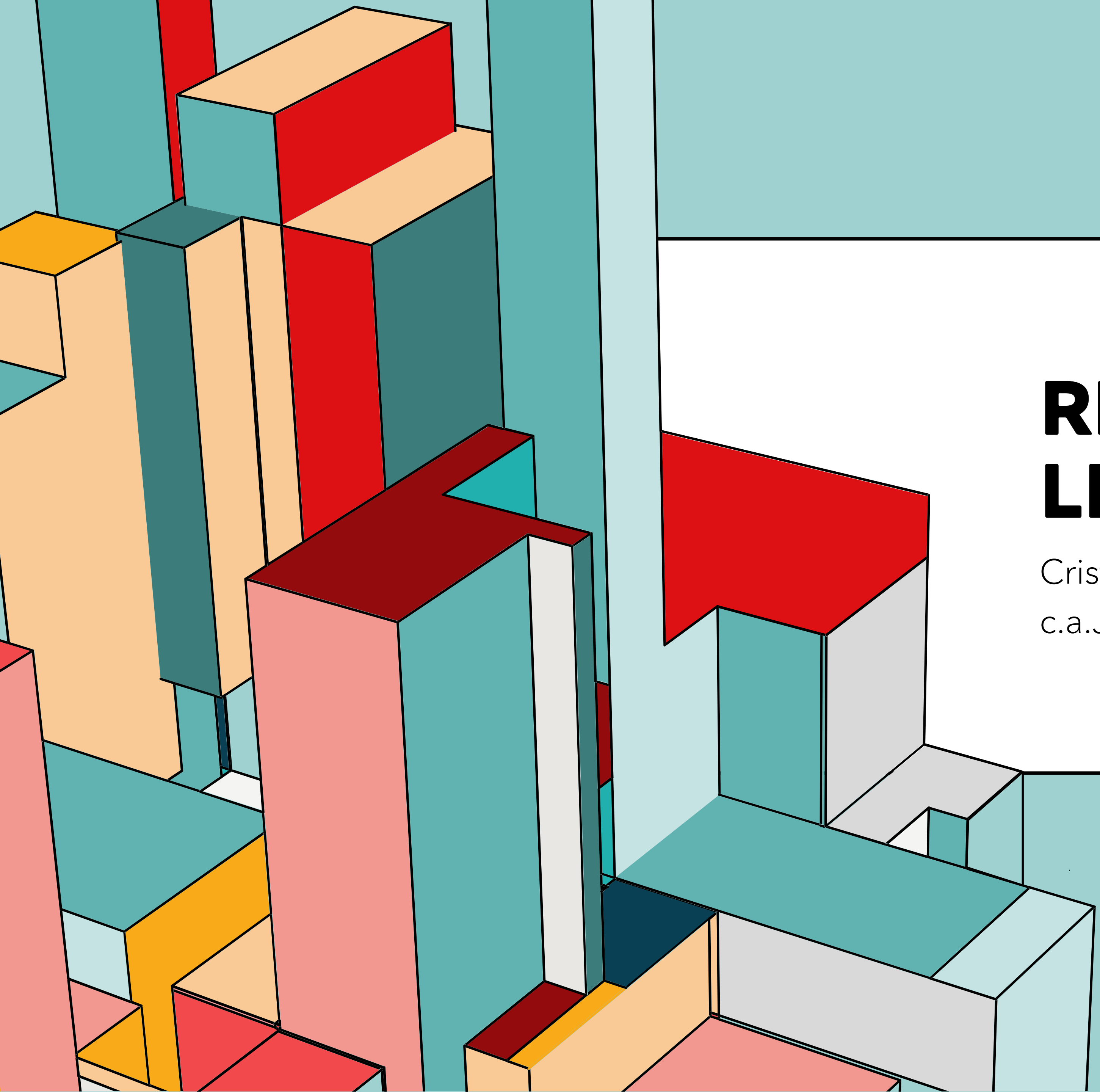


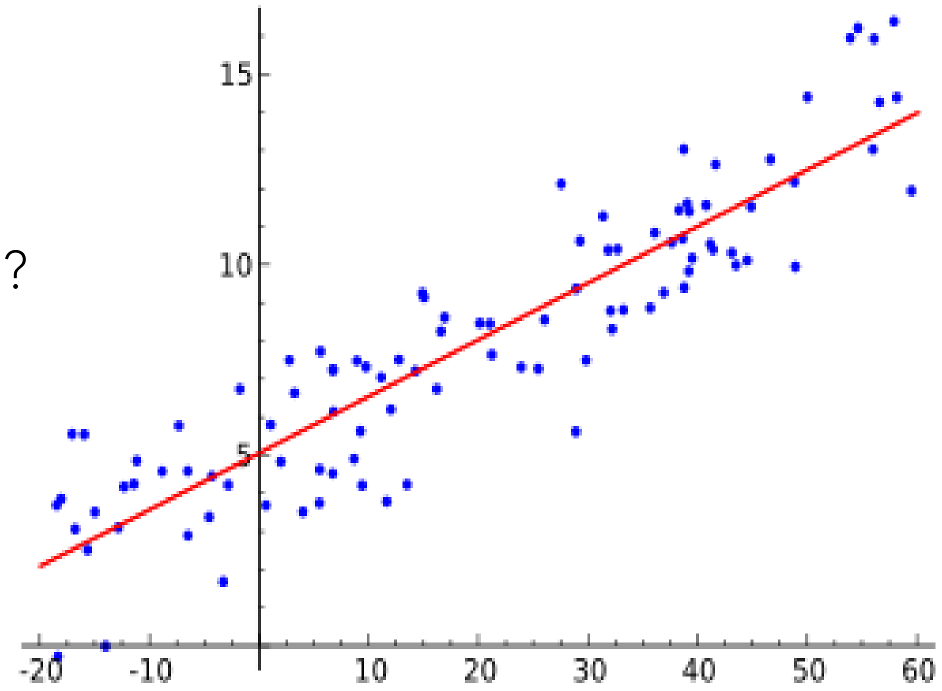
REGRESIÓN LINEAL

Cristhian Jaramillo
c.a.Jaramillo-huaman@lse.ac.uk

Objetivos de la sesión

Repasar <u>de manera sencilla</u>:

- ¿Qué son los modelos matemáticos?
- ¿Qué es la regresión lineal?
- ¿Cuál es el modelo matemático de la regresión lineal?
- ¿Cómo se leen los resultados del modelo?
- ¿Qué es el método de mínimos cuadrados?
- ¿Cuáles son los requisitos para la regresión lineal?
- ¿Qué es el contraste de regresión?



¿Por qué necesitamos conocer regresión?

La regresión permite responder preguntas como:

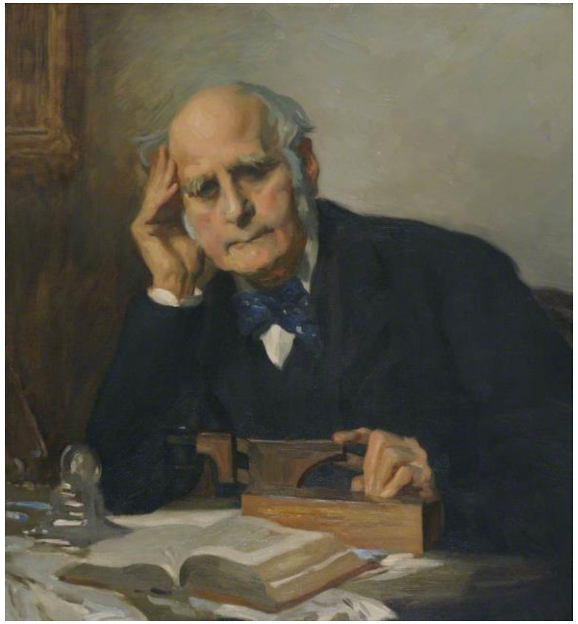
- ¿Cómo influyen los factores económicos y de salud en los casos de COVID?
- ¿El nivel de pobreza explica el voto por un partido político?
- ¿Podemos predecir resultados electorales usando encuestas y características demográficas?

Sirve tanto para **explicar relaciones** como para **hacer predicciones** fundamentadas en datos.

Breve historia

La idea de regresión surge en 1886 con **Francis Galton**, quien estudió la relación entre la altura de los padres y la altura de los hijos. Observó que los hijos de padres muy altos tendían a ser más bajos y los hijos de padres muy bajos tendían a ser más altos, fenómeno que llamó **"regression to the mean"**.

Posteriormente, **Karl Pearson** desarrolló las bases matemáticas de la correlación y la regresión, estableciendo los cimientos de la estadística moderna aplicada en ciencias sociales y economía.



¿Qué son los modelos matemáticos?

