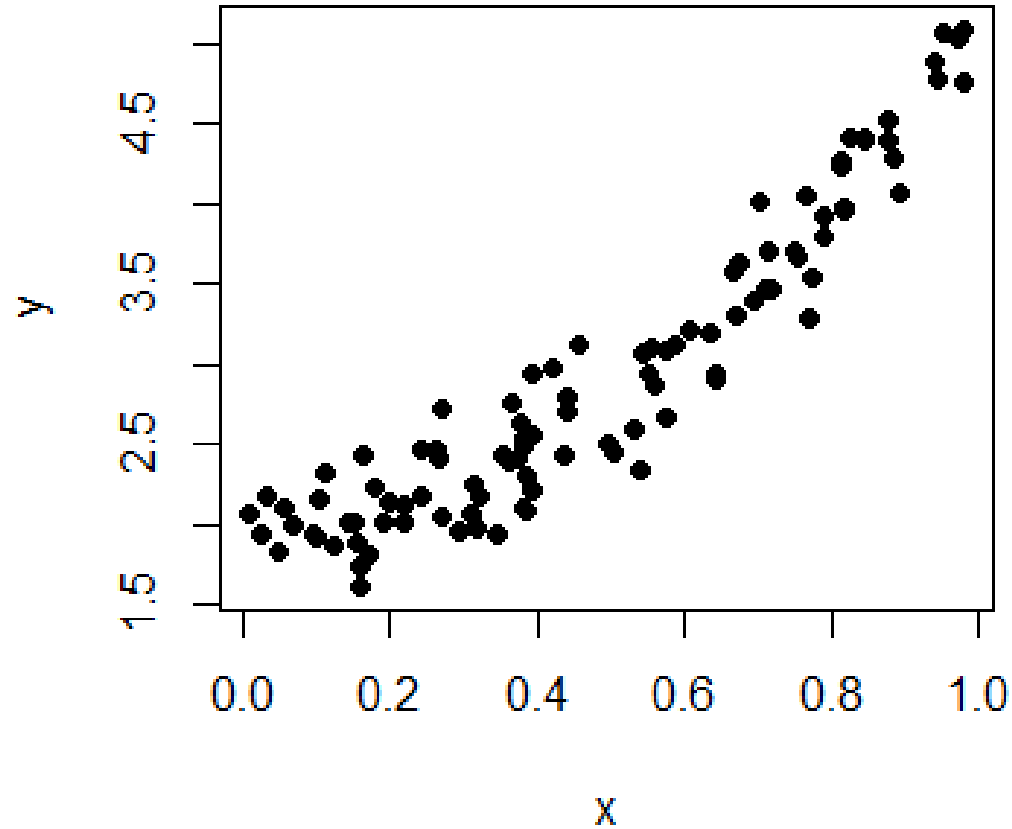


Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.



Existe relación directa entre las variables x e y

Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.

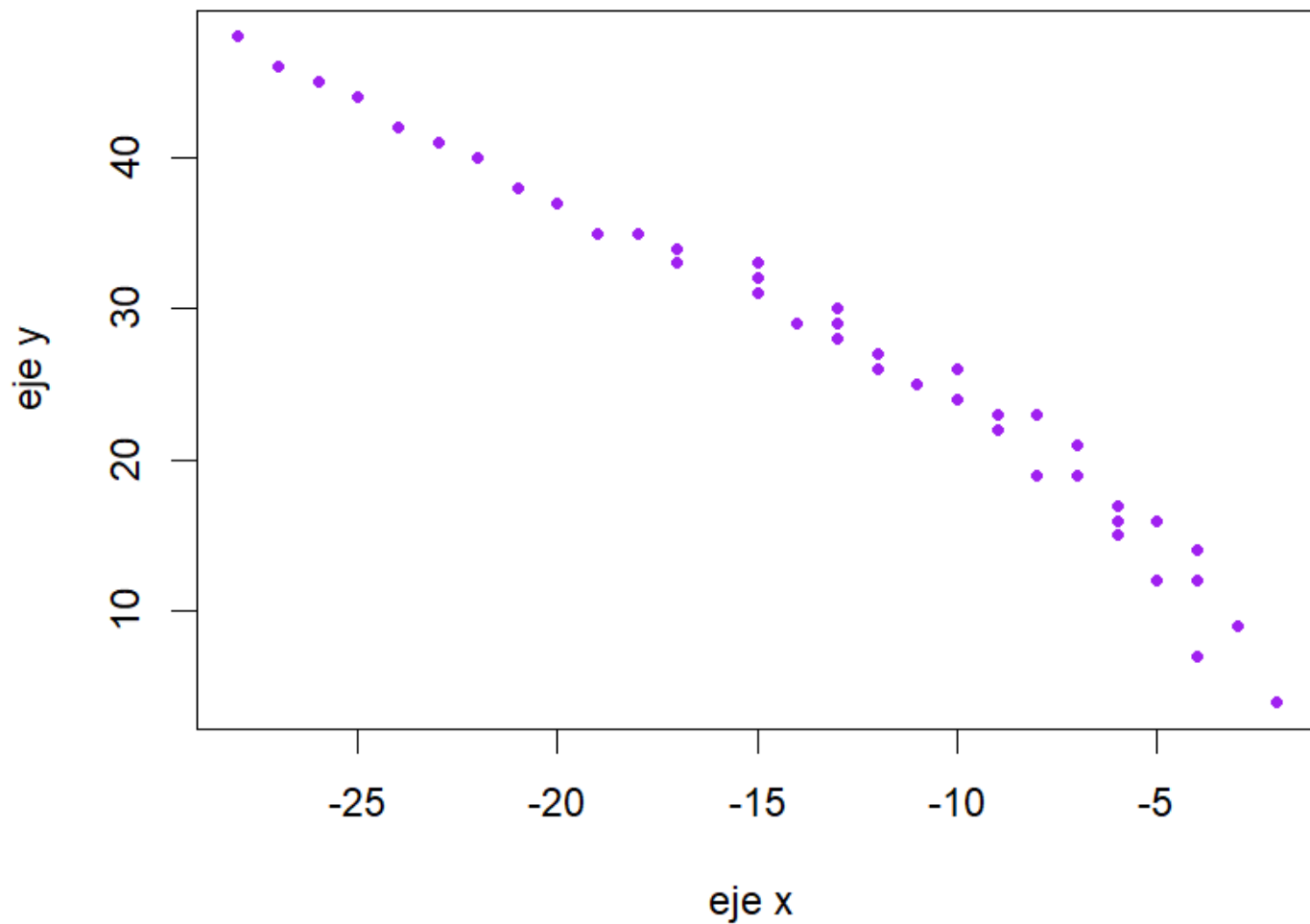
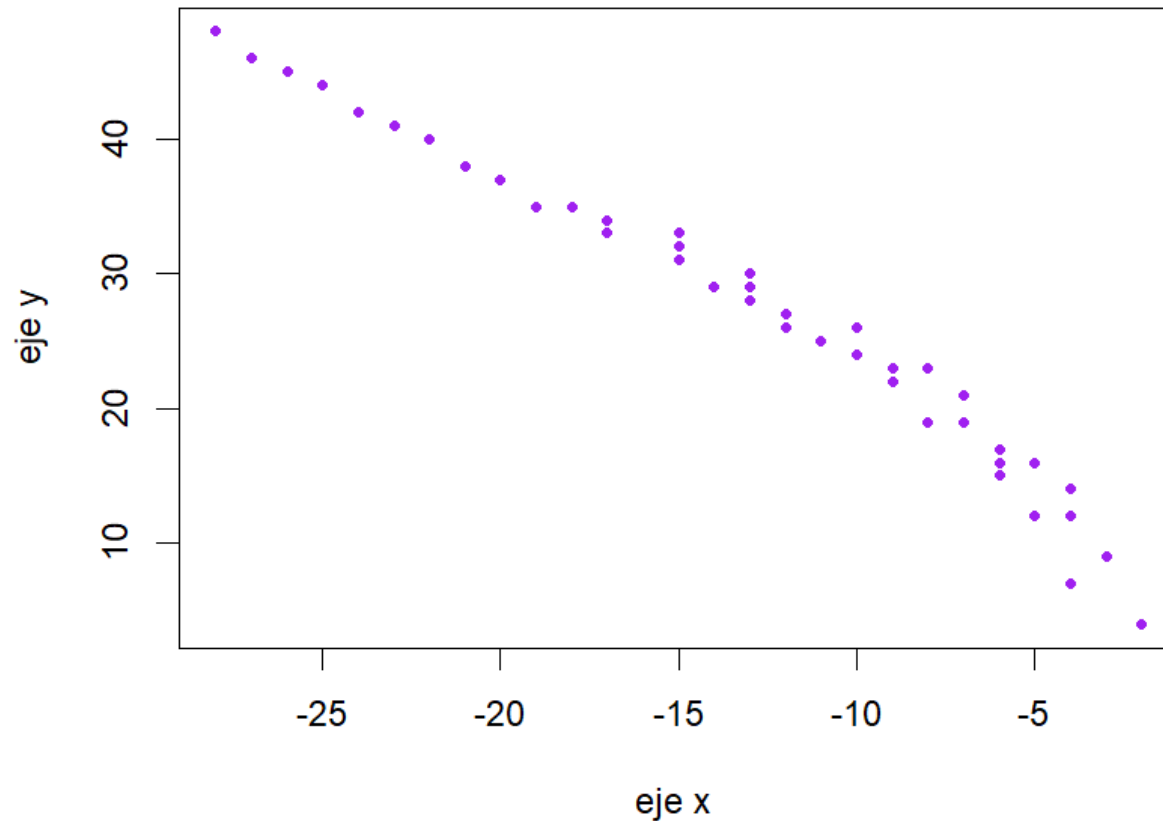


Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.



Existe relación inversa entre las variables x e y

Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.

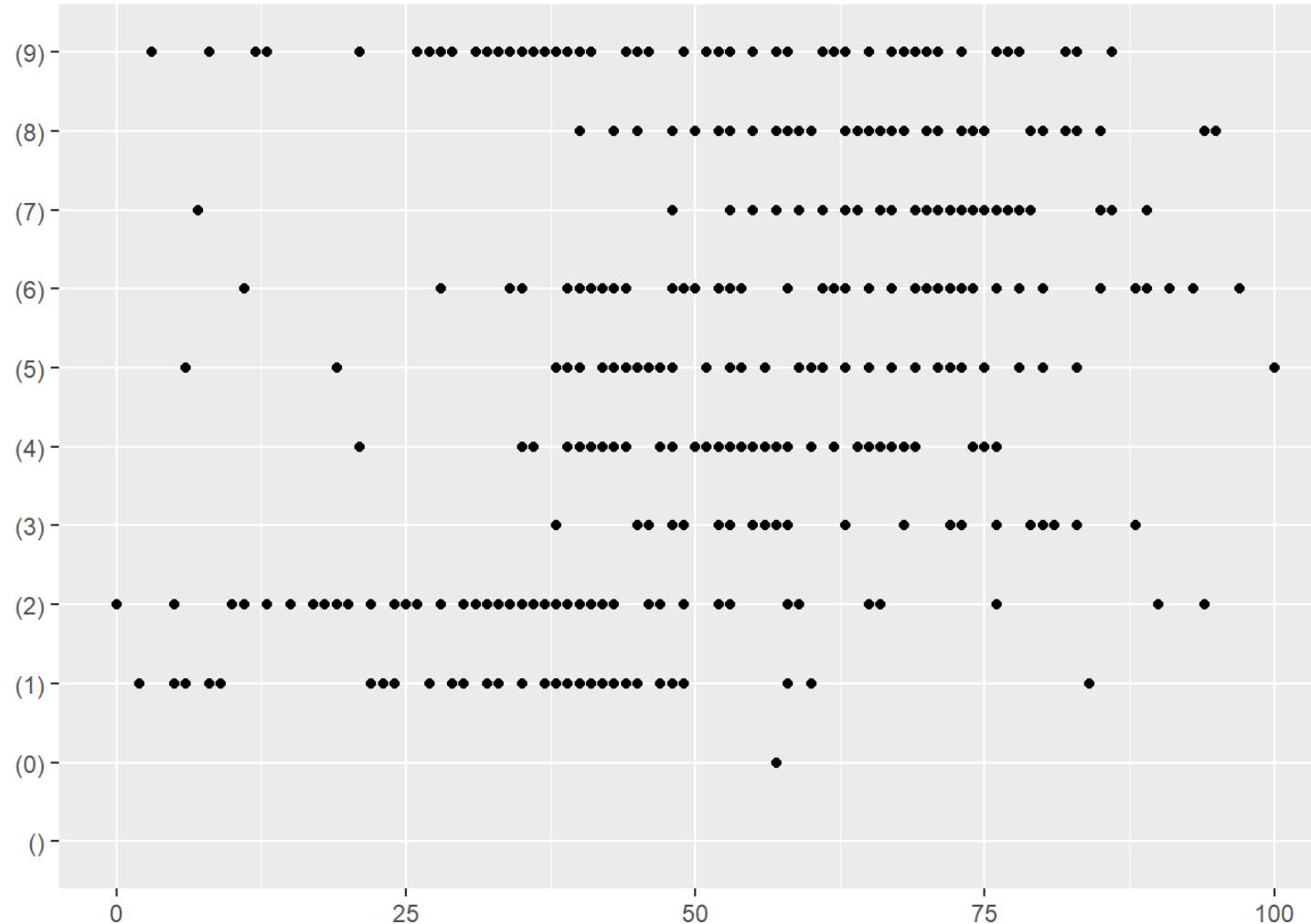
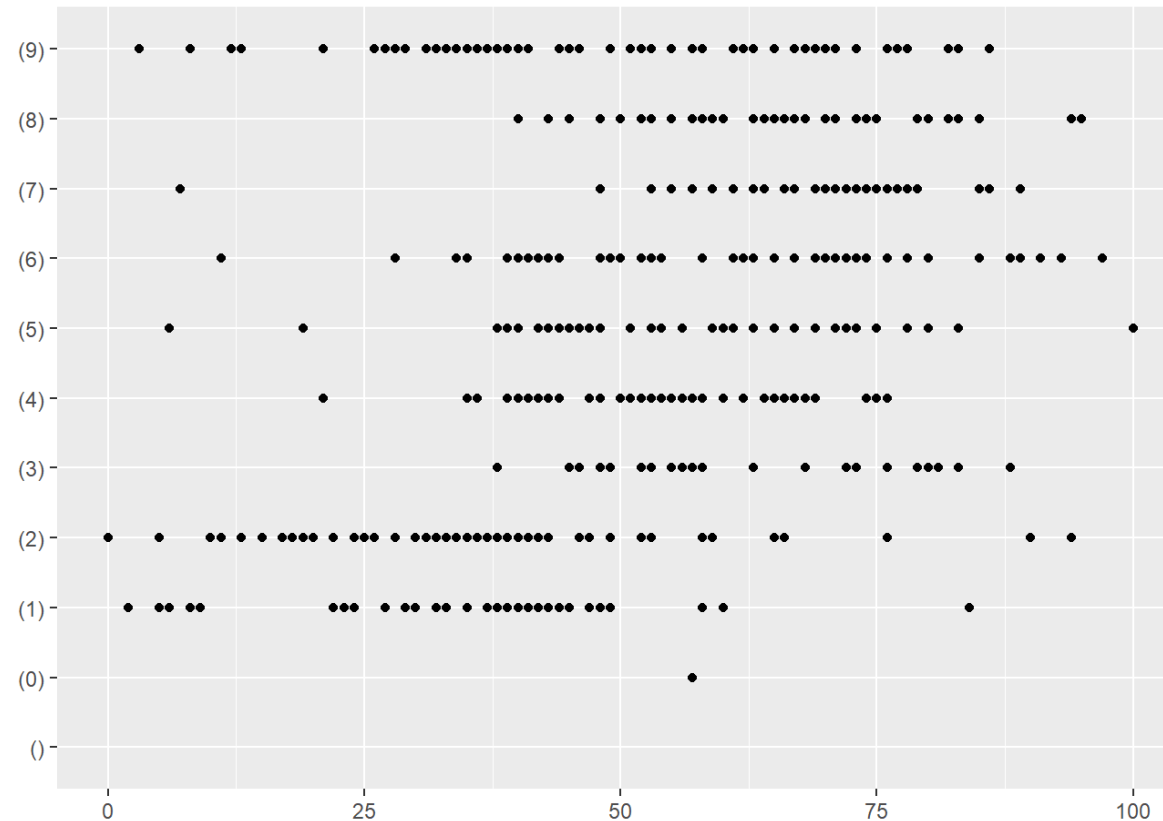


