

# Laboratorio de datos, clase 6

## Inferencia, predicción, y regresión lineal

Prof. Enzo Tagliazucchi

[tagliazucchi.enzo@gmail.com](mailto:tagliazucchi.enzo@gmail.com)

[www.cocuco.org](http://www.cocuco.org)



“Laboratorio de datos... quiero ver más allá de lo evidente”

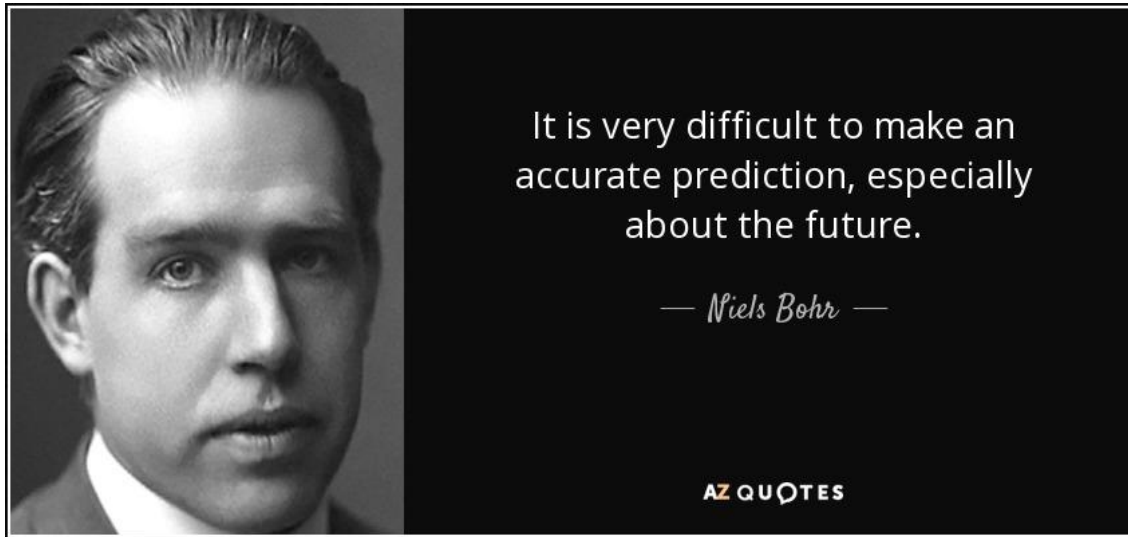
$Y$  : variable dependiente

$X = (\bar{X}_1, X_2, \dots, X_p)$  : variables independientes (o predictores)

$$Y = f(X) + \epsilon$$

Puedo *modelar* la variable  $Y$  mediante la función  $f()$  más un término de error.

¿Para qué sirve?



# Predicción

$Y$  : variable dependiente

$X = (\bar{X}_1, X_2, \dots, X_p)$  : variables independientes (o predictores)

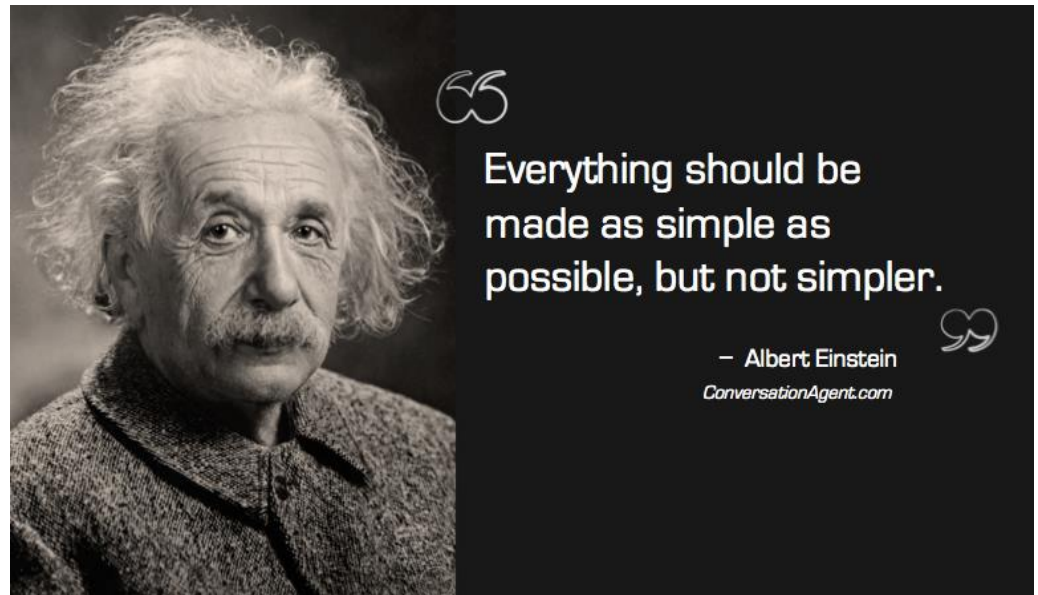
} En el presente

$$Y = f(X) + \epsilon$$

Puedo usar esta relación para calcular el valor de  $Y$  para  $X$  que todavía no tengo, es decir, para *predecir* el valor de  $Y$ .

Típicamente no me interesa la forma de  $f()$ , ni su simplicidad, siempre que pueda demostrar que la predicción funciona (enfoque tipo “machine learning”)

# Inferencia



$$Y = f(X) + \epsilon$$

¿Qué podemos decir sobre la relación entre X e Y obteniendo e interpretando la función f()?

¿Todas las variables X importan para determinar Y? ¿Cuánto?

¿Cómo es la relación entre las variables X e Y? (e.g. creciente, decreciente...)

¿Qué me dice la forma función de f() sobre el proceso que relaciona los datos?

# Inferencia: un ejemplo práctico

$Y$  : cantidad de ventas de un producto


$X = (\bar{X}_1, X_2, \dots, X_p)$  : variables independientes (o predictores), por ejemplo:

Precio de venta del producto

Inversión en publicidad (por medio)

Precio de venta de la competencia

Unidades vendidas por la competencia


$$Y = f(X) + \epsilon$$

¿Conviene subir el precio para vender más?

¿Estamos gastando poco o mucho en publicidad?

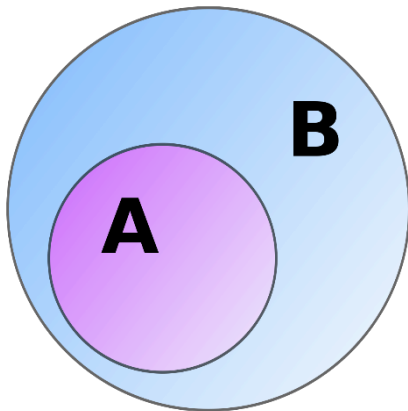
¿Sirve la publicidad radial, o mejor solo invertir en televisión?

¿Qué pasa si a nuestros competidores les va mejor o peor?



**Inferencia:**

“Las ovejas negras del país X son negras con Y grado de confianza”



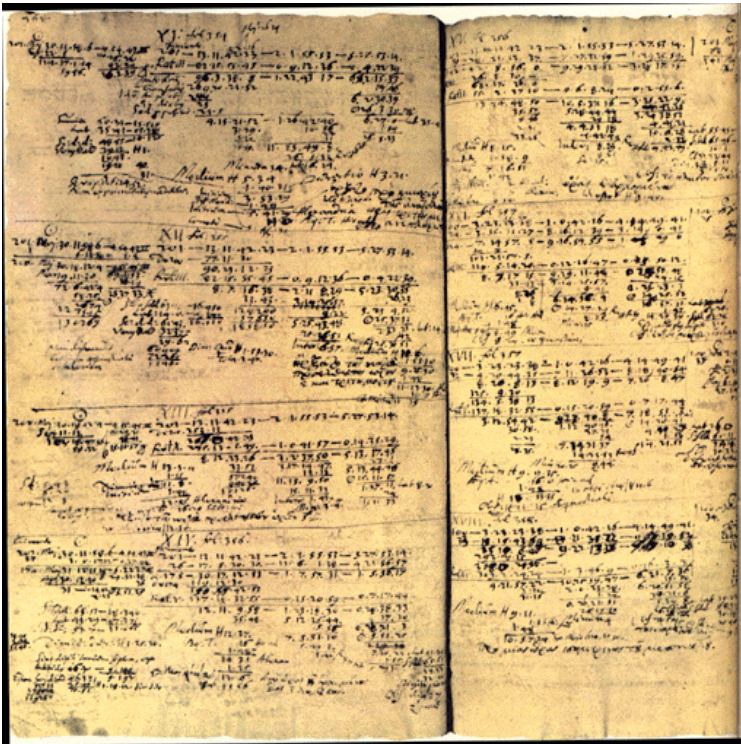
**Estadística descriptiva:**

“En el país X hay al menos tres ovejas al menos cuyo costado derecho es negro”