Es la función matemática que propone un tipo de relación entre una variable dependiente (Y) y una o más variables independientes:

- MODELO DETERMINÍSTICO: Supone que bajo condiciones ideales, el comportamiento de la variable dependiente puede ser totalmente descrito por una función matemática de variables independientes. PREDICE SIN ERROR. Ejemplo: Ley de la Gravedad.

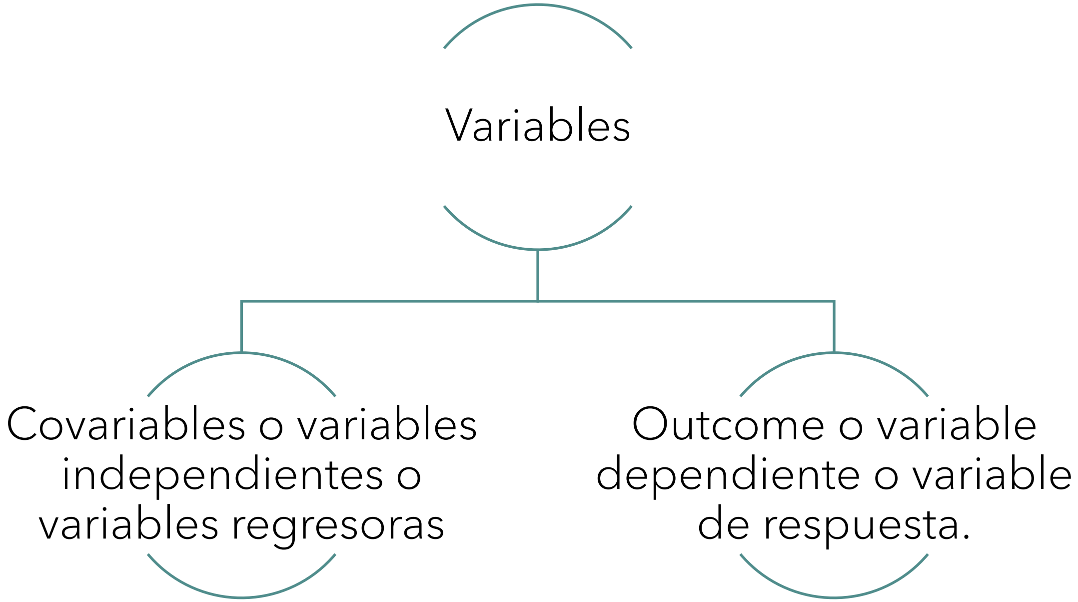
- MODELO ESTADÍSTICO: permite la incorporación de un componente aleatorio en la función. En consecuencia, las predicciones obtenidas tendrán asociado un ERROR DE PREDICCIÓN. Ejemplo: Relación de la altura con la edad en niños.

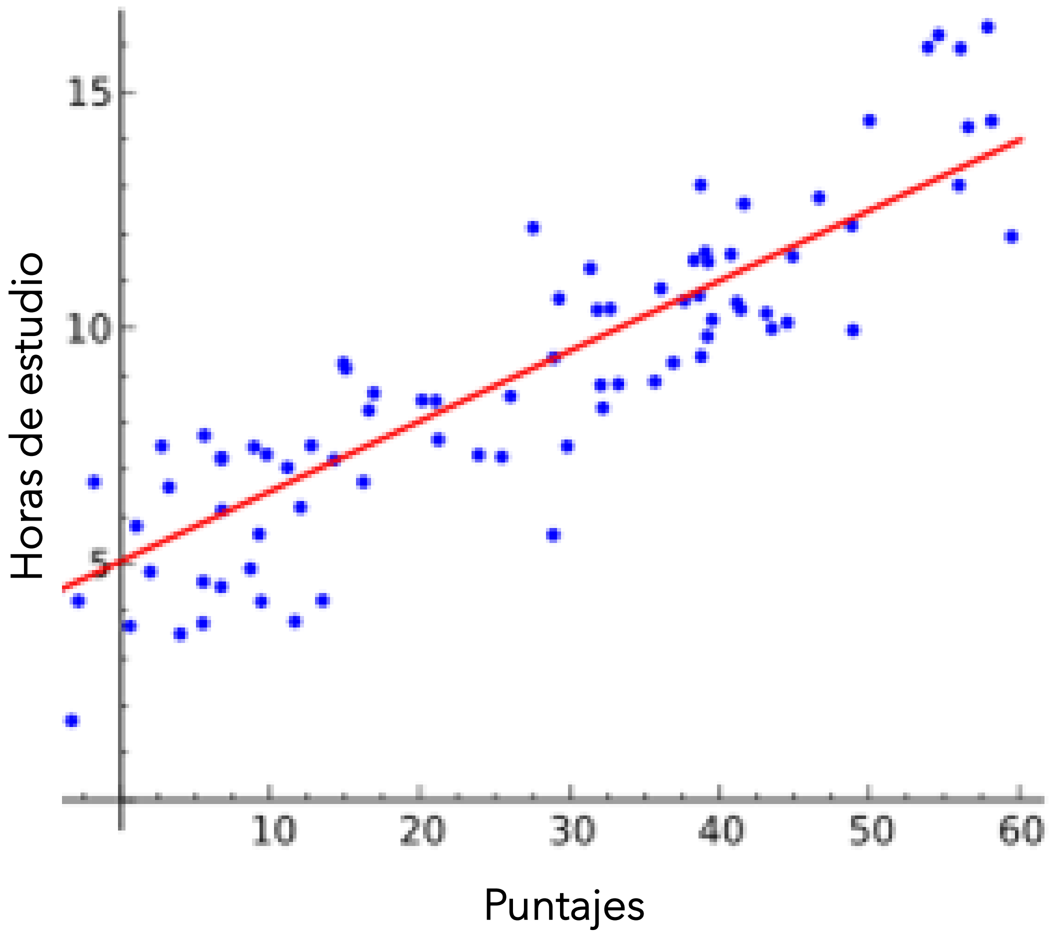
¿Qué es la regresión lineal?