Diagrama de Dispersión

Gráfico que nos permite observar la relación entre dos variables.



No existe relación entre las variables x e y

Algunos ejemplos puntuales se describen a continuación:

¿ El ingresos de los hogares incide en el porcentaje de analfabetismo?

Variable dependiente:

Variable regresora:

¿El ingresos de los hogares incide en el porcentaje de analfabetismo?

Variable dependiente: **Porcentaje de analfabetismo**

Variable regresora: **El ingreso**

“Relación inversa entre las variables”

**¿El nivel socioeconómico del estudiante (INSE)
incide en los resultados en las pruebas Saber 11?**

Variable dependiente:

Variable regresora:

¿El nivel socioeconómico del estudiante (INSE) incide en los resultados en las pruebas Saber 11?

Variable dependiente: Resultados de las pruebas Saber 11

Variable regresora: El nivel socioeconómico del estudiante (INSE)

“Relación directa entre las variables”

¿El monto de dinero invertido en una campana de publicidad sobre celulares incide en el monto de compra de los consumidores de este producto?

Variable dependiente:

Variable regresora:

¿El monto de dinero invertido en una campaña de publicidad sobre celulares incide en el monto de compra de los consumidores de este producto?

Variable dependiente: Monto de compra de los consumidores

Variable regresora: El monto de dinero invertido en una campaña de publicidad sobre celulares.

“Relación directa entre las variables”

¿ Los años de experiencia de los empleados incide en las ventas en una empresa?

Variable dependiente:

Variable regresora:

¿ Los años de experiencia de los empleados incide en las ventas en una empresa?

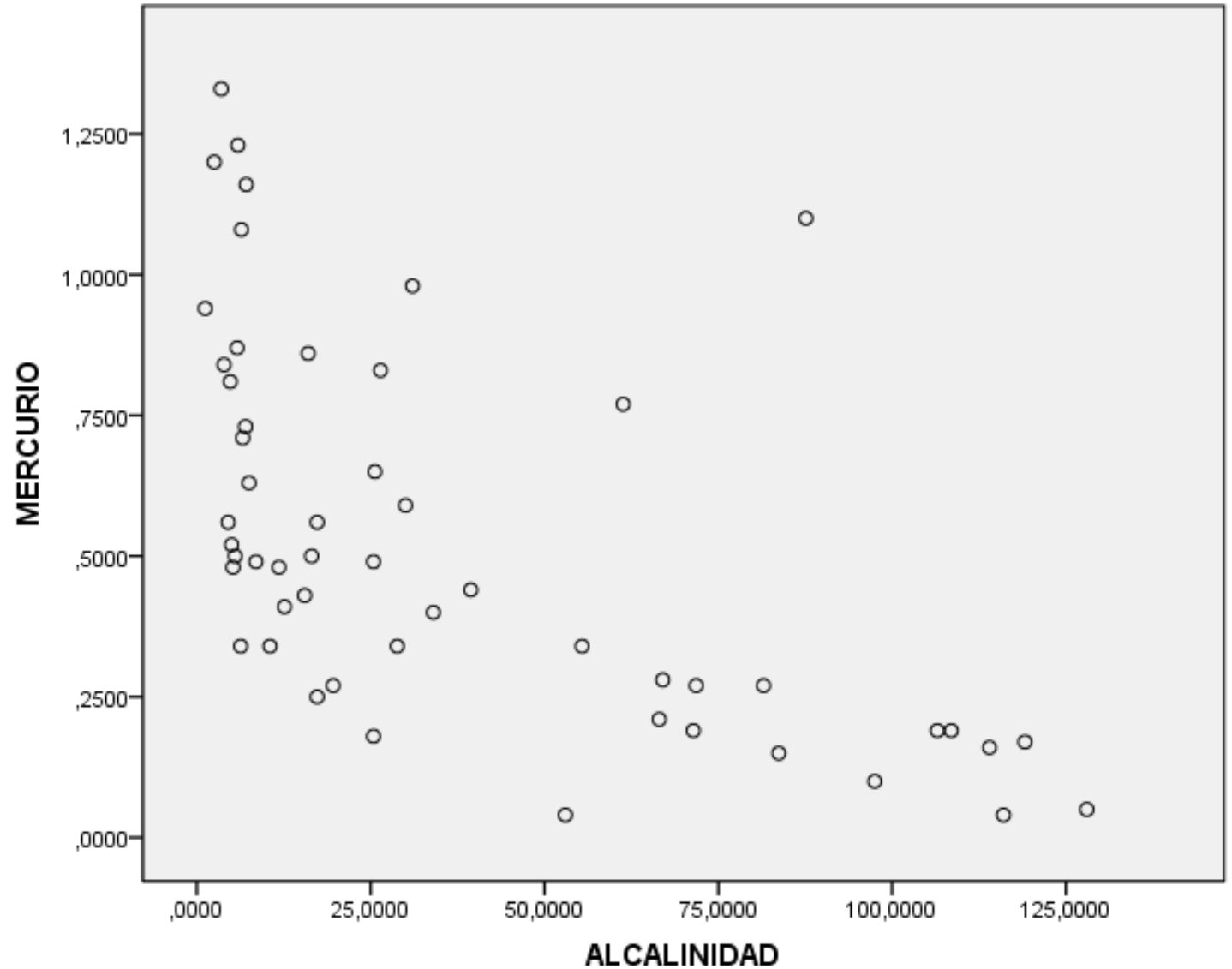
Variable dependiente: **Ventas en una empresa**

Variable regresora: **Años de experiencia de los empleados**

“Relación directa entre las variables”

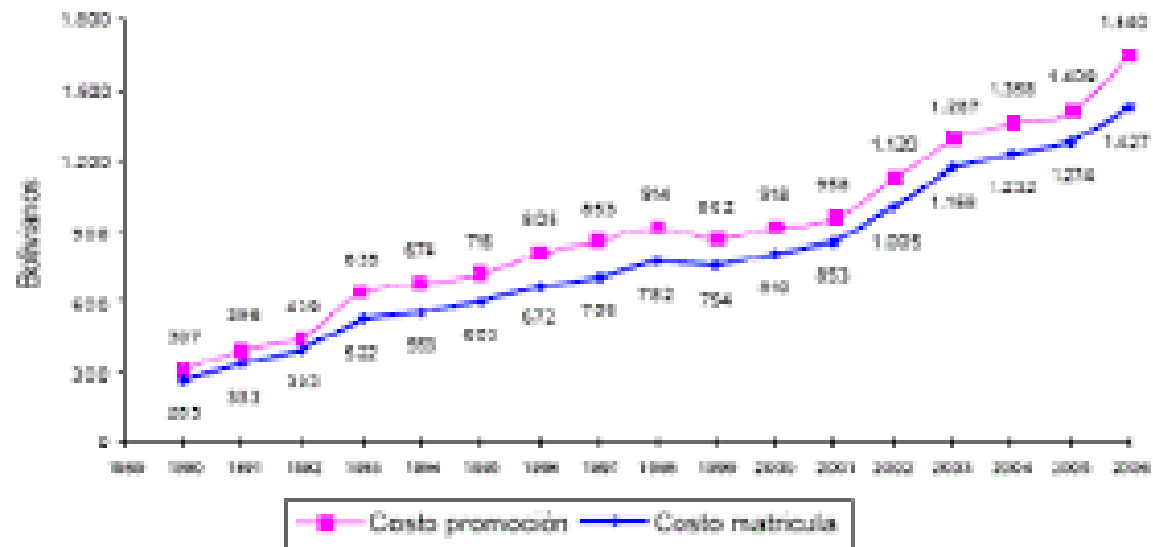
Relación entre la alcalinidad del agua y la concentración media de mercurio en peces

¿Qué parecen decir los datos sobre el comportamiento de la concentración de mercurio en peces y la alcalinidad del agua?