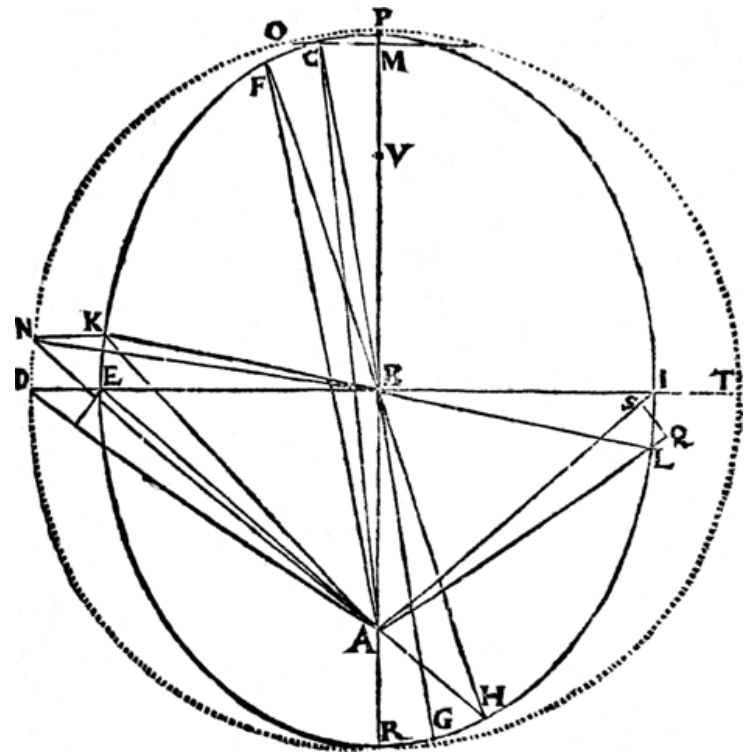


# Hace más de pocas décadas...

... había que trabajar muy duro para que los datos existan



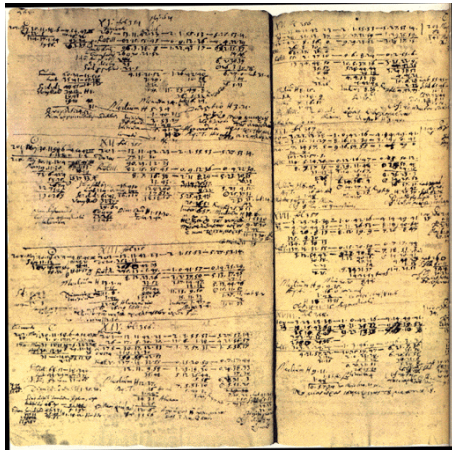
Tycho Brahe (1546 -1601)



Johannes Kepler (1571 -1630)

# Datos

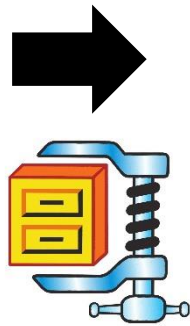
Voluminosos, difíciles de usar, entender, formarse intuiciones, etc.



# Modelo

Fácil de extrapolar a valores nuevos.  
Es clara la relación funcional entre los datos, qué importa, qué no, etc.

Inferencia



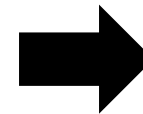
$$r = \frac{a(1-e^2)}{(1+e \cos \theta)}$$

for  $0 \leq e < 1$

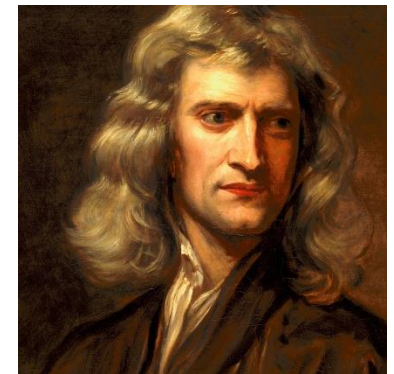


# Dinámica

Una ley dinámica resume una cantidad infinita de modelos mediante ecuaciones diferenciales.



$$\frac{dv}{dt} = \frac{F^{net}}{m}$$
$$F_g = \frac{Gm_1m_2}{r^2}$$



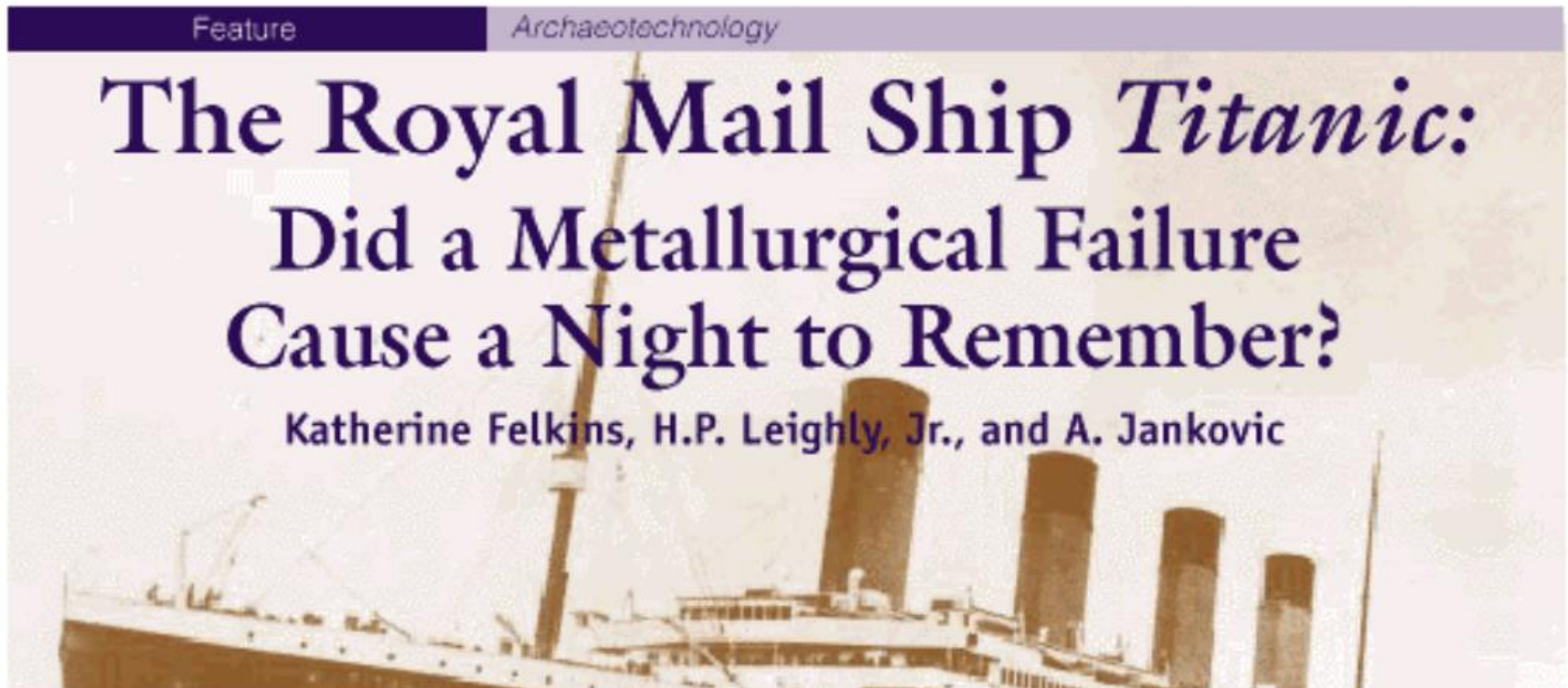


*"In general we look for a new law by the following process. First, we guess it. Then we compute the consequences of the guess, to see if this law that we guessed is right, we see that it would imply. And then we compare the computation results to Nature, or we say compare to experiments, or to experience. Compare it directly with observation, to see if it works. If it disagrees with experiments, it's wrong. In that simple statement is the key to science. It doesn't make a difference how beautiful your guess is, it doesn't make any difference how smart is who made the guess, or what his name is. If it disagrees with experiment, it's wrong. That's all there is to it".*



Richard Feynman (1918-1988)