Es un modelo estadístico que involucra el análisis de la relación entre dos variables para:

- Formalizar y entender relaciones teóricas entre variables
- Investigar si existe una asociación entre las dos variables.
- Estudiar la fuerza de la asociación (coeficiente de correlación).
- Estudiar la forma de la relación.

Se propone un modelo que mide el efecto de una variable independiente (X) en una variable dependiente (Y).





Si queremos saber si **a más horas de estudio se obtienen mejores notas**, podemos hacer un gráfico simple:

- En el eje X: horas de estudio
- En el eje Y: nota del examen

Si los puntos muestran una tendencia ascendente, una línea recta puede resumir esa relación y permitirnos predecir el desempeño de otros estudiantes.

Regresión lineal

$$\hat{y} = b_0 + b_1\ X$$

variable

variable

coeficiente
constante
no varía
un solo valor
INTERCEPTO

coeficiente
constante
no varía
un solo valor
PENDIENTE

Regresión lineal

En la vida real, muchos fenómenos tienen **más de un factor explicativo**.

Ejemplo: los casos de COVID pueden depender de la inversión en salud, el gasto en los hogares y el nivel de morbilidad.

La ecuación general es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_X X_X + \in$$

Estimación en R

Para calcular un modelo en R usamos la función lm():

```
modelo <- lm(casos_100k ~ var3 + var5 + var20, data = data)
summary(modelo)
```

Este comando nos entrega:

- Coeficientes estimados
- Errores estándar y p-valores
- Medida de ajuste R^2

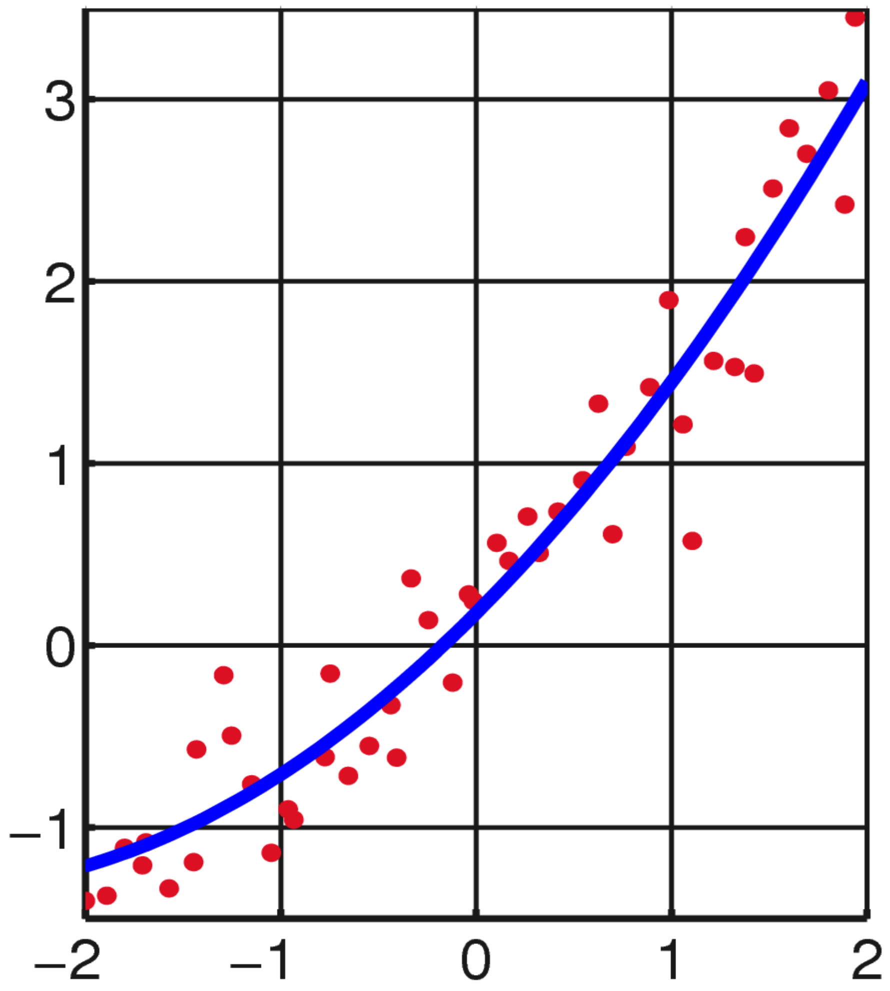
```
Call:
lm(formula = competitividad$casos_100k ~ competitividad$var3 + 
    competitividad$var5 + competitividad$var20)

Residuals:
    Min      1Q  Median      3Q     Max 
-800.31 -360.70    8.18  340.91 1331.92 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)   
(Intercept)          1.811e+03  1.206e+03   1.501  0.14986   
competitividad$var3  2.382e-02  8.178e-03   2.913  0.00892 **
competitividad$var5  1.484e+00  4.165e-01   3.563  0.00207 **
competitividad$var20 -3.678e+01  1.550e+01  -2.373  0.02833 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 548.8 on 19 degrees of freedom
Multiple R-squared:  0.7314,	Adjusted R-squared:  0.689 
F-statistic: 17.25 on 3 and 19 DF,  p-value: 1.185e-05
```

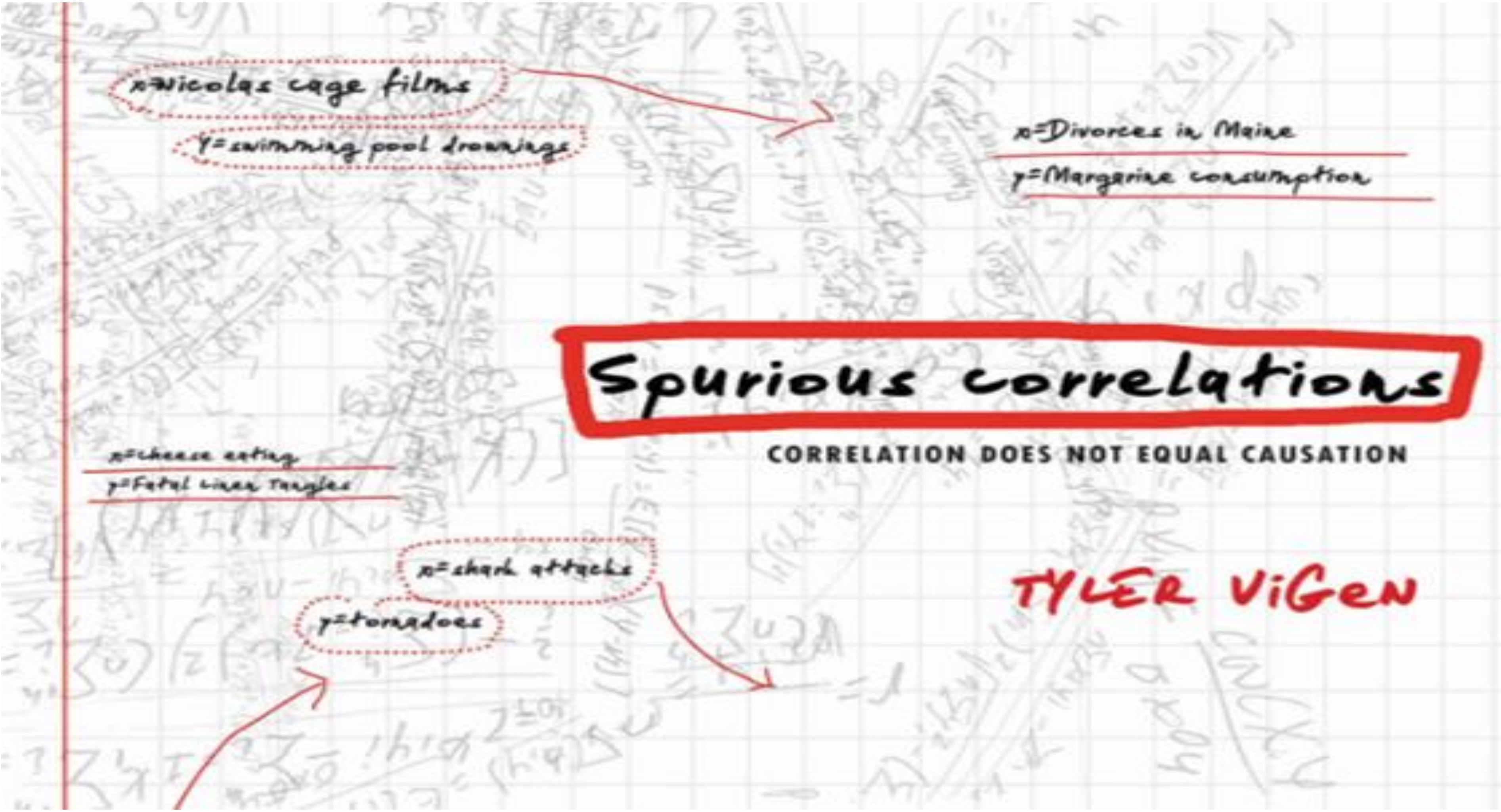
¿Qué es el Método de Mínimos cuadrados?