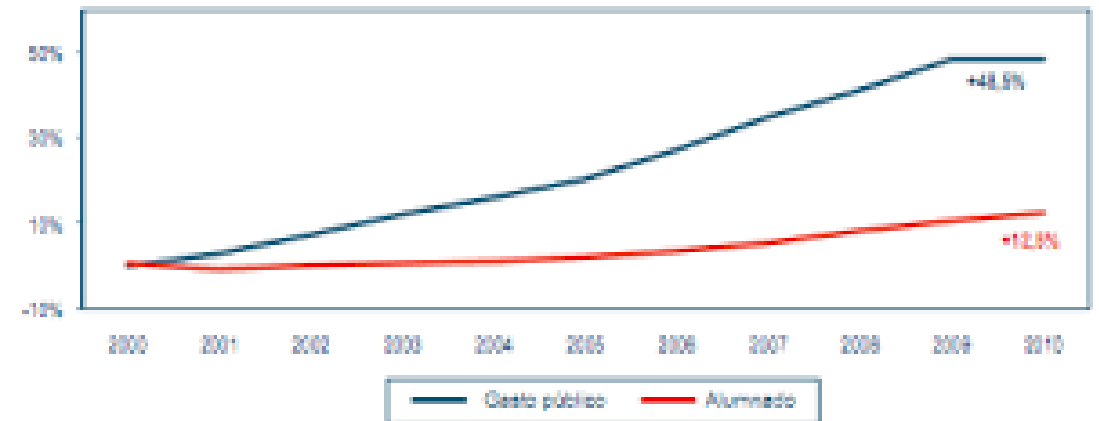


Algunos ejemplos

Costo por alumno

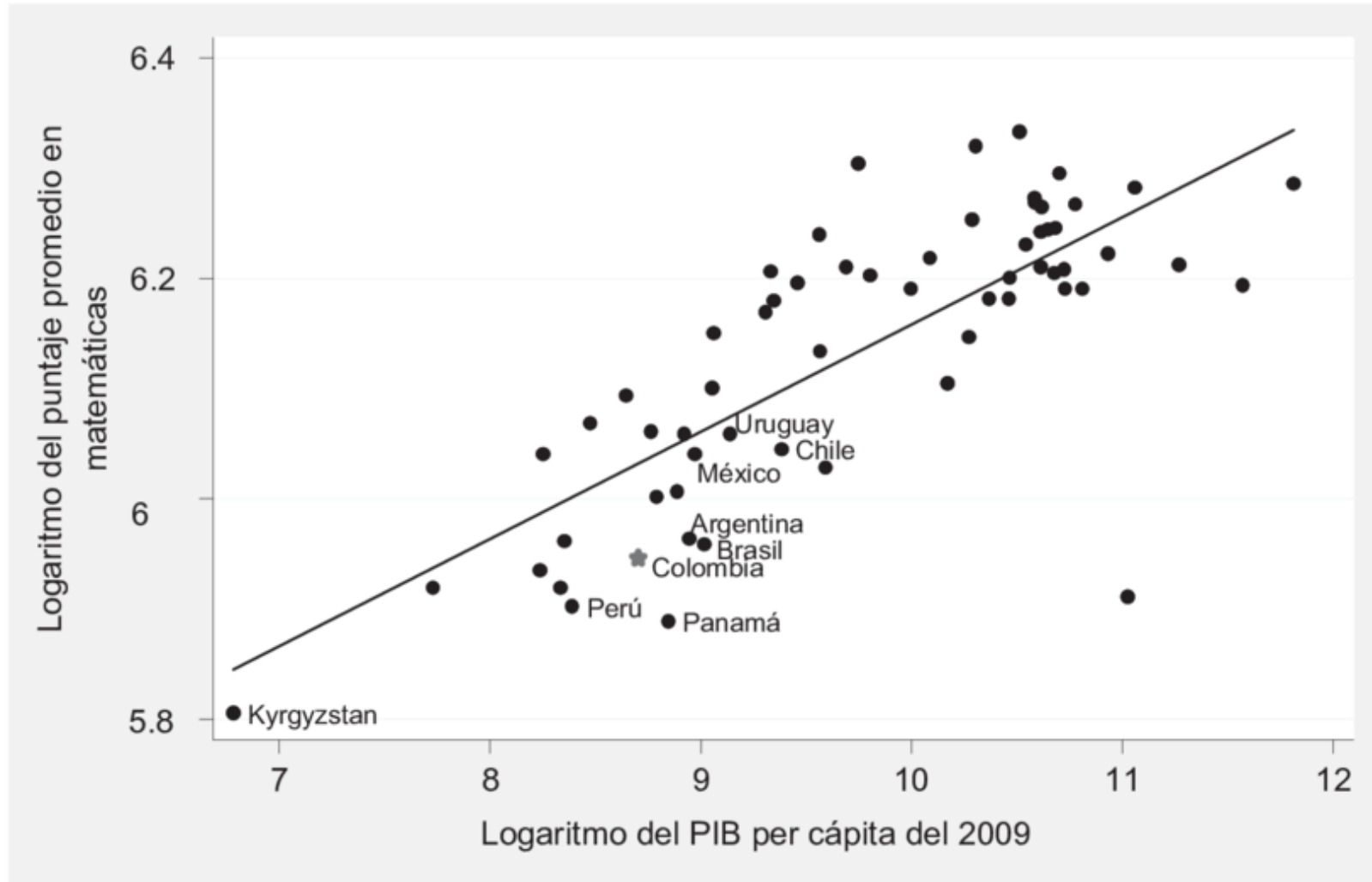


Evolución de la variación del gasto público en educación⁽¹⁾ y del alumnado en el periodo 2000 a 2010



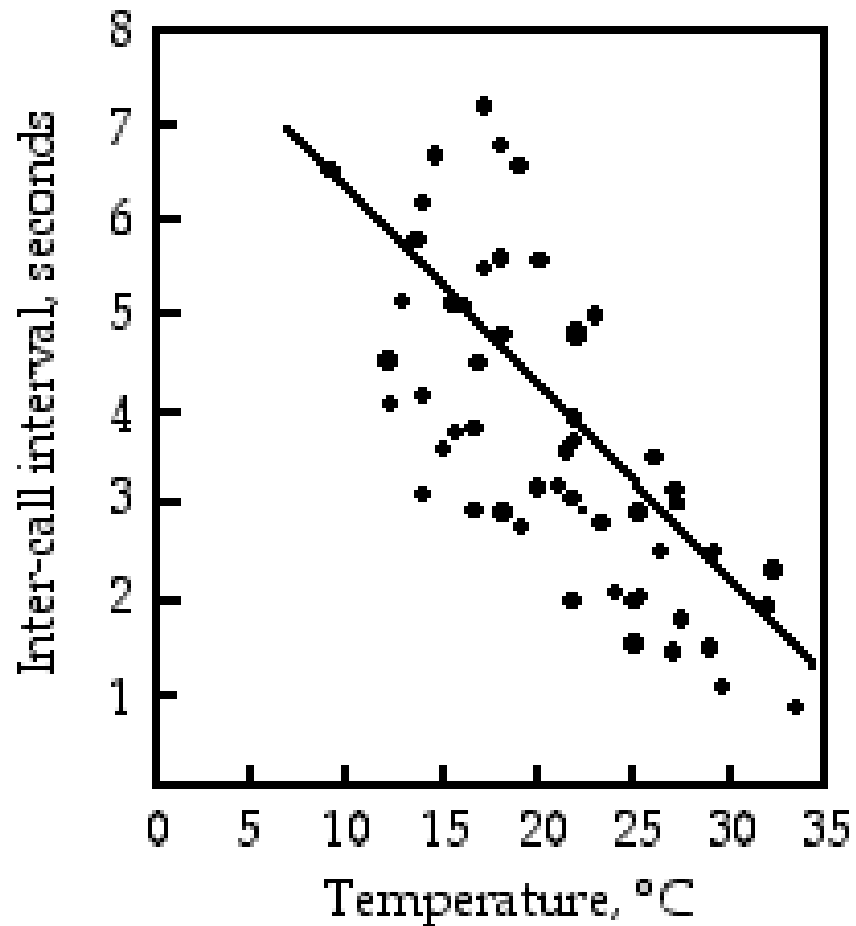
¿Qué parecen decir los datos?

Relación entre el puntaje promedio en matemáticas y el PIB por países (2009)



¿Qué parecen decir los datos?

Relación de la temperatura corporal y el intervalo entre llamadas en la rana arborícola gris.

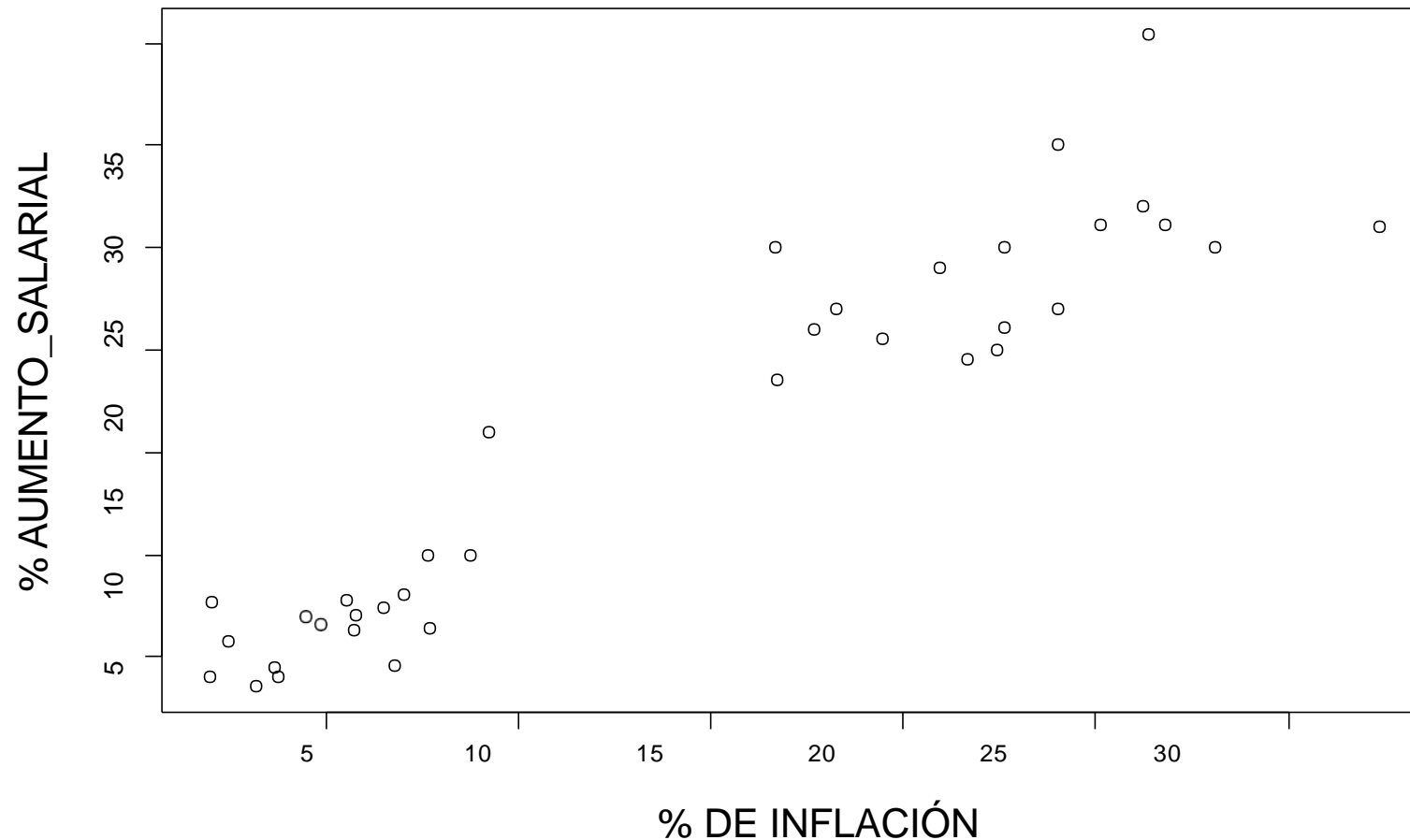


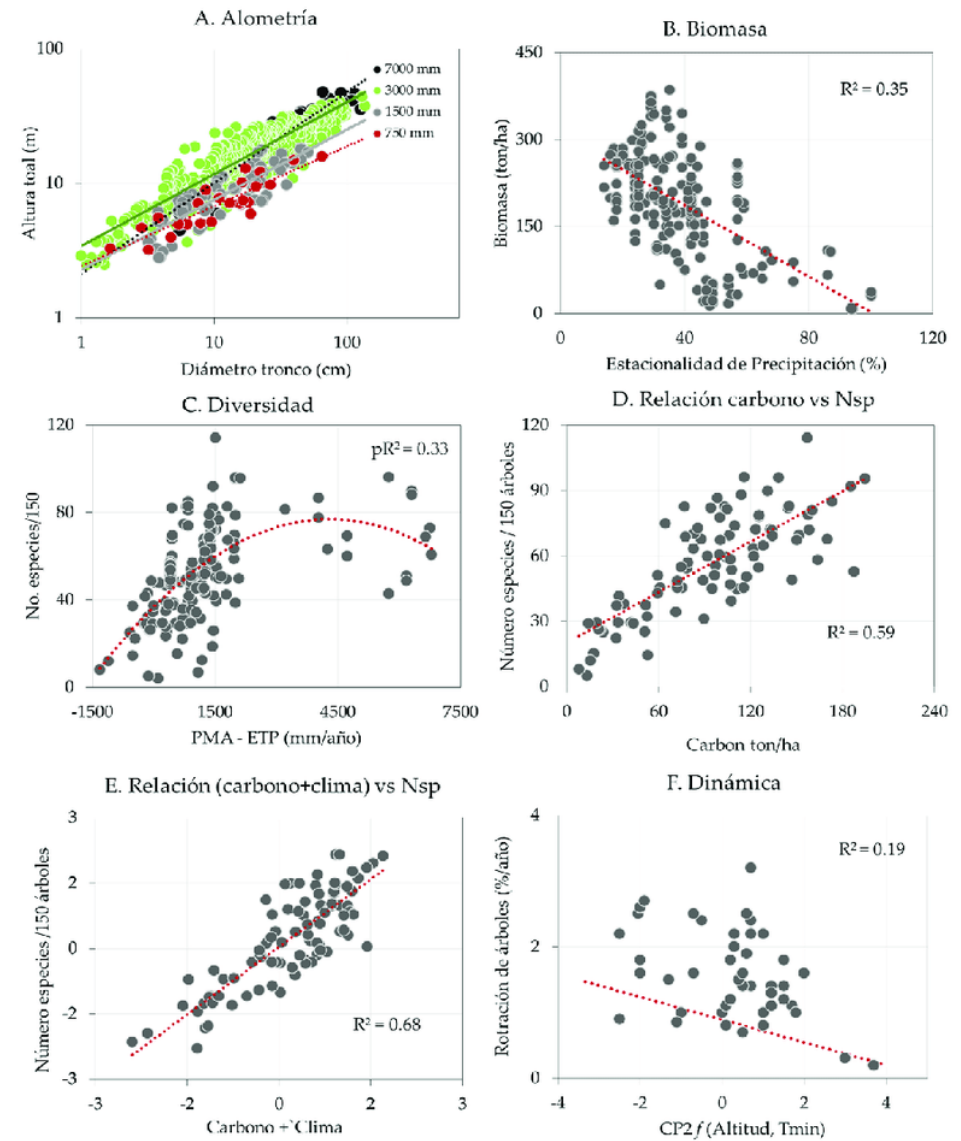
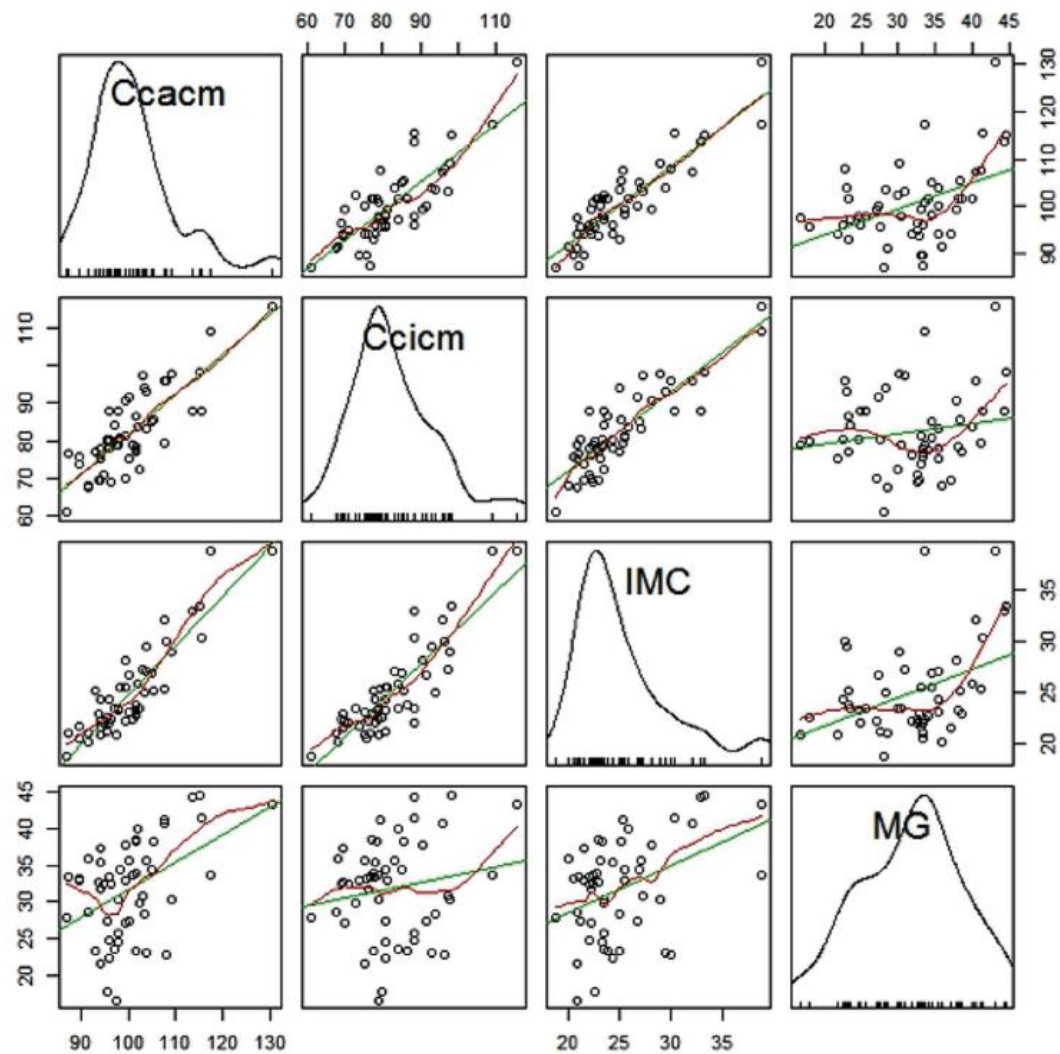
¿Qué parecen decir los datos?