# Correlación (estadística) vs. regresión (inferencia)



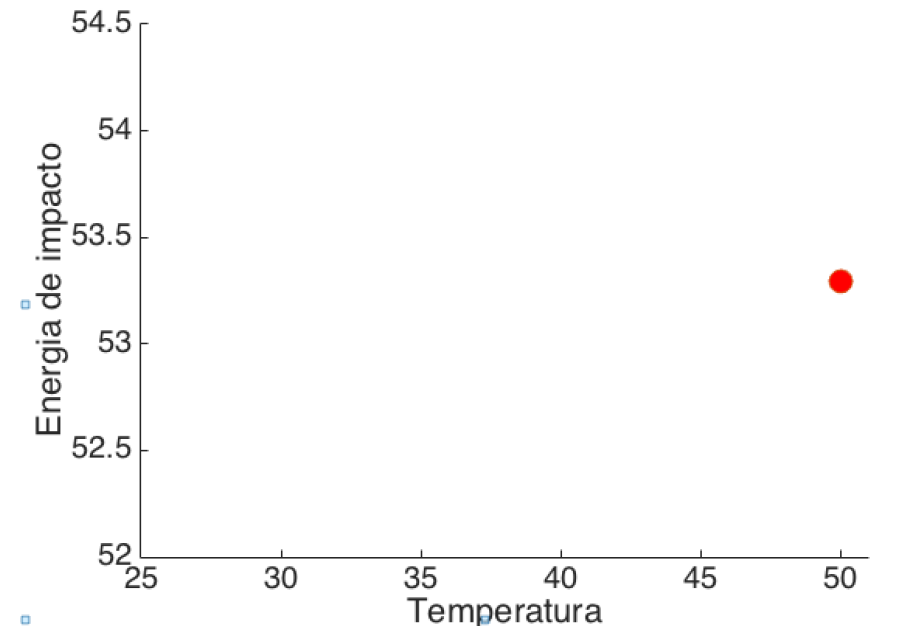
*“La fragilidad del acero (medida como la energía de un golpe necesaria para romperlos) depende de la temperatura”*

Medimos la fragilidad (F) de una pieza de acero y medimos la temperatura de dicha pieza (T):

T: 50°

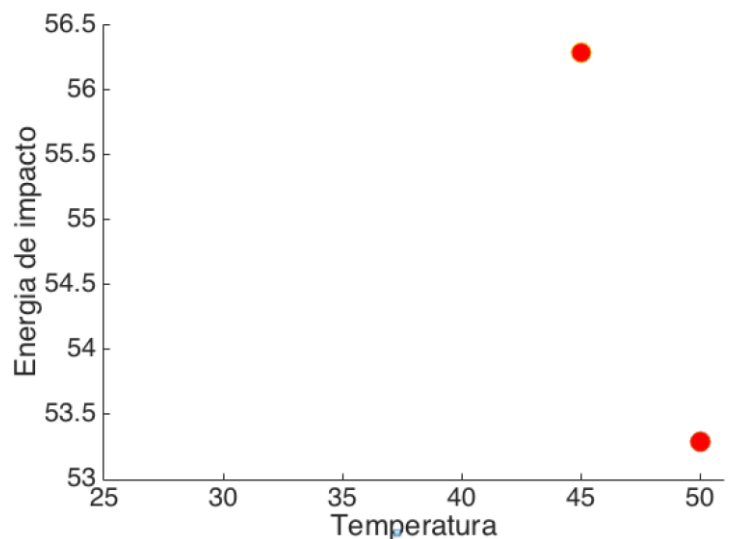
F: 53.2

... e indicamos la medición con un punto de coordenadas x=50, y=53.2.



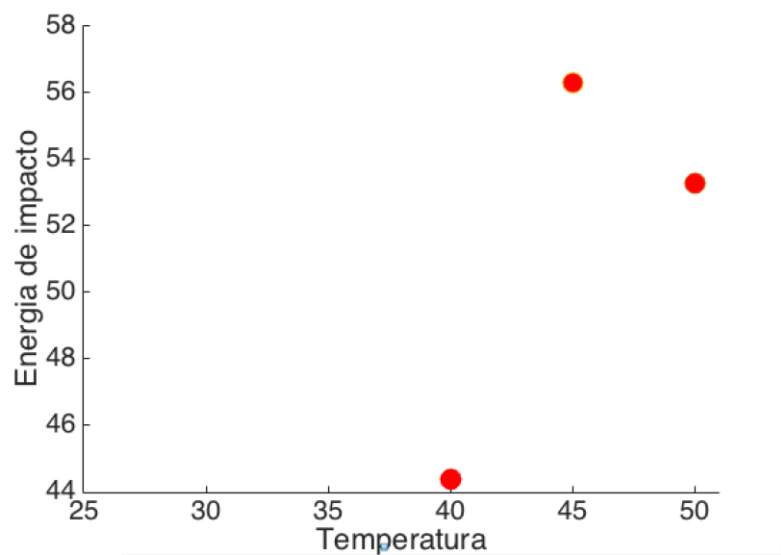
T: 50°, 45°

F: 53.2, 56.2



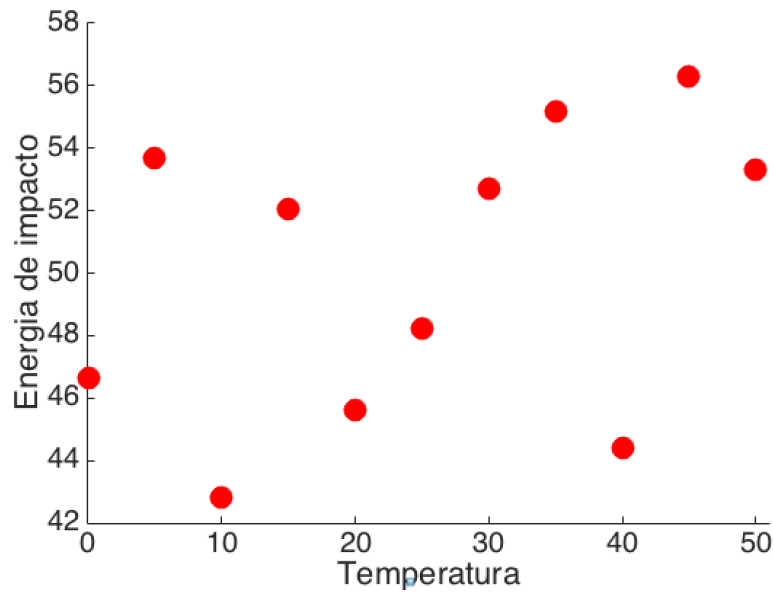
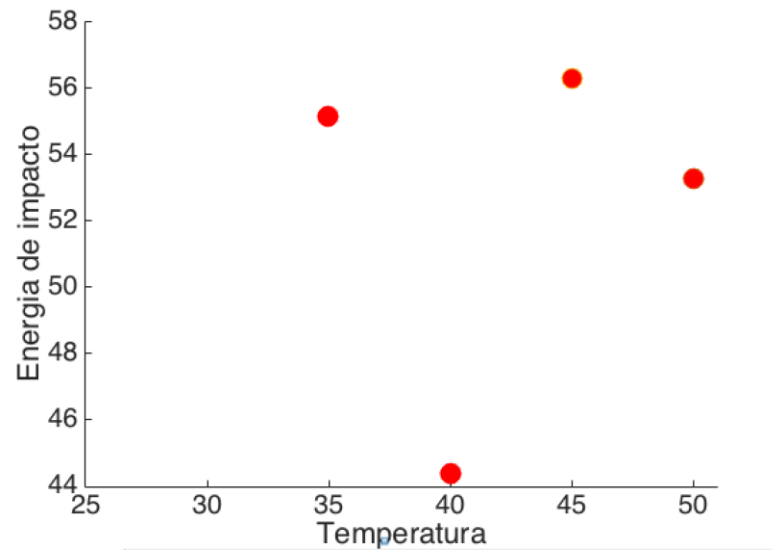
T: 50°, 45°, 40°