Es aquella recta en la cual la ecuación que predice los cambios es la "mejor" línea en cuanto a la reducción de las distancias entre los valores observados y los valores que se predicen.

Si la línea está cerca de las observaciones, los residuales tienden a ser pequeños.

El R cuadrado: Cuanta variación podemos explicar en la variable dependiente a partir de la(s) explicativa(s). Cuanto más se acerca R2 a 1, más fuerte es la asociación lineal y más efectiva es la línea recta $y = \alpha + bx$ para predecir la variable dependiente o de respuesta




x=Nicolas cage films
y=swimming pool drownings
x=Divorces in Maine
y=Margarine consumption
Spurious correlations
CORRELATION DOES NOT EQUAL CAUSATION
x=cheese eating
y=Fatal linen tangles
x=shark attacks
y=tornadoes
TYLER VIGEN

Hipótesis en regresión

Cada coeficiente se contrasta con la hipótesis:

- H0: el coeficiente es cero (la variable no tiene efecto).
- HA: el coeficiente es distinto de cero (la variable sí influye).

Si el p-valor es menor a 0.05, se considera estadísticamente significativo.

```
Call:
lm(formula = competitividad$casos_100k ~ competitividad$var3 +
    competitividad$var5 + competitividad$var20)

Residuals:
    Min      1Q  Median      3Q     Max
-800.31 -360.70    8.18  340.91 1331.92

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.811e+03  1.206e+03   1.501  0.14986
competitividad$var3  2.382e-02  8.178e-03   2.913  0.00892 **
competitividad$var5  1.484e+00  4.165e-01   3.563  0.00207 **
competitividad$var20 -3.678e+01  1.550e+01  -2.373  0.02833 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 548.8 on 19 degrees of freedom
Multiple R-squared:  0.7314,    Adjusted R-squared:  0.689
F-statistic: 17.25 on 3 and 19 DF,  p-value: 1.185e-05
```