Rana arborícola gris, *Hyla versicolor*.

Relación entre la inflación y el aumento del salario en Colombia en los últimos años





TRAS ENTREVISTAR A MILES DE PERSONAS, HE ENCONTRADO UNA FUERTE CORRELACIÓN ENTRE SER INTELIGENTE Y ESTAR DE ACUERDO CONMIGO.

Aplicación del análisis de regresión lineal simple para la estimación de los precios de las acciones de Facebook, Inc.

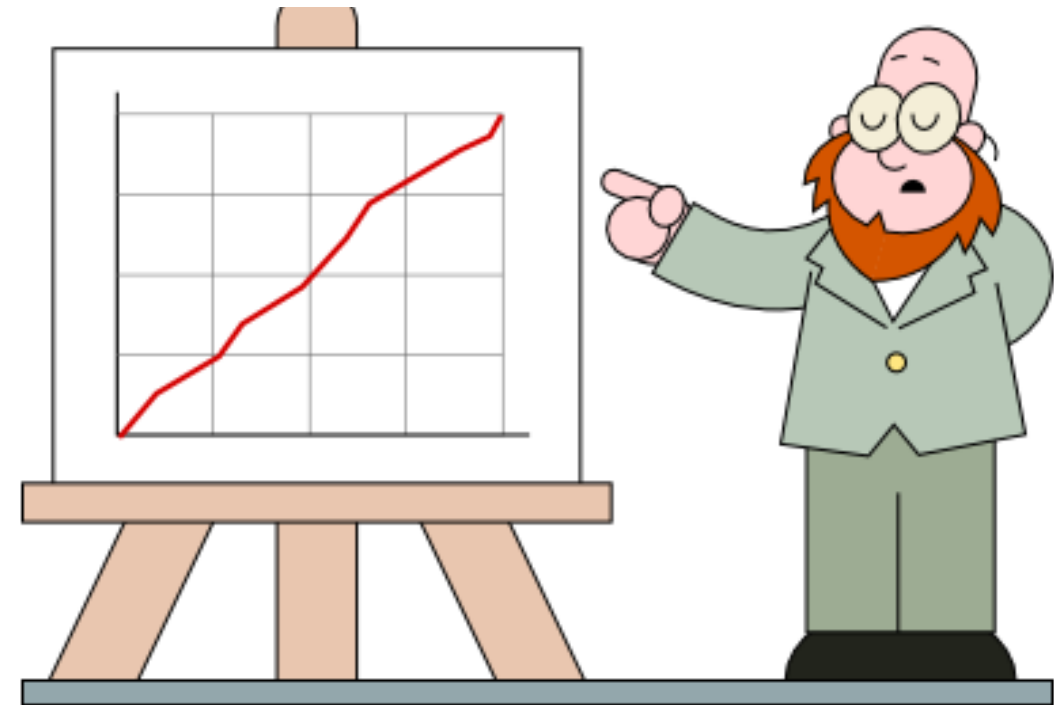
MUNDO

Las pruebas PISA mostraron una correlación entre el bullying y el bajo desempeño escolar

El informe de bienestar y satisfacción de los estudiantes reveló que un 18,7% de los encuestados dijo haber sufrido algún tipo de acoso escolar, una proporción aún mayor entre quienes tuvieron un bajo desempeño académico

APLICACIÓN DEL MÉTODO DE REGRESIÓN LINEAL EN EL ANÁLISIS DE LOS DETERMINANTES DE LA INVERSIÓN EXTRANJERA EN COLOMBIA

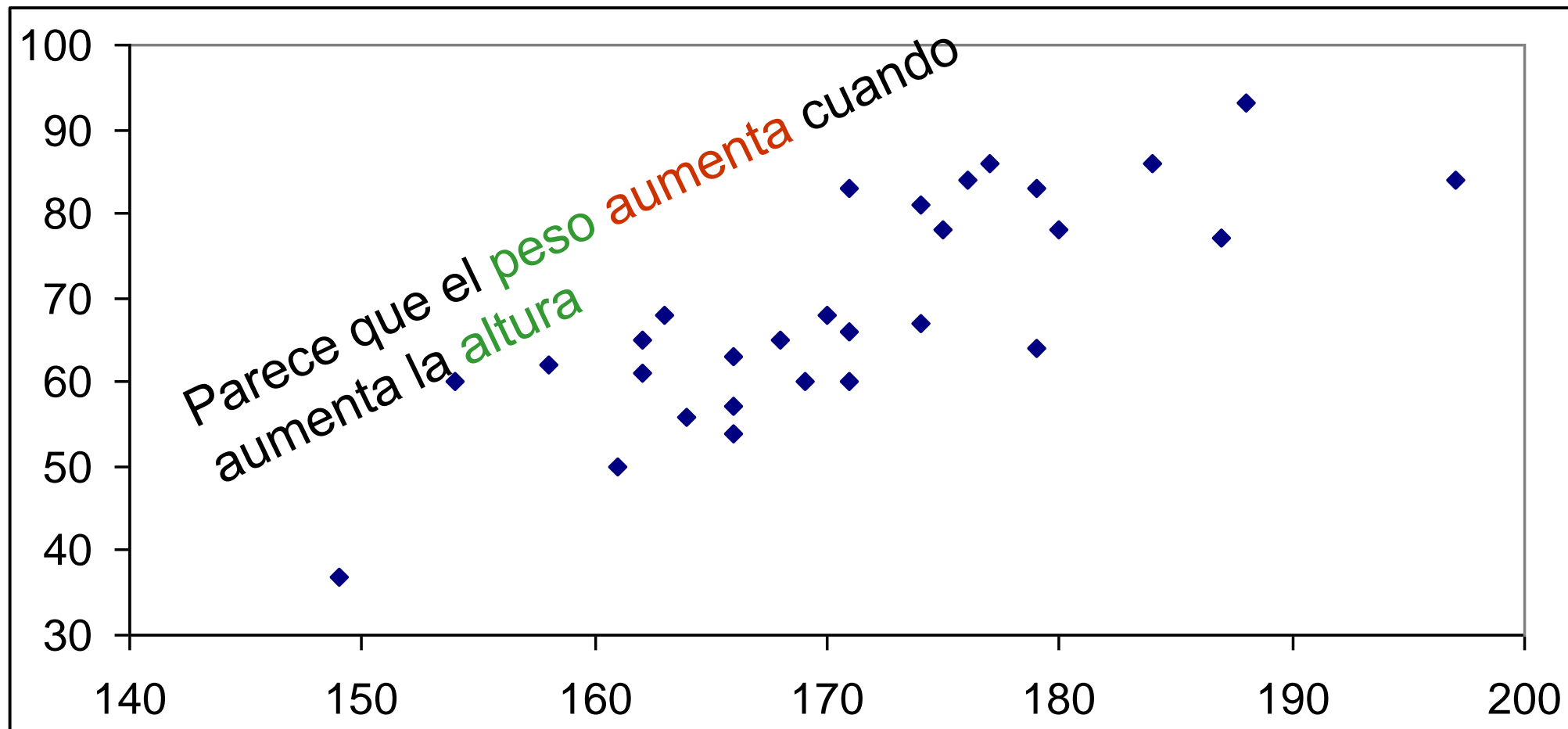
El análisis de regresiones lineales entre el tipo de cambio peso – dólar y diferentes variables económicas de México y Estados Unidos durante los últimos 20 años



Empecemos a analizar la
relación existente entre dos o
más variables numéricas



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



La Correlación es una técnica estadística usada para cuantificar que tan fuerte es la relación entre dos o más variables numéricas

- Altura y peso de niños.
- pH y biomasa
- tiempo de exposición a un pesticida y mortalidad de una plaga
- Resultados académicos y tiempo de estudio.

La relación puede ser claramente causal o no

- La potencia del motor de un auto es la causa de que alcance una mayor velocidad,
- La relación altura – peso tiene parte de causalidad, pero también existen otros factores.

Cuando se estudian variables correlacionadas hay que analizar bien el fenómeno para no caer en errores

Tipos de correlación

Correlación positiva o directa

Cuando hay valores altos o bajos, simultáneamente en dos variables.

Ejemplo:

Peso y altura en una muestra de niños de 5 a 12 años: los mayores son también los más altos y pesan más, y los más jóvenes pesan menos y son más bajos; decimos que peso y altura son dos variables relacionadas porque los más altos pesan más y los más bajos pesan menos.

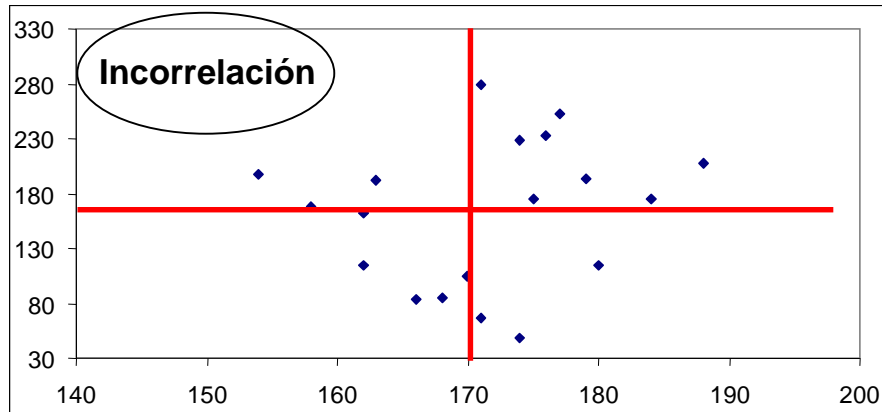
Correlación Negativa o inversa

Es cuando los valores altos en una variable coinciden con valores bajos en otra variable.