F: 53.2, 56.2, 44

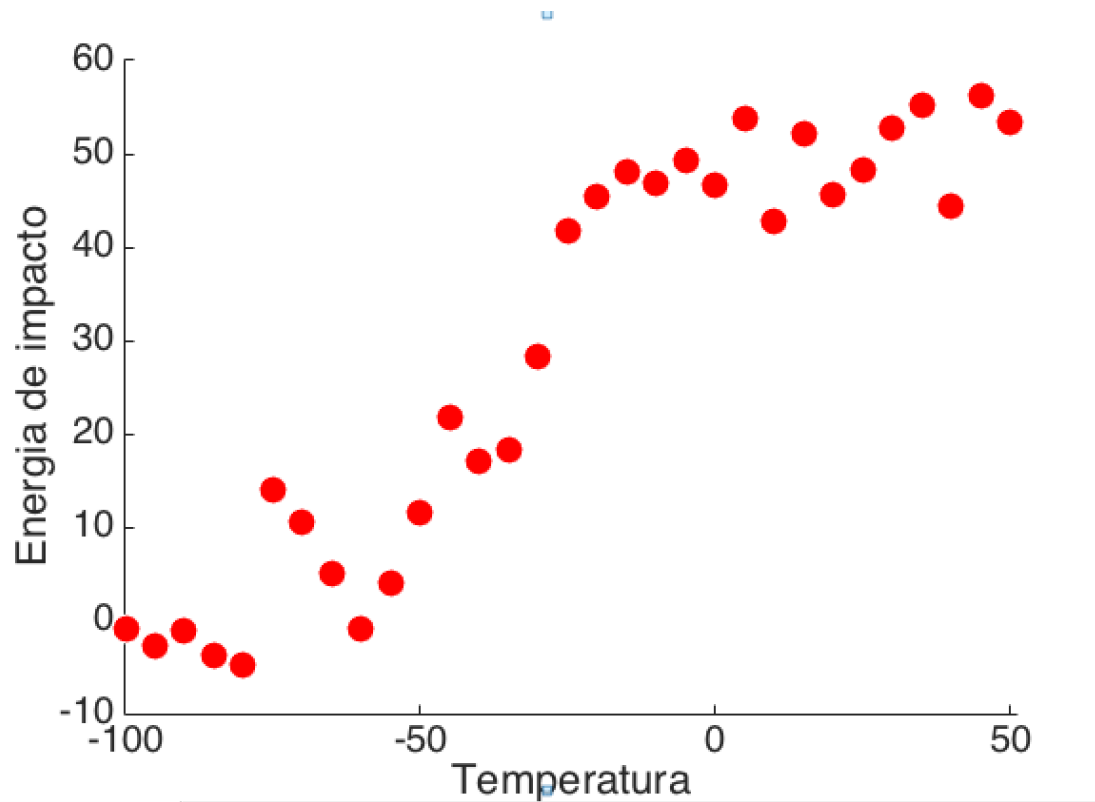


T: 50°, 45°, 40°, 35°

F: 53.2, 56.2, 44, 55

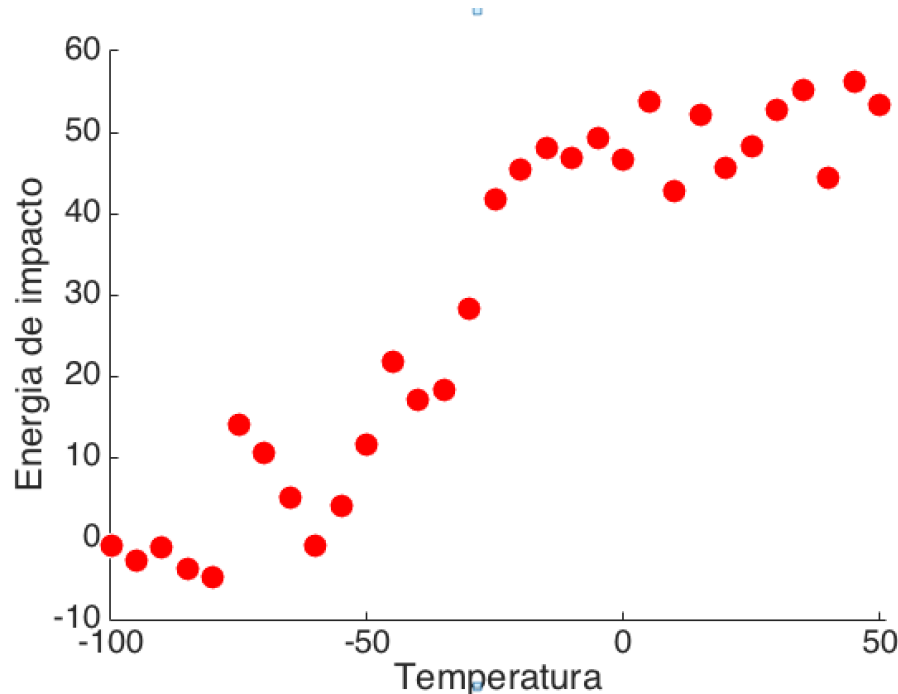






T:	-100	-95	-90	-85	-80	-75	-70	-65	-60	-55	-50	-45	-40	-35	-30	-25	-20	-15	-10	-5	0	5	10	15	20	25	30	35	40	45	50
F:	-1	-2.8	-1.2	-3.7	-4.8	14	10.4	5	-0.8	4.1	11.6	21.7	17.1	18.3	28.3	41.8	45.5	48.1	46.7	49.3	46.6	53.7	42.8	52.1	45.6	48.2	52.7	55.2	44.4	56.3	53.3

“La fragilidad del acero (medida como la energía de un golpe necesaria para romperlos) depende de la temperatura”



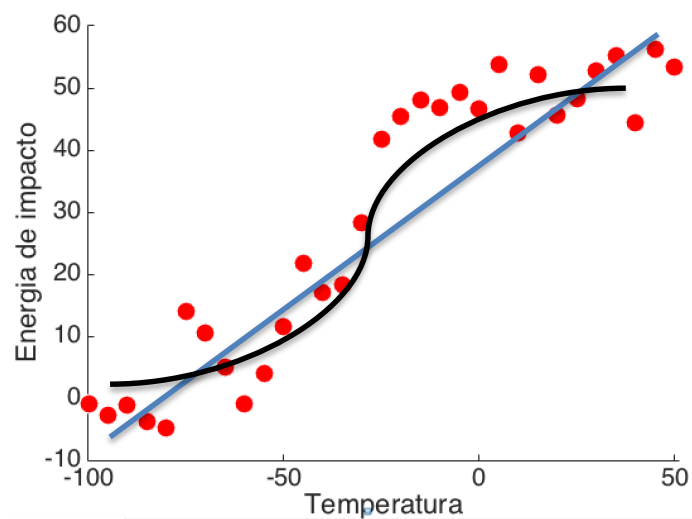
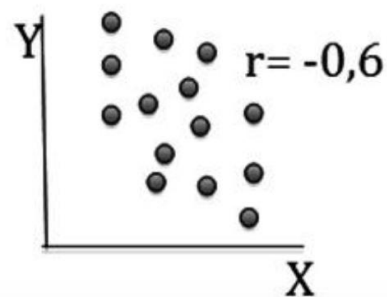
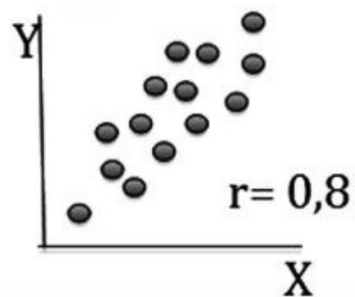
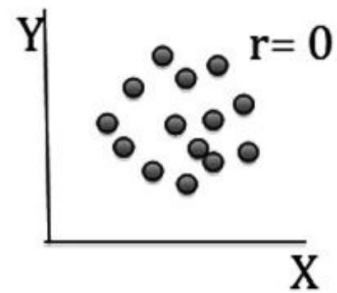
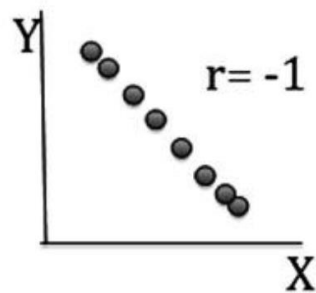
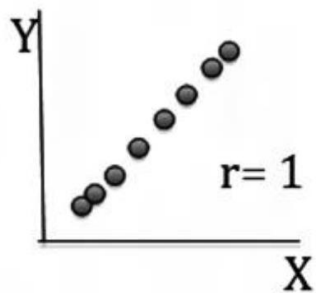
= ¿están las variables *correlacionadas*?