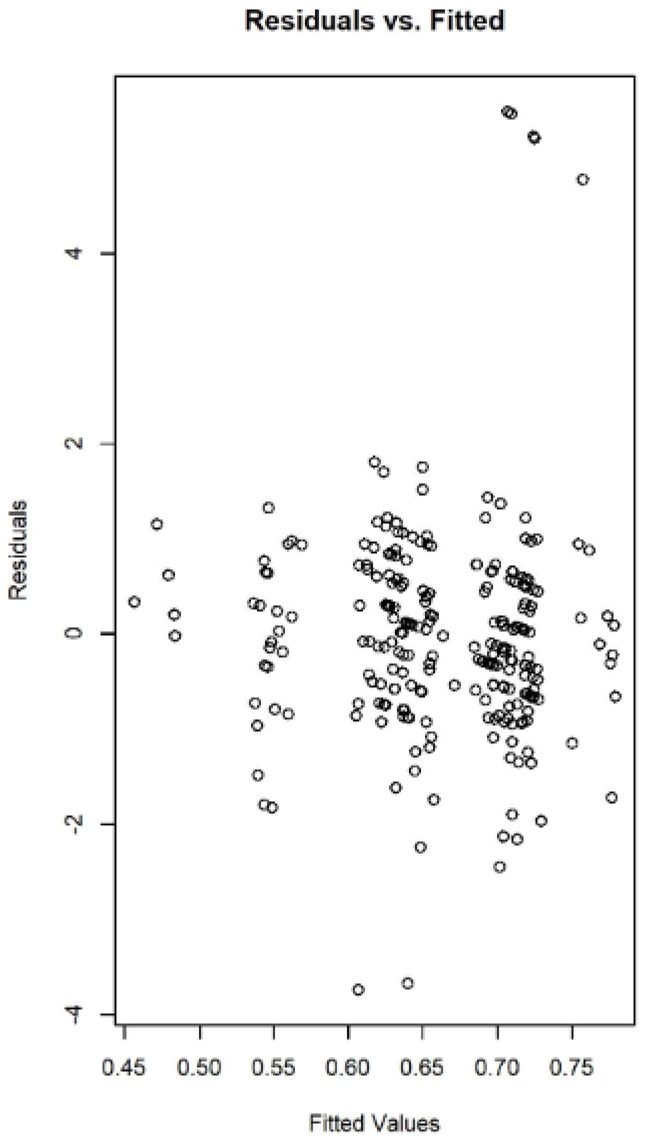
Supuestos del modelo

- Relación lineal entre predictores y respuesta
- Los residuos tienen distribución normal
- Homocedasticidad: varianza constante de los residuos
- Independencia de los residuos
- Ausencia de multicolinealidad entre variables explicativas

Supuesto de linealidad

La relación entre las variables explicativas y la respuesta debe ser **lineal**.

Se verifica graficando los residuos contra los valores ajustados: si los puntos se dispersan alrededor de cero sin un patrón claro, el supuesto se cumple.

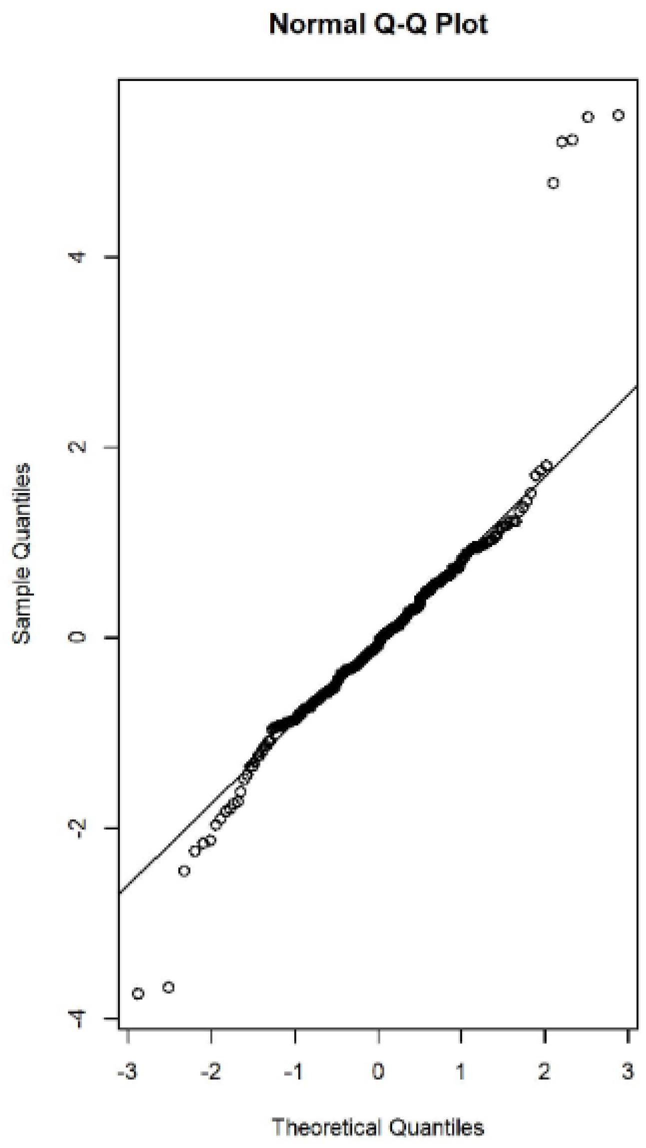


Supuesto de normalidad

Los errores del modelo deben seguir una **distribución normal**.