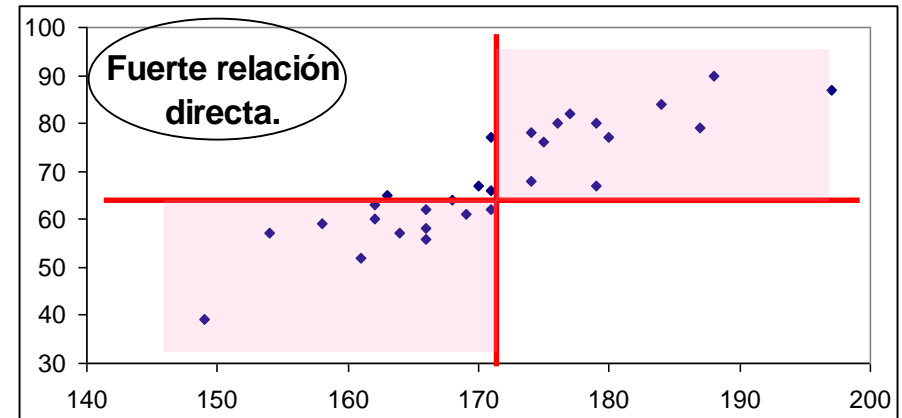
Ejemplo:

La edad y fuerza física en una muestra de adultos de 30 a 80 años de edad: los mayores son los menores en fuerza física; hay una relación, que puede ser muy grande: según los sujetos aumentan en una variable (edad) disminuyen en la otra (fuerza física).

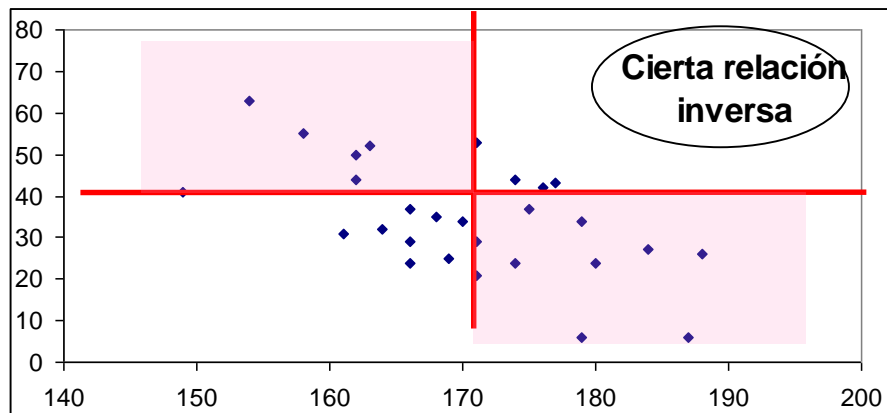
En resumen:



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



- 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.
- Esto se llama **relación directa**.



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

Practiquemos en RStudio

¿CÓMO MEDIR LA
CORRELACIÓN ENTRE
DOS VARIABLES?



Covarianza de dos variables X e Y

- La **covarianza** entre dos variables, S_{xy} , nos indica si la posible *relación lineal* entre dos variables es directa o inversa.

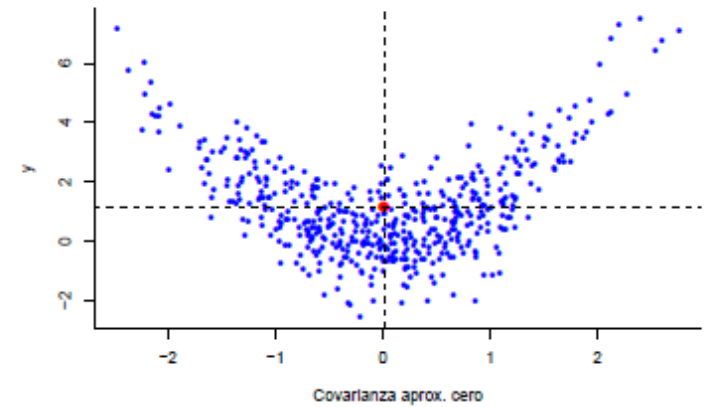
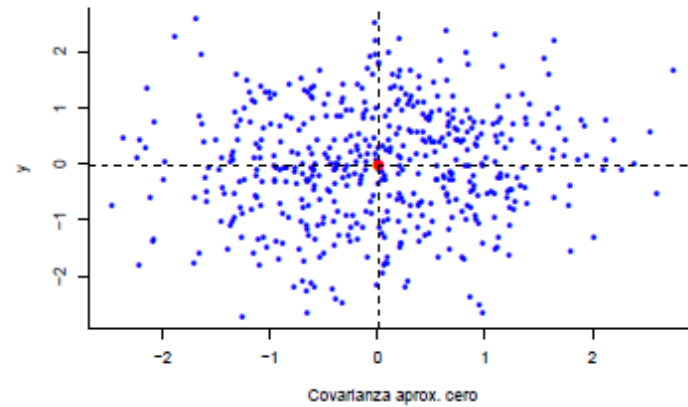
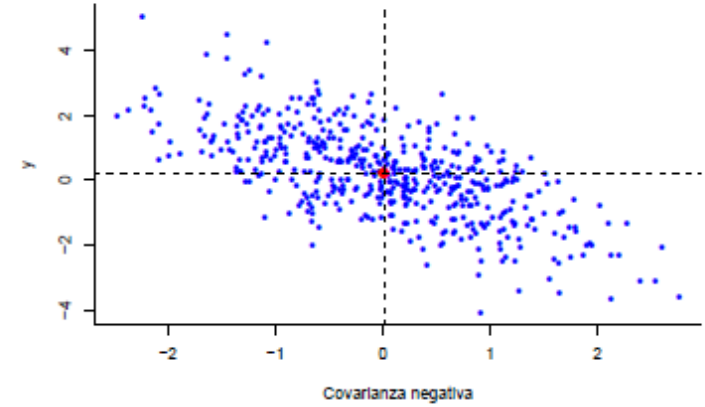
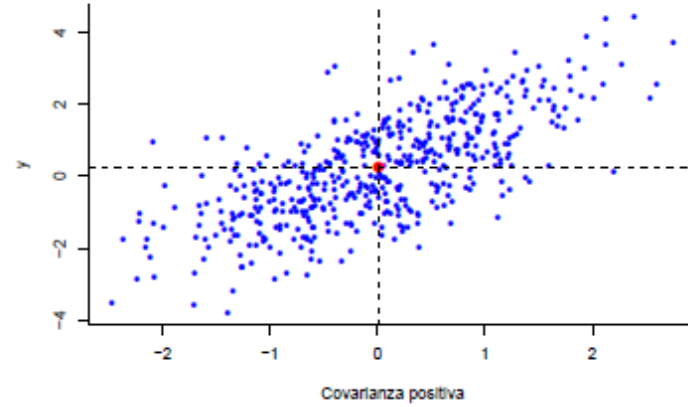
- **Directa**: $S_{xy} > 0$
- **Inversa**: $S_{xy} < 0$
- **Incorreladas**: $S_{xy} = 0$

$$\sigma_{xy} = \frac{1}{n} \sum_i (x_i - \mu_x)(y_i - \mu_y)$$

$$S_{xy} = \frac{1}{n-1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- 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.

Ejemplos covarianza



Ejemplos usando Excel

Actividad 1

“Trabajo en clase”

Coeficiente de correlación lineal de Pearson

- La **coeficiente de correlación lineal de Pearson** de dos variables, r , nos indica si los puntos tienen una **tendencia a disponerse alineadamente** (excluyendo rectas horizontales y verticales).
- tiene el mismo signo que S_{xy} por tanto de su signo obtenemos el que la posible relación sea directa o inversa.
- r es útil para determinar si hay relación **lineal** entre dos variables, pero **no servirá para otro tipo de relaciones** (cuadrática, logarítmica,...)

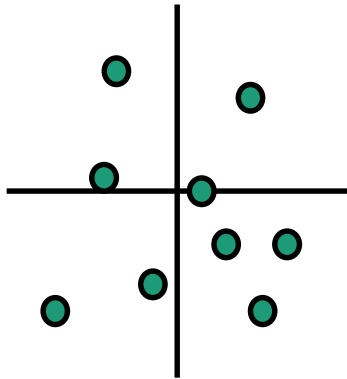


$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad r = \frac{S_{xy}}{S_x S_y}$$

Propiedades de r

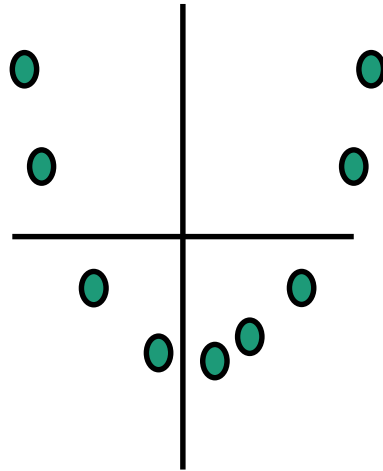
- Puede tener signo positivo o negativo, dependiendo del signo de la *covariación* muestral de las variables.
- Cae entre los límites de -1 y 1
- Es simétrico por naturaleza; es decir, el coeficiente de correlación entre X y Y (r_{xy}) es el mismo que entre Y y X (r_{yx}).
- Es independiente del origen y de la escala
- Si X y Y son estadísticamente independientes, el coeficiente de correlación entre ellos es cero; pero si $r = 0$, esto no significa que las dos variables sean independientes. En otras palabras, una correlación igual a cero no necesariamente implica independencia.
- Es una medida de *asociación lineal* o *dependencia lineal* solamente; su uso en la descripción de relaciones no lineales no tiene significado.