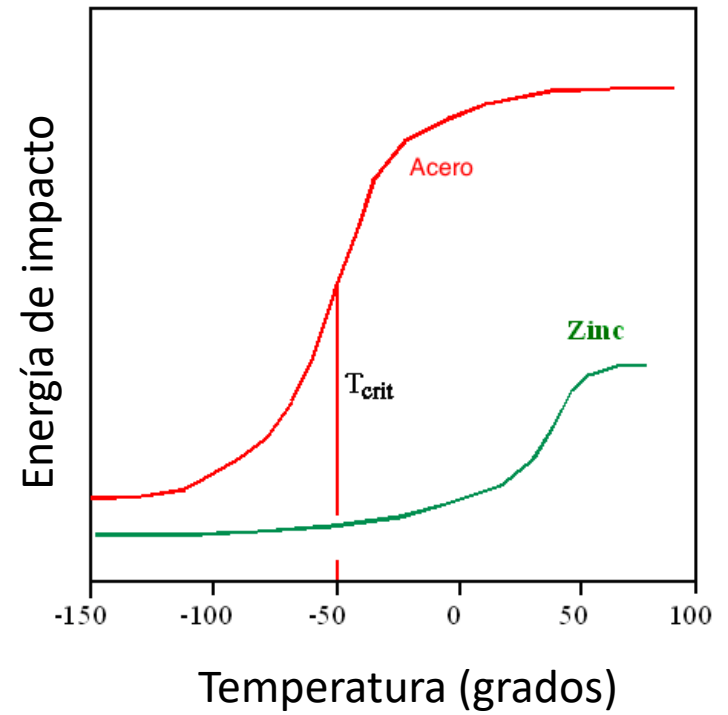
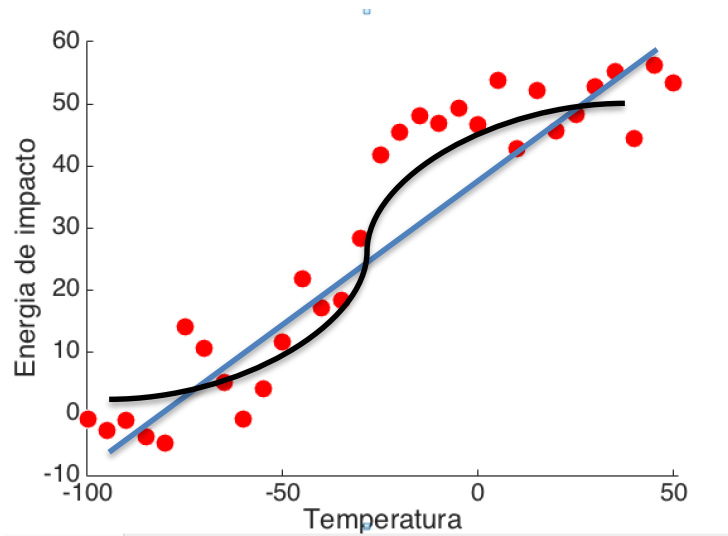
Correlación lineal o de Pearson:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$R=0.93$$

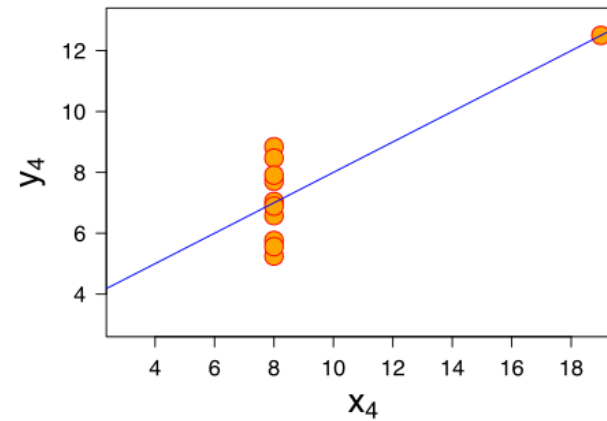
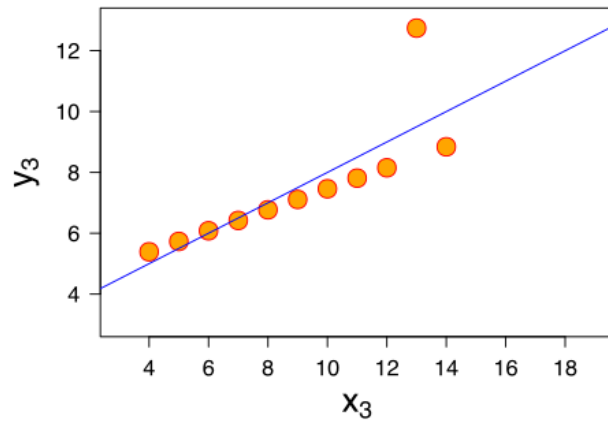
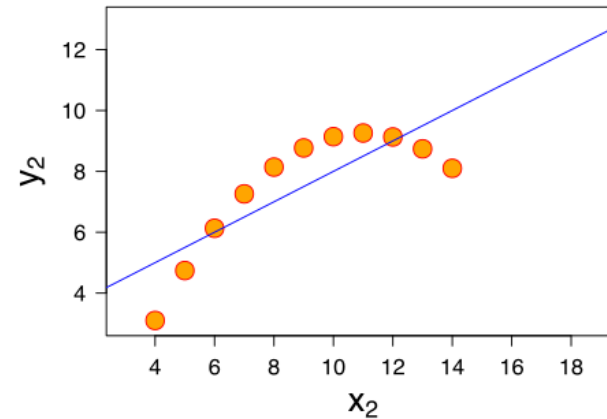
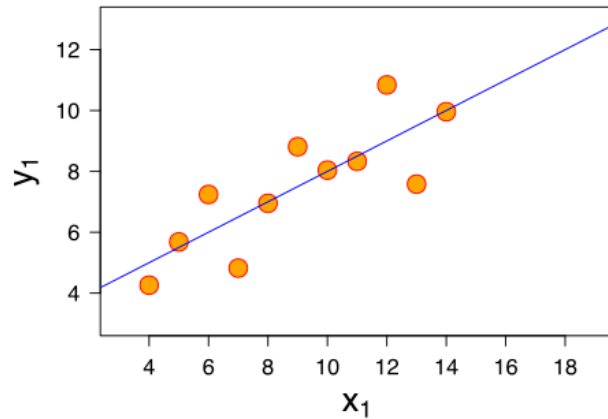
$$r = \cos(\alpha) = \frac{\sum_{i=1}^N (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^N (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^N (y_i - \bar{y})^2}}$$





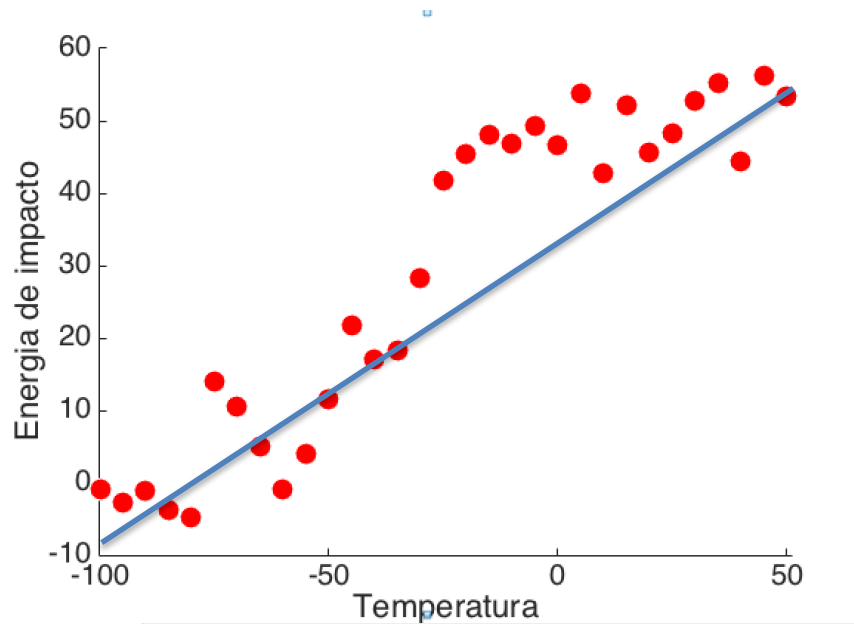
# Cuarteto de Anscombe

(misma correlación lineal, distintas relaciones funcionales)

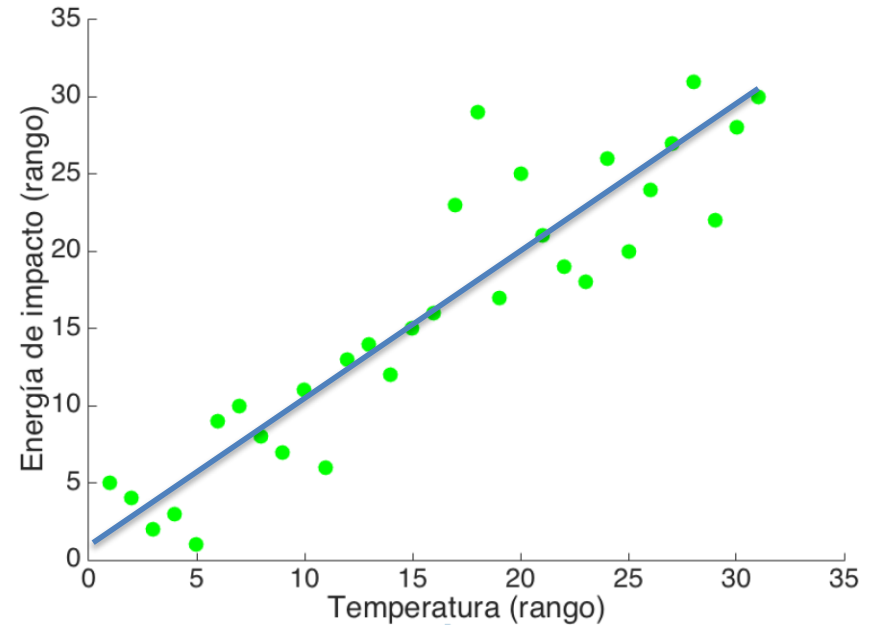




## Pearson



## Spearman



Esto me sugiere que existe una relación en los datos y que además es monótona creciente. Pero ¿cuál es?