Se revisa con un gráfico QQ-Plot: los puntos deberían alinearse sobre la diagonal.

También puede comprobarse con la prueba de Shapiro-Wilk.



Supuesto de homocedasticidad

La **varianza de los residuos debe ser constante**.

Si la dispersión de los residuos aumenta o disminuye a medida que cambian los valores ajustados, hay heterocedasticidad.

El test Breusch-Pagan permite comprobarlo formalmente.

$$\sigma^2 = {\sum \mu_i^2}/{n}$$

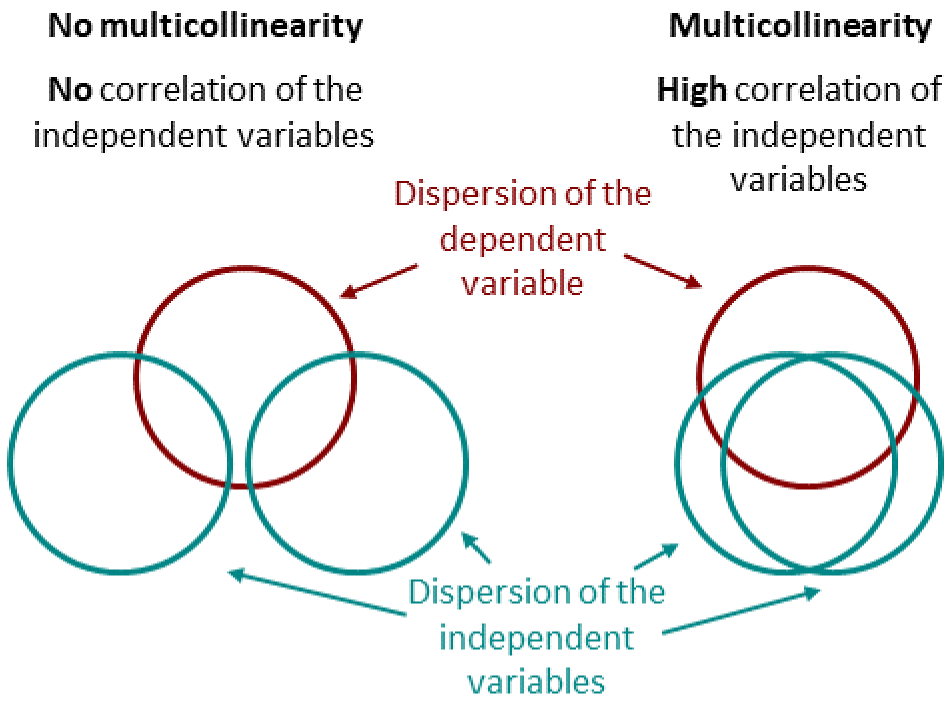
$where,$ $\quad n \text{ is the number of observations}$

$\sum \mu_i^2 \text{ is the sum of squared residuals}$

Supuesto de multicolinealidad

Cuando dos o más variables explicativas están fuertemente correlacionadas, los coeficientes pueden volverse inestables.

Se detecta con el **Factor de Inflación de Varianza (VIF)**: valores mayores a 5 o 10 indican problema.



Supuesto de independencia

Los residuos deben ser independientes entre sí.

Si hay patrón en el tiempo o en el orden de los datos, se viola el supuesto.

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

Se usa el **test de Durbin-Watson** para evaluar autocorrelación.

¿Qué pasa si fallan los supuestos?

Si los supuestos no se cumplen:

- Transformar variables (logaritmos, potencias).
- Usar estimadores robustos frente a heterocedasticidad.
- Revisar colinealidad y eliminar variables redundantes.
- Considerar otros modelos como regresión logística o modelos no lineales.

Tipo de regression	Cuándo se usa	Ejemplo
Lineal simple	Una X y Y continua	Horas de estudio → nota
Lineal múltiple	Varias X y Y continua	Factores socioeconómicos → COVID
Logística	Y binaria (0/1)	Votar o no votar
Poisson / Neg. Binomial	Datos de conteo	Número de protestas por año
Cox (supervivencia)	Tiempo hasta evento	Duración de gobiernos
Multinivel (mixto)	Datos jerárquicos	Grado educativo

GRACIAS

Cristhian Jaramillo

https://cristhianjaramillo.com/