Grado de Correlación



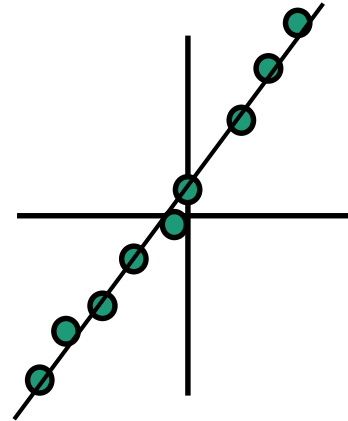
No hay
correlación

$$r \approx 0$$



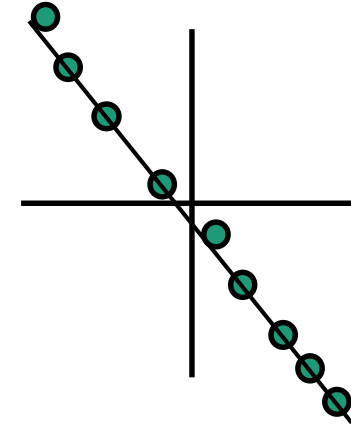
Hay correlación
no lineal

$$r \approx 0$$



Correlación lineal
positiva

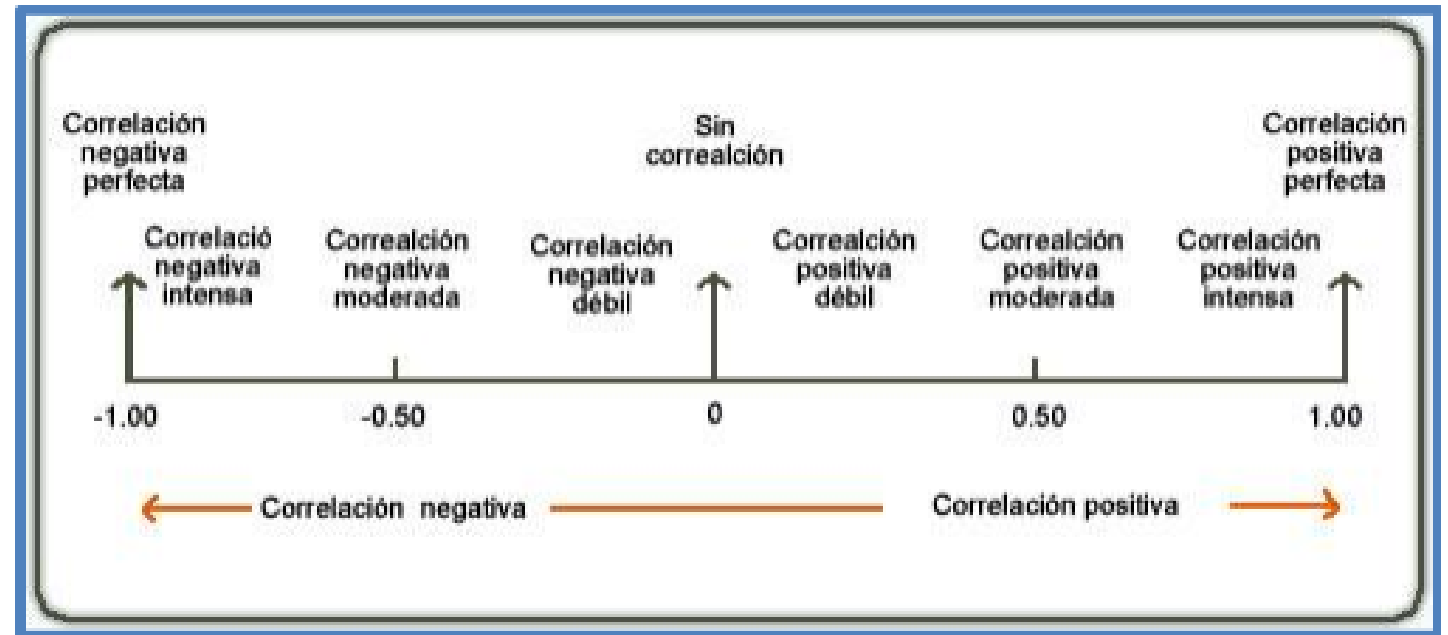
$$r \approx +1$$

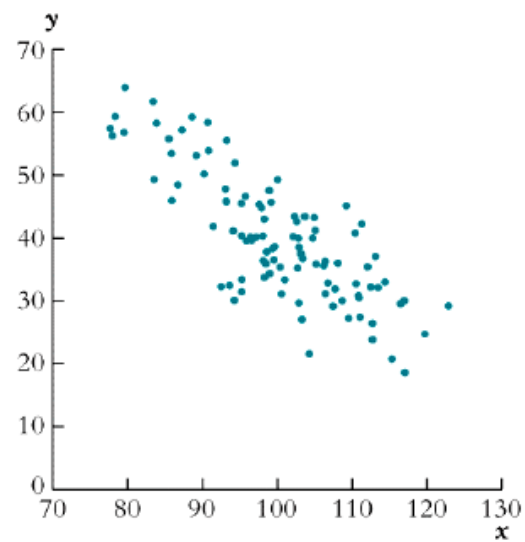


Correlación lineal
negativa

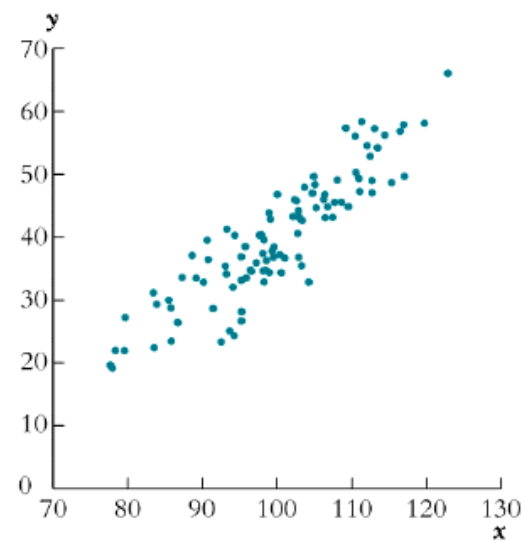
$$r \approx -1$$

¿Qué tan buena es la correlación lineal?

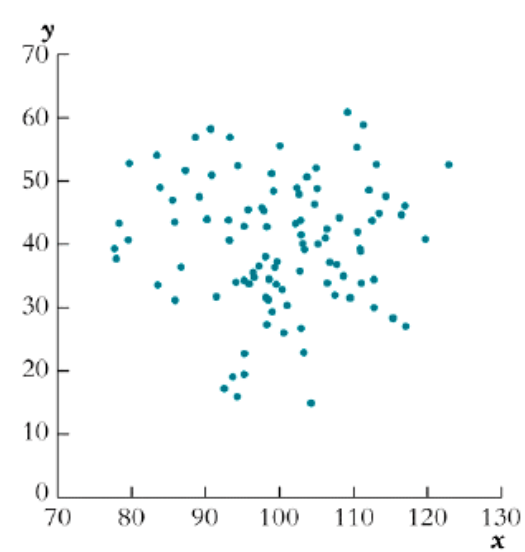




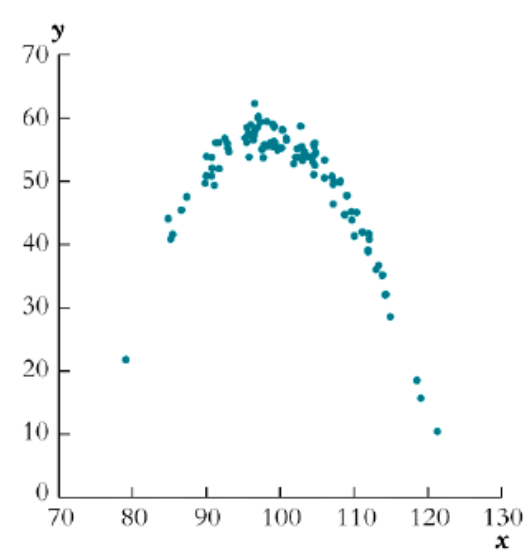
(b) Correlation = -0.8



(a) Correlation = $+0.9$



(c) Correlation = 0.0



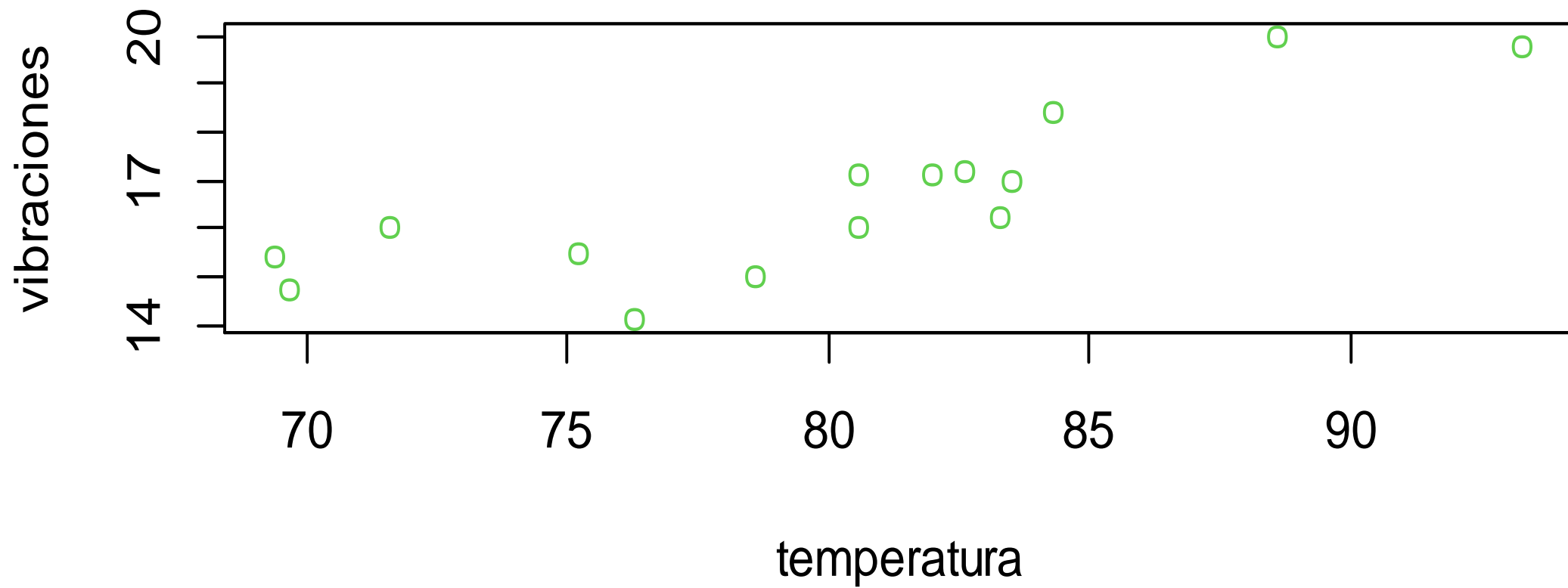
(d) Correlation = 0.0 (quadratic)

Vibraciones/seg.	Temp.
20.0	88.6
16.0	71.6
19.8	93.3
18.4	84.3
17.1	80.6
15.5	75.2
14.7	69.7
17.1	82.0
15.4	69.4
16.2	83.3
15.0	78.6
17.2	82.6
16.0	80.6
17.0	83.5
14.1	76.3



Ejemplo:
temperatura y
vibración de las alas

Los grillos son ectotermos, por lo que sus procesos fisiológicos y su metabolismo están influidos por la temperatura. Con el fin de estudiar estas cuestiones se ha medido el número de vibraciones por segundo de las alas de un grupo de grillos a varias temperaturas. Graficar, estudiar la posible correlación entre las variables. Hallar el coeficiente de correlación



```
cor(temperatura,vibraciones)  
1] 0.8364793
```

Ejemplo Salario

Para estudiar la relación entre el salario y los gastos de un grupo de personas, se obtuvieron 51 muestras en diferentes departamentos de una empresa.

- Realice un análisis exploratorio de los datos
- Encuentre la media, mediana, cuartiles, varianza, desviación estándar
- Realice un diagrama de dispersión
- Que podemos interpretar
- Encuentre el coeficiente de correlación lineal de Pearson para análisis realizado

Correlación no implica causalidad

El hecho de que dos eventos se den habitualmente de manera consecutiva no implica que uno sea causa del otro.



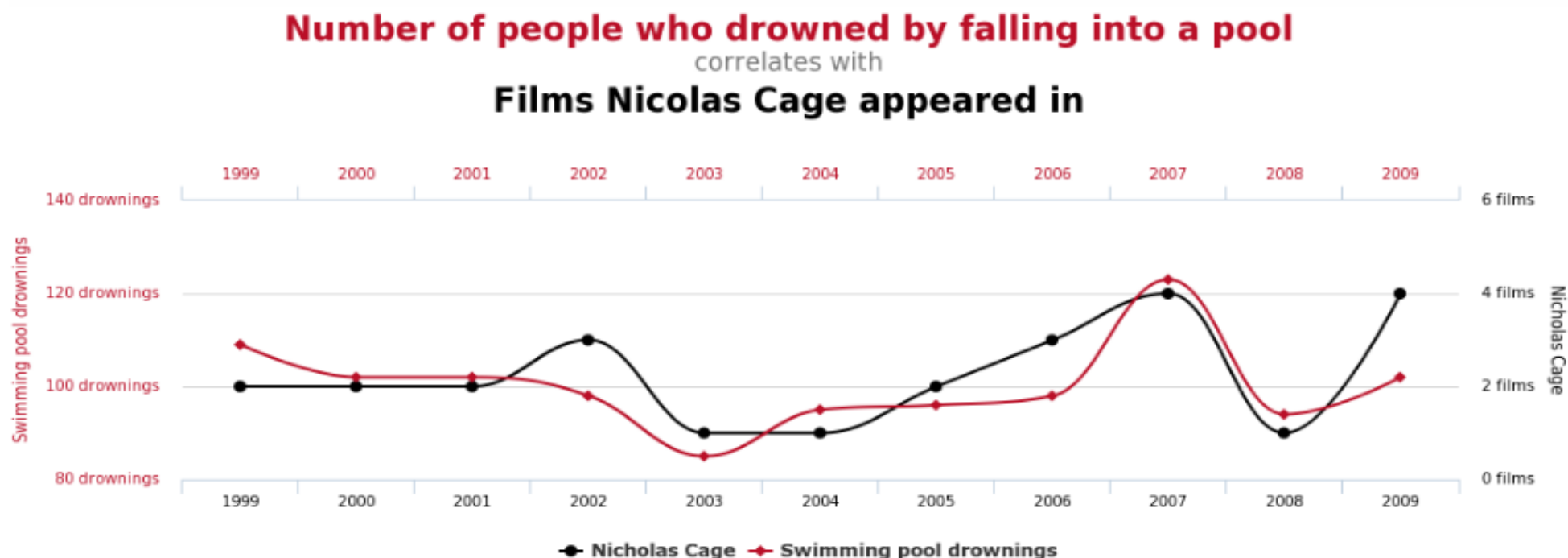
¿Cuál de estas
afirmaciones
genera
mayor
credibilidad?

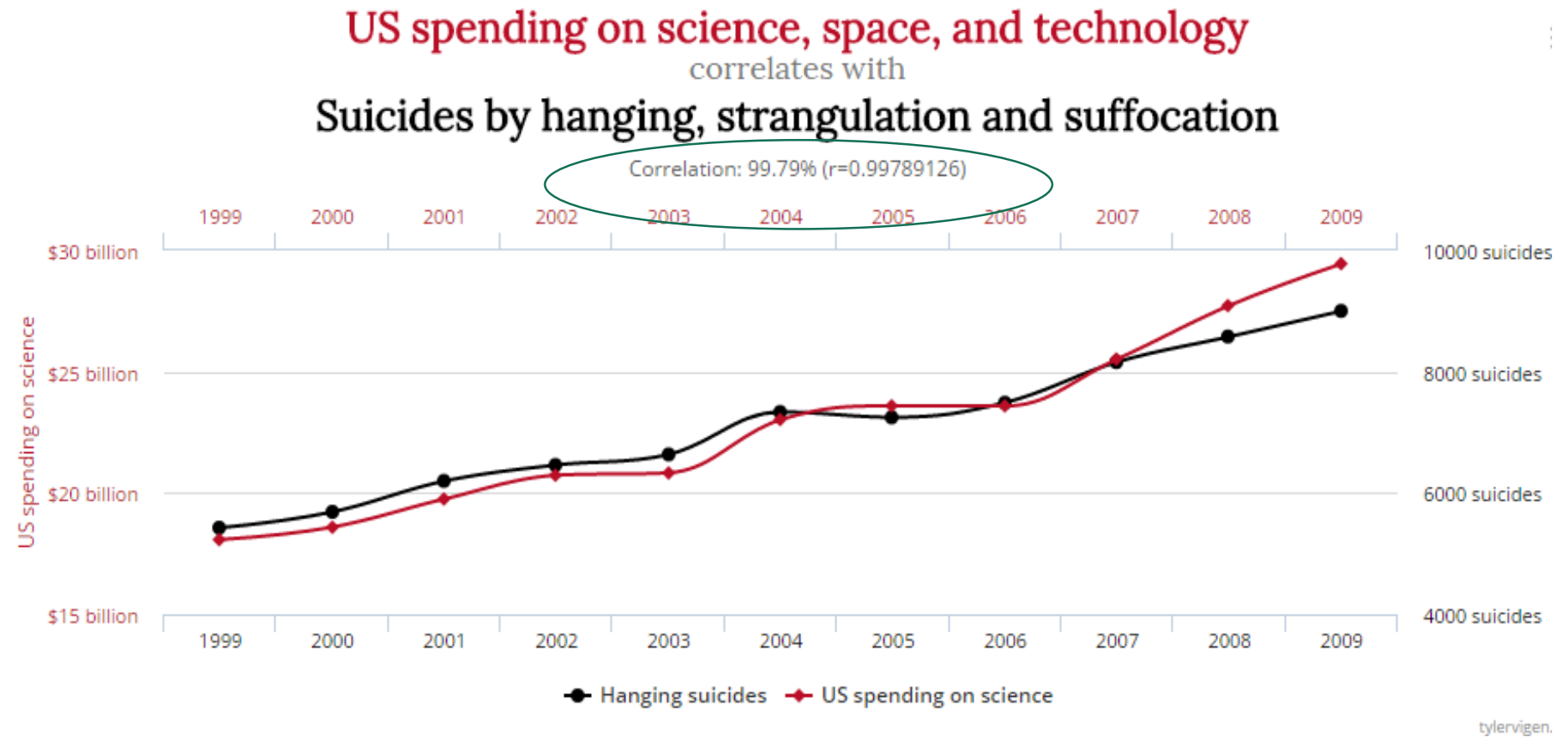
- ¿Sabía usted que hay una correlación muy alta entre el número de películas en las que aparece **Nicholas Cage** y el número de personas que se ahogan al caerse en una piscina en EEUU?
- Un estudio afirma que los niños que duermen con la luz encendida tienen más posibilidades de padecer miopía