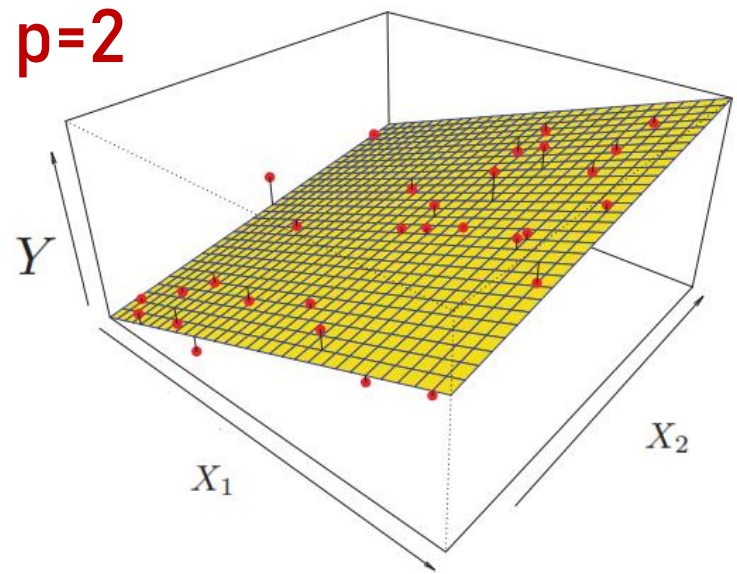
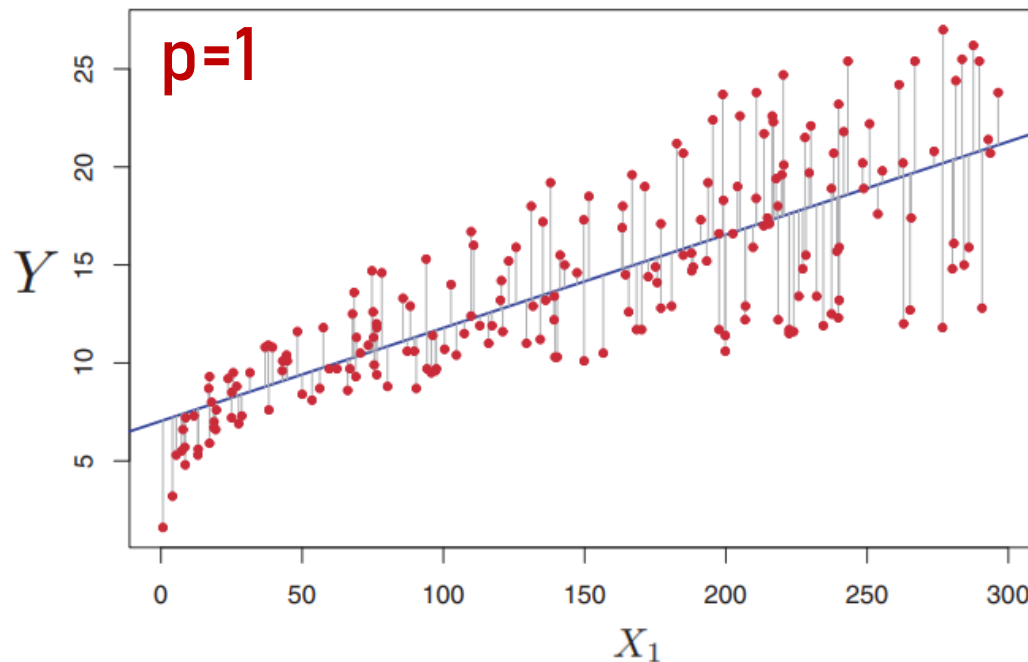
$$Y = \boxed{f(X)} + \epsilon$$

$$Y = f(X) + \epsilon$$

Regresión lineal: la función  $f()$  es una función lineal

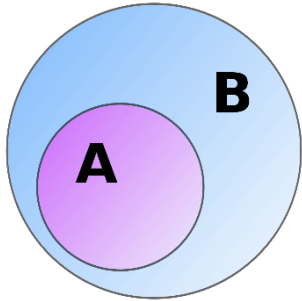
$$f(X) = \underline{\beta_0} + \underline{\beta_1}X_1 + \underline{\beta_2}X_2 + \dots + \underline{\beta_p}X_p$$

p variables  
p+1 parámetros



$Y \approx \beta_0 + \beta_1 X$  : esta es la relación matemática real entre las variables

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  : esto es lo que yo estimo a partir de mi muestra



¿Qué validez tiene mi inferencia de los parámetros?

Pero antes... ¿cómo infiero los parámetros?

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  : los datos disponibles

$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  : si tuviese los parámetros podría estimar la variable dependiente a partir de la variable independiente

$e_i = y_i - \hat{y}_i$  : luego podría computar el error del estimativo (que depende de los parámetros estimados). Se conocen como residuos.

$RSS = e_1^2 + e_2^2 + \dots + e_n^2$  : y luego, la suma de los residuos al cuadrado

Esta suma depende de los dos parámetros del modelo lineal:

$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

Más chica es, el modelo mejor reproduce los datos de los que ya dispongo.

¿Cuáles son los parámetros tal que esta suma es lo más pequeña posible?

Modelos lineales:

Podemos derivar respecto de los parámetros (por eso computamos el error cuadrático), igual a cero y despejar:

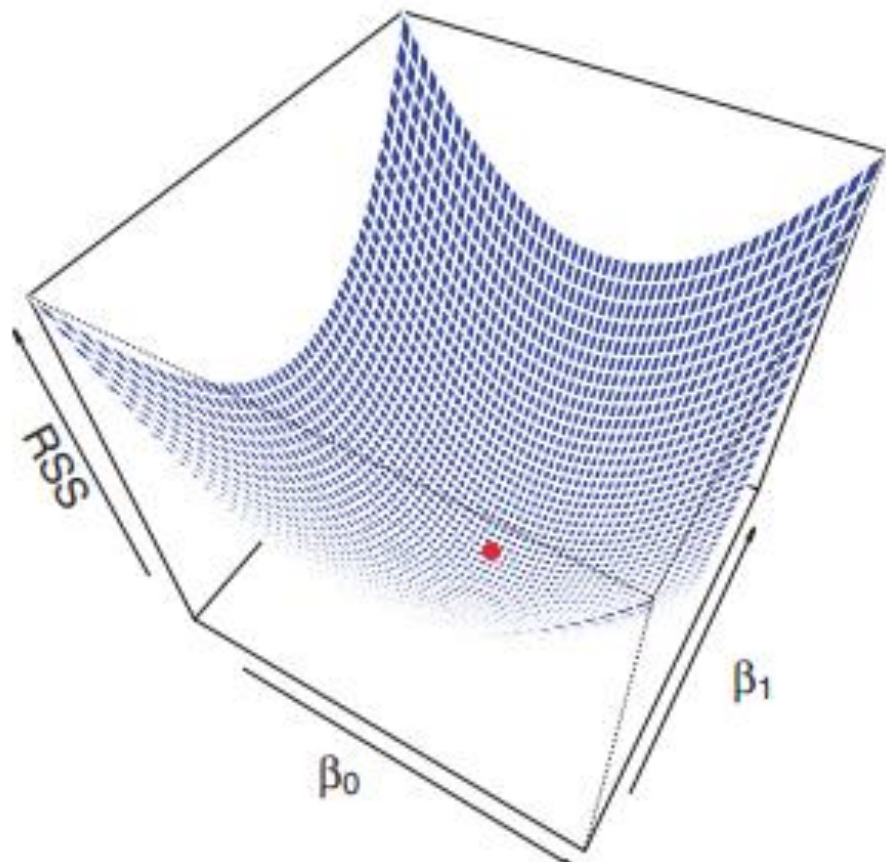
$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} & \bar{y} &\equiv \frac{1}{n} \sum_{i=1}^n y_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} & \bar{x} &\equiv \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

Modelos más complicados  
(casi todos salvo este):

Va a ser necesario recurrir a un proceso de optimización (por ejemplo, *gradient descent*)

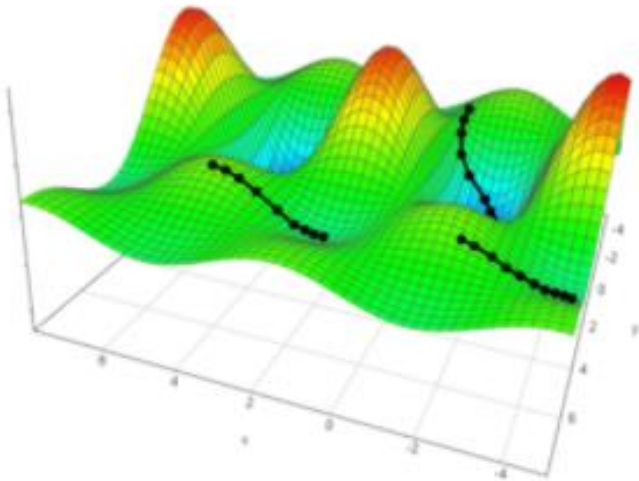
Aprendemos los parámetros de los datos: *machine learning*

Cuadrados mínimos (Gauss, 1822)



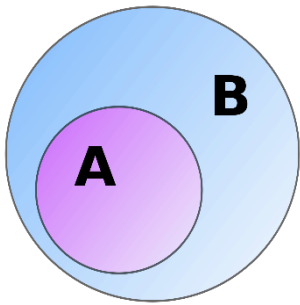
Los parámetros resultan de minimizar la suma de los errores cuadráticos.