SALUD ›

Ver películas de Nicolas Cage aumenta el riesgo de ahogarse en la piscina

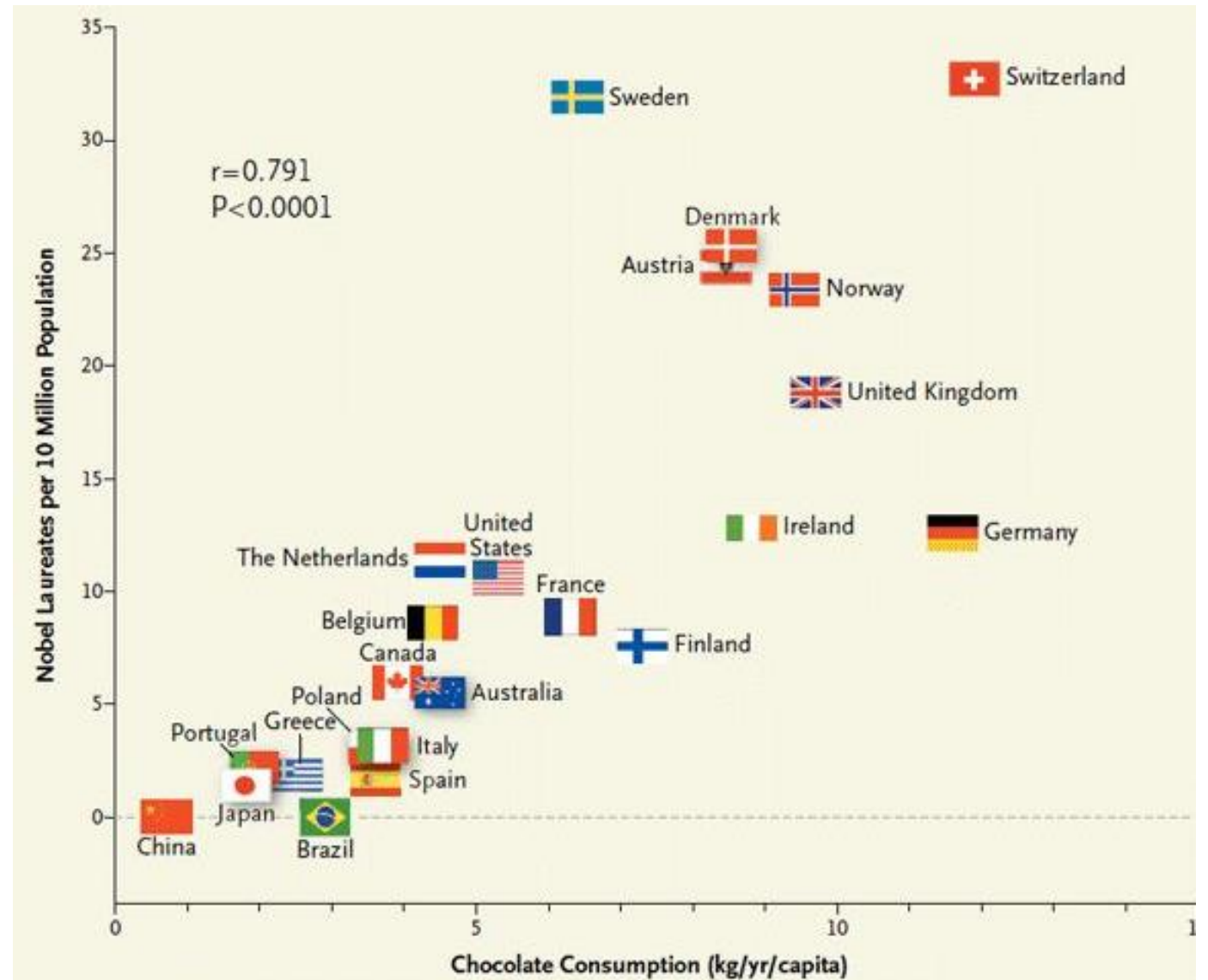
Saber leer las estadísticas y entrenar el espíritu crítico servirá para no confundir casualidad con causalidad



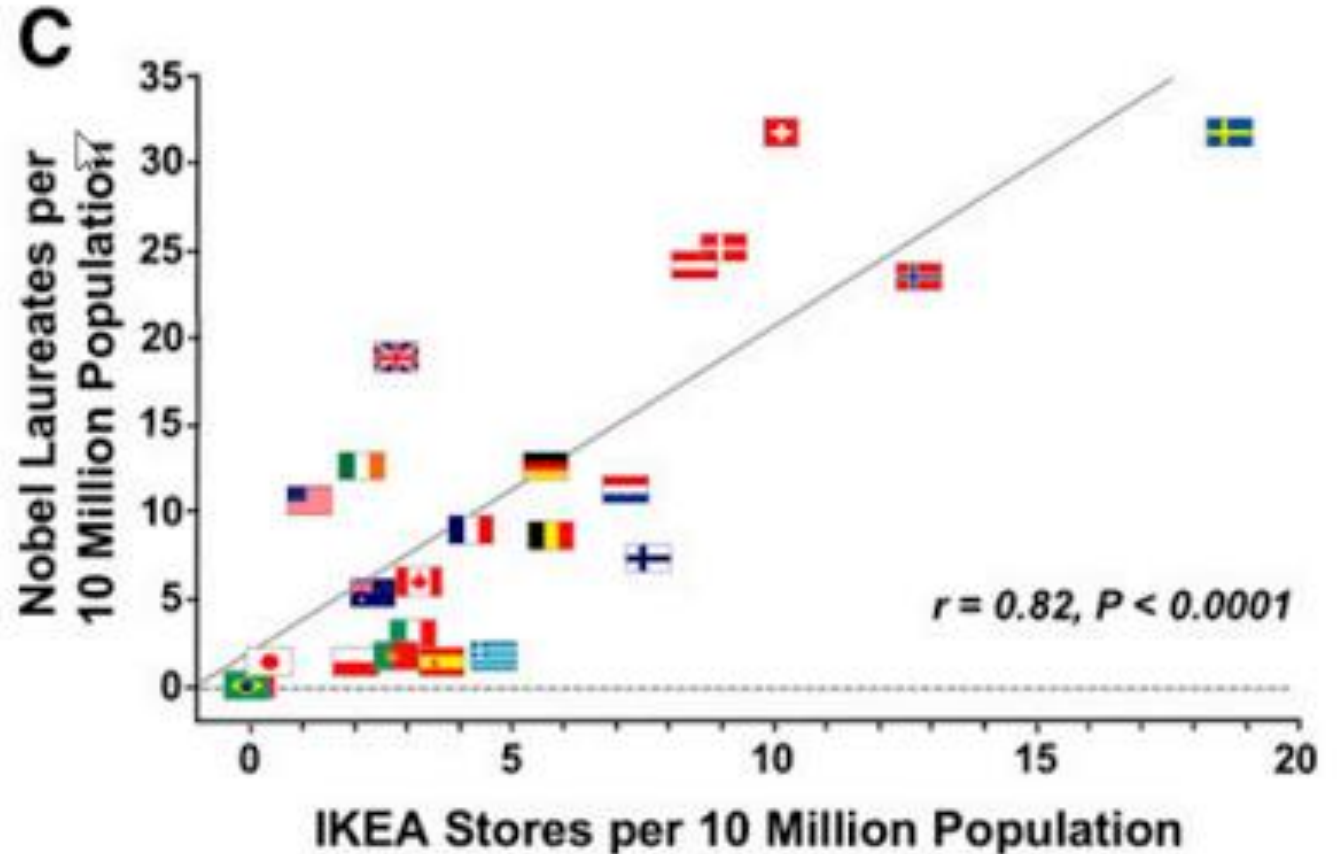


Inversión en ciencia y tecnología vs Suicidios por ahorcamiento y asfixia

Comer más
chocolate,
¿el secreto
para ganar
más premios
Nobel?





Posible
relación entre
número de
premios Nobel
y número de
tiendas de
IKEA en cada
país





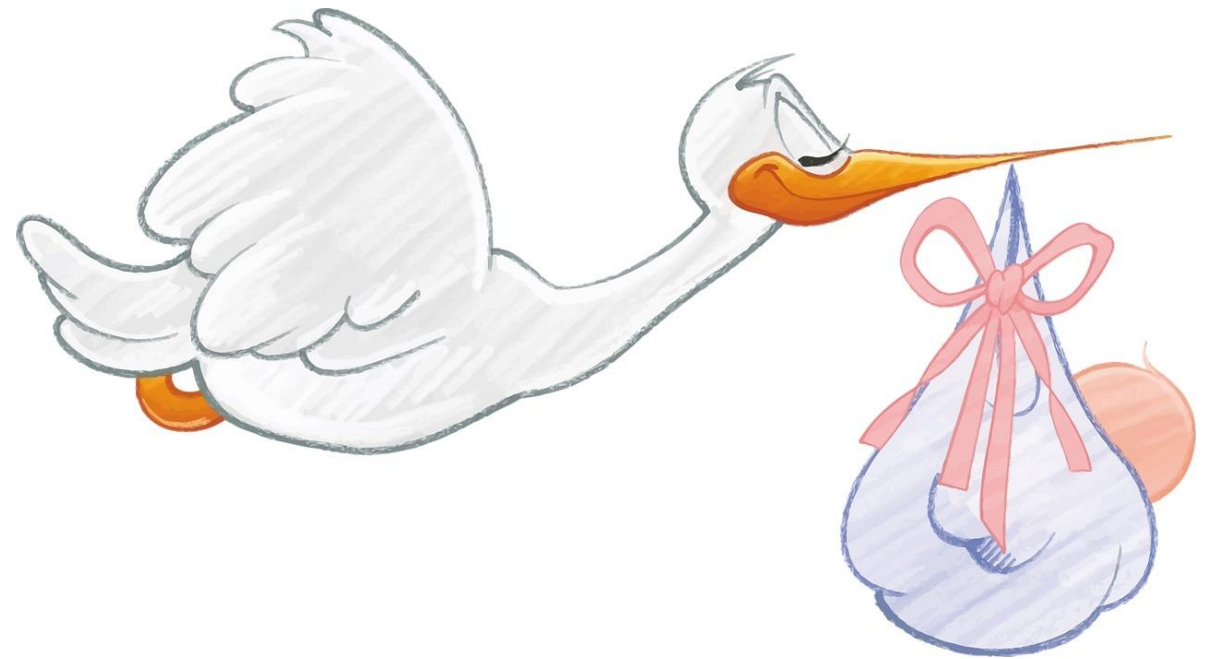
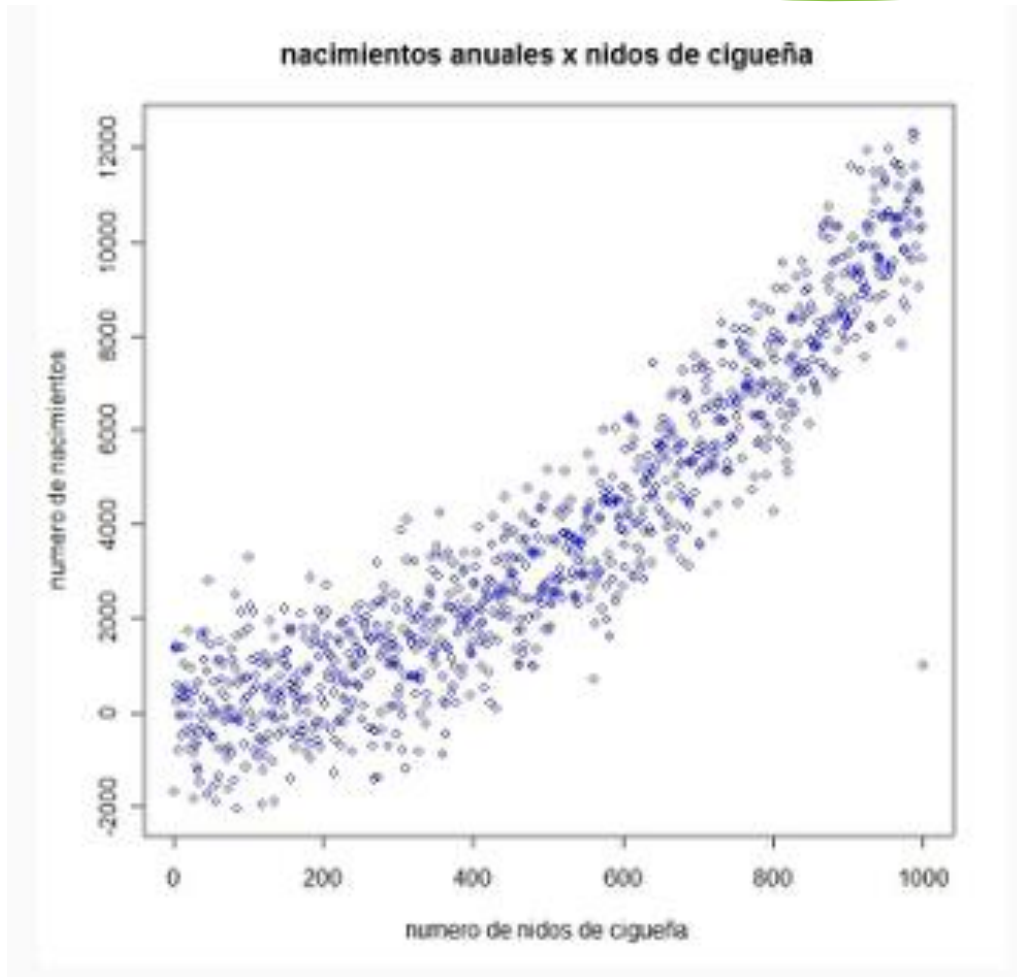
¿Y, si no es A la causa de B,
por qué se dan los dos
fenómenos a la vez de
forma repetida?



En general, si hay una fuerte correlación entre los fenómenos A y B, tenemos cuatro posibilidades:

- **Que A cause B** (que los ahogamientos en piscinas hagan que el bueno de Nicolas quiera hacer más cine para animar a las familias).
- **Que B cause A** (Que la gente quiera ahogarse después de ver alguna de sus películas)
- **Que haya un tercer fenómeno, C**, que provocara tanto A como B (es complicado imaginar alguno, pero a lo mejor el Orden Mundial conspira para reducir la población humana tanto mediante el ahogamiento como mediante el aburrimiento).
- **Puro azar.** Hay muchos datos en el mundo, así que si los comparamos todos más tarde o más temprano encontraremos este tipo de correlaciones que no significan nada.

Correlación entre el número de cigüeñas y el número de nacimientos



Cuanto más habitantes hay, la construcción de edificios (iglesias, campanarios, pisos...) es mayor y por lo tanto aumenta la cantidad de cigüeñas. A su vez, al haber más habitantes, también se incrementa el número de nacimientos.



Causalidad
siempre implica
correlación, pero
la correlación no
necesariamente
implica causalidad

La clave para poder **asociar correlación con causalidad de manera rotunda** es estar seguros de que la única causa posible de B es A... o que tiene más causas pero **todas ellas son independientes de A.**

El problema de creer que una fuerte correlación implica una cierta relación causal es extraer conclusiones equivocadas, es atribuir ciertas propiedades irreales a un producto

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

Para estimar ρ usamos r

*X y Y variables aleatorias
Normales*

Coeficiente de correlación poblacional

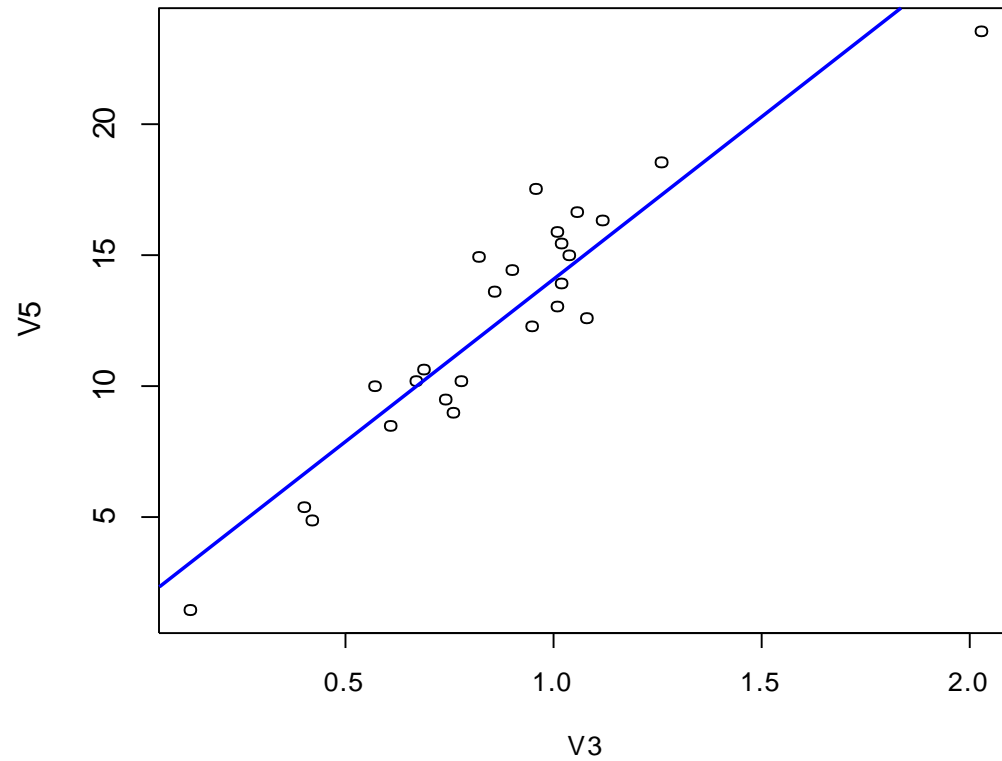
Inferencia sobre r

$H_0 : \rho = 0$ (no existe correlación lineal)

$H_1 : \rho \neq 0$ (existe correlación lineal)

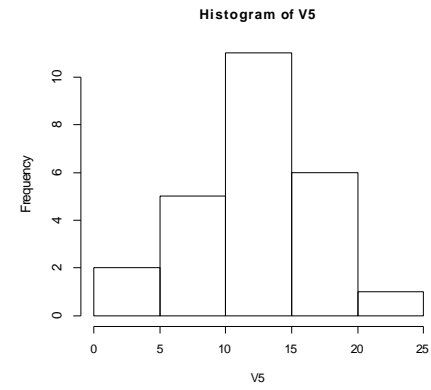
Supuestos:

- La relación entre las variables es lineal
- La población sigue la distribución Normal.
- Los datos son independientes entre sí



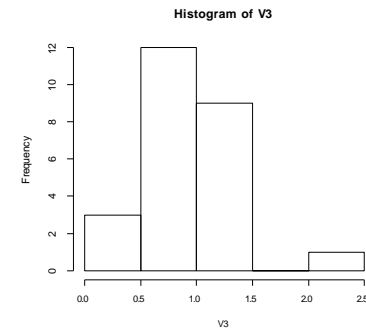
Pearson's product-moment correlation

```
data: V5 and V3
t = 11.7587, df = 23, p-value = 3.312e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8370787 0.9672061
sample estimates:
      cor
0.9259473
```



Shapiro-Wilk normality test

```
data: V5
W = 0.9803, p-value = 0.8904
```



Shapiro-Wilk normality test

```
data: V3
W = 0.9023, p-value = 0.02065
```

Cuando las dos variables bajo del estudio de correlación no tienen distribución normal se procederá con los rangos de mediciones para cada variable. Hay dos métodos de rango de correlación, uno de Spearman y otro de Kendall

Ecuación.

$$\tau = (S_a - S_b) / [n(n - 1) / 2]$$

Donde,

τ = Estadística de Kendall

n = # de casos en el ejemplo

S_a = Sumatoria de rangos más altos

S_b = Sumatoria de rangos más bajos

**Coeficiente
de Kendall**

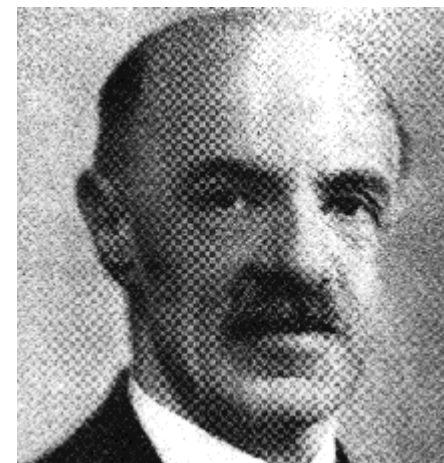


Maurice George Kendall

$$r_s = 1 - [6 \sum d_i^2 / (n^3 - n)]$$

Donde, d_i = diferencia entre rangos de X y Y.

**Coeficiente
de Spearman**



Charles Edward Spearman

Otros coeficientes de correlación

Coeficiente de correlación de Spearman

Lo que tenemos ahora son 2 sucesiones de valores ordinales.

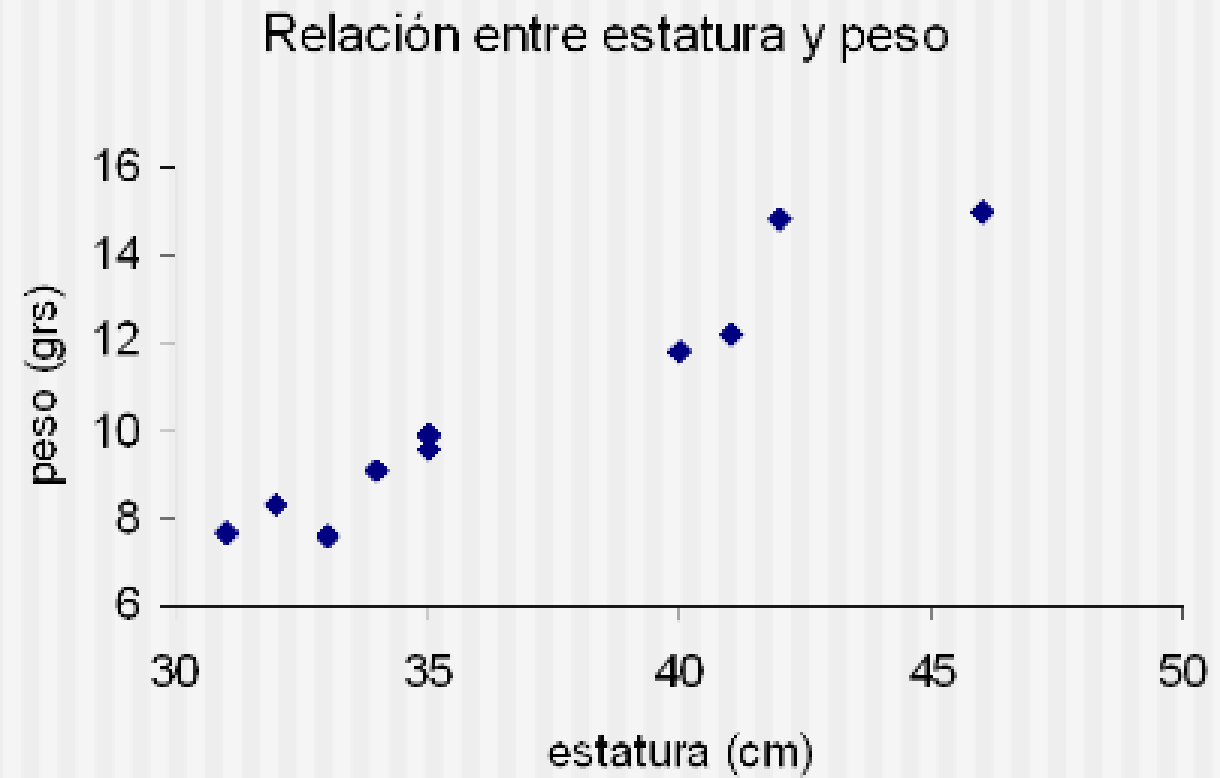
El coeficiente de Spearman es un caso especial del coeficiente de correlación de Pearson aplicada a dos series de los n primeros números naturales (cuando no hay empates; si hay –muchos- empates hay otra fórmula)

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

d_i es la diferencia entre el valor ordinal en X y el valor ordinal en Y del sujeto i

Ejemplo: Relación entre estatura y peso

centímetros	gramos
31	7.7
32	8.3
33	7.6
34	9.1
35	9.6
35	9.9
40	11.8
41	12.2
42	14.8
46	15.0



Ejemplo: Relación entre estatura y peso

estatura		peso		diferencia de rango d
centímetros	rango	gramos	rango	
31	1	7.7	2	-1
32	2	8.3	3	-1
33	3	7.6	1	2
34	4	9.1	4	0
35	5.5	9.6	5	0.5
35	5.5	9.9	6	-0.5
40	7	11.8	7	0
41	8	12.2	8	0
42	9	14.8	9	0
46	10	15.0	10	0

Calculemos ahora: $r_s = 1 - \frac{6 \cdot \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$