*En este caso, hay un único mínimo local (igual al global) y se puede computar analíticamente, pero no siempre.*

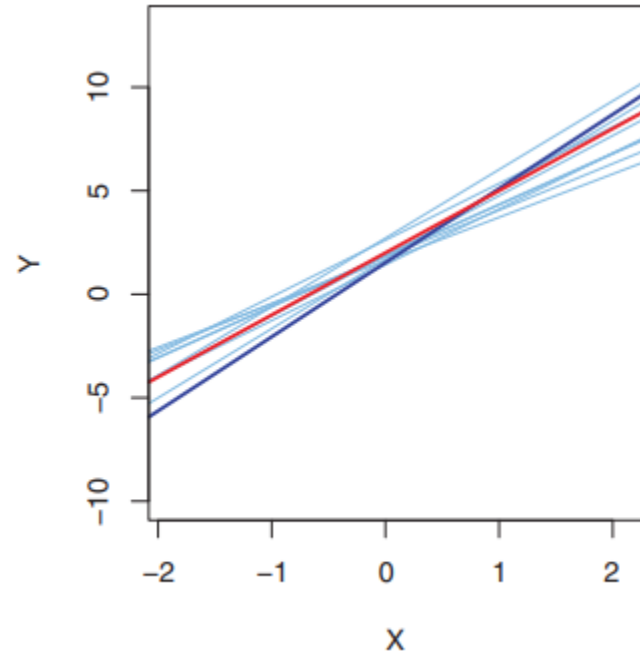
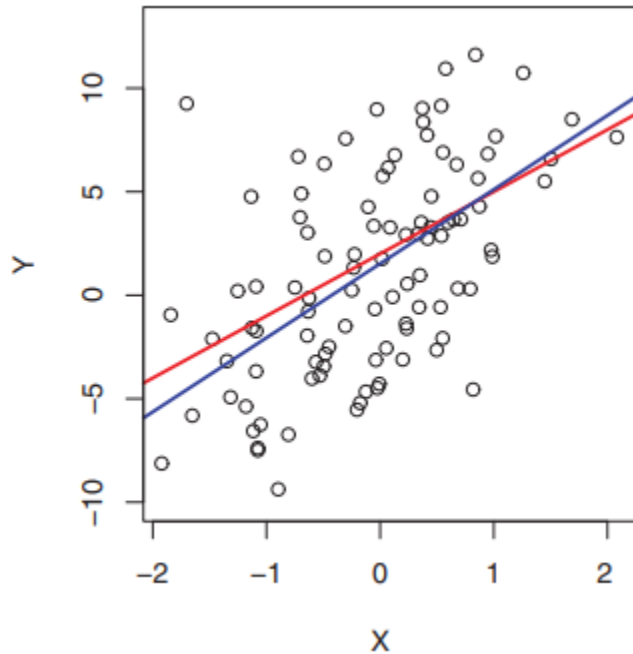


Sucede en algunos algoritmos de *machine learning* (e.g. redes neuronales) pero en otros no (e.g. *support vector machines*)





¿Qué validez tiene mi inferencia de los parámetros?



— Relación lineal real  
o poblacional,  $Y \approx \beta_0 + \beta_1 X$

— Relación lineal  
estimada,  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

—  
—  
—  
—  
— } Relaciones lineales  
estimadas utilizando  
distintos subconjuntos de  
los datos

# Analogía (clase de 4, estadística descriptiva)

$Y$  : una variable aleatoria

$\mu$  : el valor medio de la variable aleatoria

$\hat{\mu} = \bar{y}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  : un estimador *no sesgado* de este valor medio

$SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$  : diferencia entre el valor medio y el estimador

---

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

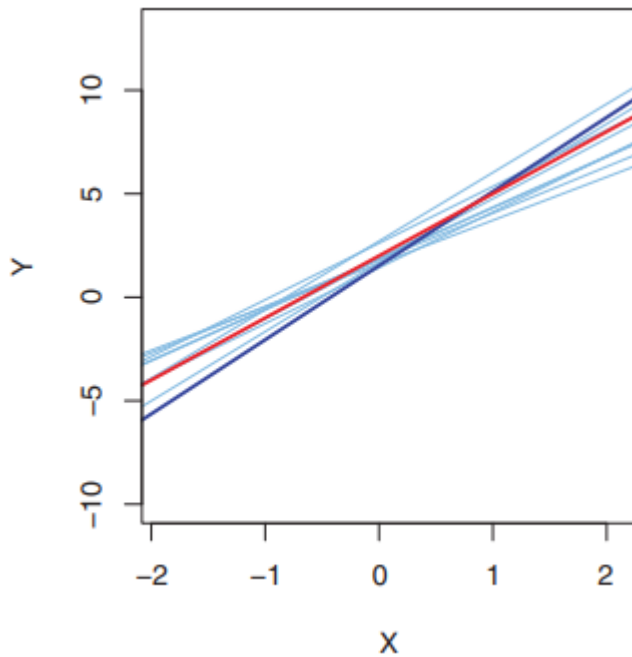
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Estimadores no sesgados para los parámetros del modelo

$$SE(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Diferencia entre los estimadores y los valores reales



Se van aproximando a la línea roja a medida que el número de datos usados para obtener el estimador aumenta

Hay un 95% de probabilidades de que los valores reales de los parámetros se encuentren dentro de los intervalos:

$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

$$\hat{\beta}_0 \pm 2 \cdot SE(\hat{\beta}_0)$$

Suma total de cuadrados

$$R^2 = \frac{\boxed{\text{TSS}} - \text{RSS}}{\text{TSS}}$$

Es la proporción de la variabilidad en Y explicada por X.

En el caso lineal, es igual al coeficiente de Pearson al cuadrado.

# Caso multilineal

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Para una estimación de parámetros, podemos calcular el Y estimado,

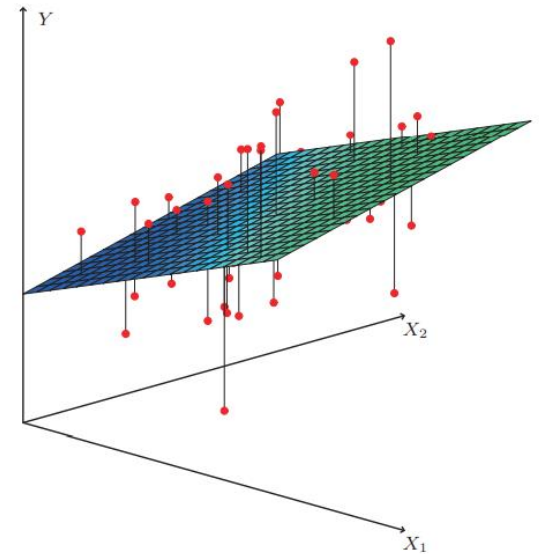
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

La suma de los errores el cuadrado se computa igual,

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

De la misma forma que en el caso univariado, es posible derivar para encontrar una estimación no sesgada de los parámetros del modelo.

Además,  $R^2 = \text{Cor}(Y, \hat{Y})^2$



# ¿Cómo acomodamos otros tipos de datos a modelos de regresión?

## Tipos de Datos

### Tipo de Dato

Numérico continuo



Numérico discreto



Categorico



Categorico ordenable

Fechas

Texto

Por ejemplo, puede interesarnos conocer la relación entre cantidad de unidades vendidas y:

Precio por unidad

Gastos de publicidad por medio

Precio por unidad (competencia)

Versión del producto (A vs. B)





$$x_i = \begin{cases} 1 & \text{el i-ésimo producto es tipo A} \\ 0 & \text{el i-ésimo producto es tipo B} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{el i-ésimo producto es tipo A} \\ \beta_0 + \epsilon_i & \text{el i-ésimo producto es tipo B} \end{cases}$$

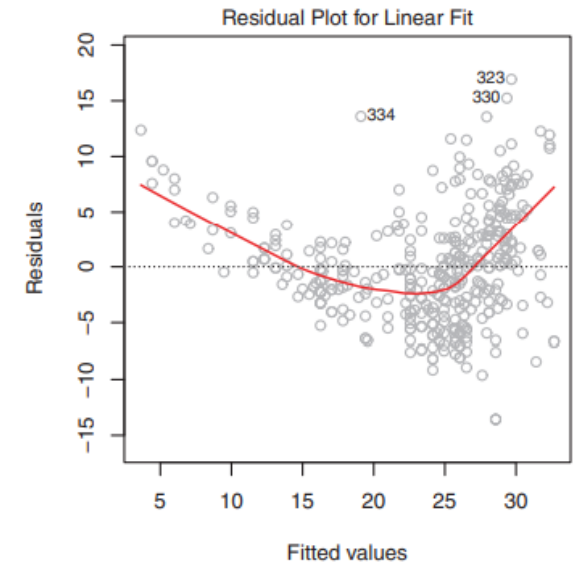
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{el i-ésimo es tipo A} \\ \beta_0 + \beta_2 + \epsilon_i & \text{el i-ésimo es tipo B} \\ \beta_0 + \epsilon_i & \text{el i-ésimo es tipo C} \end{cases}$$

# Hipótesis y limitaciones de la regresión lineal

La relación entre los datos es lineal.

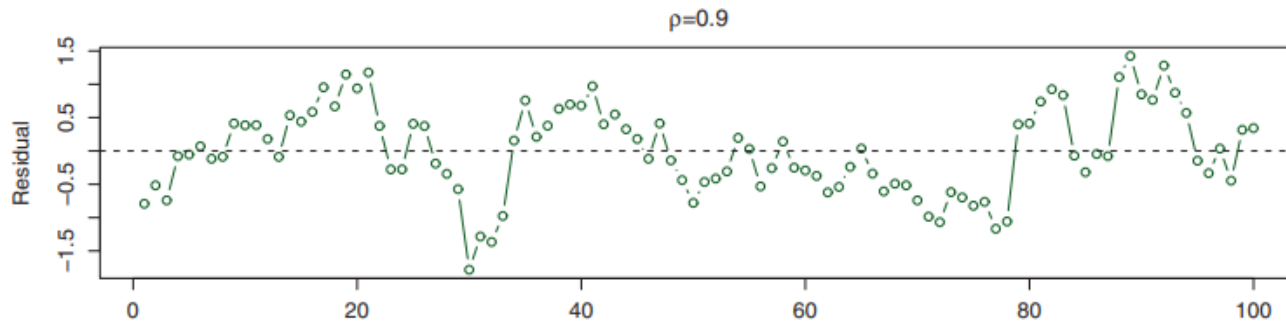
Si no es el caso, vemos que los residuos graficados contra la variable independiente resultan en una curva no-lineal.

Vamos a ver cómo abordar esto en la próxima clase.



Los errores no están correlacionados

Esta hipótesis suele no valer para series de tiempo con correlaciones temporales:



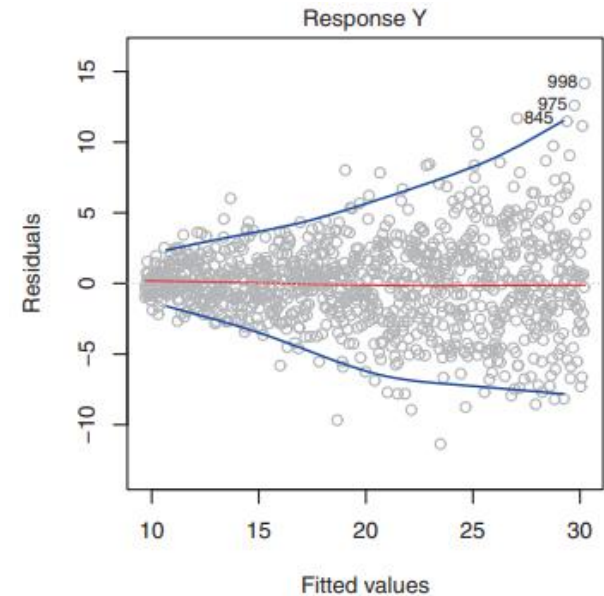
En ese caso, los intervalos de confianza exageran la probabilidad de que los parámetros se encuentren en el intervalo.

# Hipótesis y limitaciones de la regresión lineal

La varianza del término de error es constante

Esto puede ocurrir si el error depende de la magnitud de la variable. Típicamente el error depende logarítmicamente.

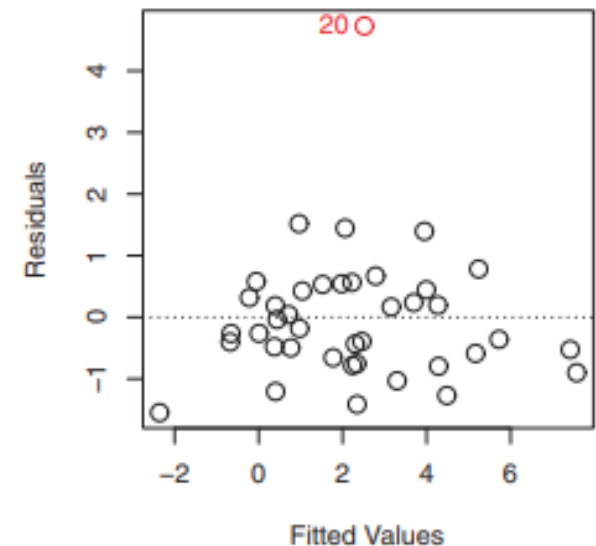
Puede resolverse usando cuadrados mínimos pesados.



## Outliers

Podemos detectarlos del gráfico de los residuos

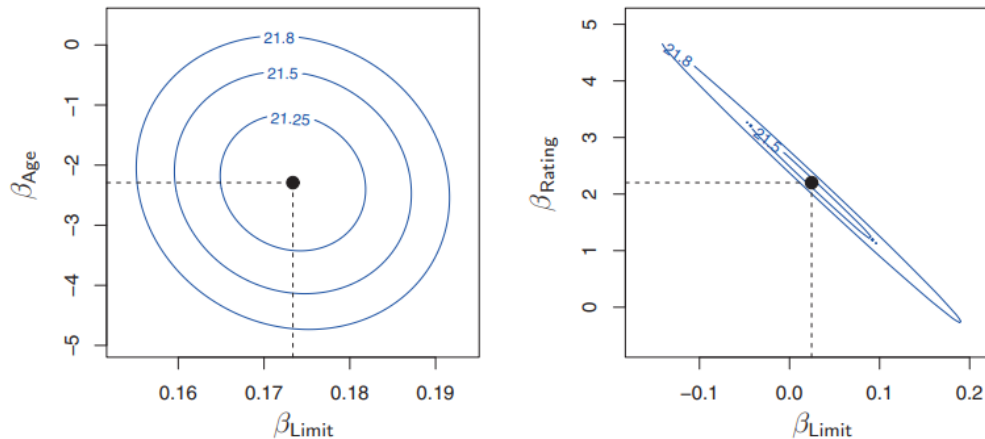
Necesitamos aplicar algún criterio para eliminarlos (por ejemplo, estar N desvíos estándar por arriba o abajo de la media)



# Hipótesis y limitaciones de la regresión lineal

Las variables independientes no son colineales

En caso de que las variables independientes muestren correlaciones altas, va a haber problemas a la hora de estimar los parámetros. En la práctica, con dos variables correlacionadas, el gráfico del error medio al cuadrado deja de tener un único mínimo y empieza a tener un mínimo en una región extensa del plano:



Más adelante en la materia vamos a ver formas de eliminar este problema, por ejemplo, utilizar análisis de componentes principales.

# Que no vimos:

1. Comparación de modelos. Una vez que estimamos los parámetros de un modelo, podemos construir una hipótesis nula sobre el valor de los parámetros y calcular la probabilidad de observar nuestros resultados asumiendo que la hipótesis nula es cierta (p-valor). Esto lo van a ver en una materia de estadística más avanzada.
2. Selección de variables. Si tenemos muchas variables independientes, ¿cuál es el subconjunto óptimo de variables para incluir en el modelo? Esto lo vamos a ver más adelante en la materia.
3. Entrenamiento y validación. Si nuestro interés está en la predicción de nuevos valores de la variable dependiente, ¿cómo podemos estimar el error de dicha predicción? Esto lo vamos a ver más adelante en la materia.
4. ¿Es posible modelar una relación no lineal entre los datos? Esto lo vamos a ver en la próxima clase.
5. ¿Y una relación no aditiva? No lo vemos con este modelo.

# Próxima clase:

Vamos a ver como superar algunas de las limitaciones de la regresión lineal multivariada, e introducir uno de los conceptos centrales de la materia: sobreajuste (*overfitting*)