



UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Diseño y Análisis de Experimentos en Ingeniería y Ciencias Ambientales

Regresión Lineal Simple

Dr. Christian R. Encina Zelada

cencina@lamolina.edu.pe

Objetivos del capítulo

Dar a conocer los fundamentos y aplicaciones de análisis de la Regresión Lineal Simple (RLS), así como su utilidad para estudiar la dependencia de una variable (variable dependiente) respecto a otra variable (variable independiente). Se incluye la aplicación del coeficiente de correlación para medir el grado de asociación de dos variables aleatorias

Antecedentes

En 1889 en su libro "Herencia Natural", Francis Galton se refirió a la "ley de la regresión universal". Él dijo que "cada peculiaridad en un hombre es compartida por sus parientes, pero en promedio, en un grado menor".

En 1903, Karl Pearson, amigo de Galton, colectó más de 1000 registros de tallas de padres e hijos y con esta información estimó la siguiente línea para explicar la talla del hijo en función a la del padre (en pulgadas):

$$\text{Talla del hijo} = 33.73 + 0.516 \text{ talla del padre}$$

Análisis de Regresión Lineal Simple

El análisis de regresión lineal simple trata el problema de predecir o estimar una variable, llamada respuesta o variable dependiente, a partir de otra variable llamada predictora, explicativa o variable independiente.

Análisis de Regresión Lineal. Ejemplo 1

- Conforme los quesos maduran, ocurren varios procesos químicos que determinan el sabor del producto final. Es un estudio en queso cheddar, 10 muestras de queso fueron analizadas en su composición química. Además, una medida subjetiva del sabor fue obtenida combinando los puntajes asignados por varios sujetos que probaron el queso. Los datos se dan a continuación:

Análisis de Regresión Lineal. Ejemplo 1

Muestra	1	2	3	4	5	6	7	8	9	10
Sabor	12.3	47.9	37.3	21	0.7	40.9	18	15.2	16.8	0.7
AA	4.543	5.759	5.892	5.242	4.477	6.365	5.247	5.298	5.366	5.328
H ₂ S	3.135	7.496	8.726	4.174	2.996	9.588	6.174	5.22	3.664	3.912
AL	0.86	1.81	1.29	1.58	1.06	1.74	1.63	1.33	1.31	1.25

Las variables son:

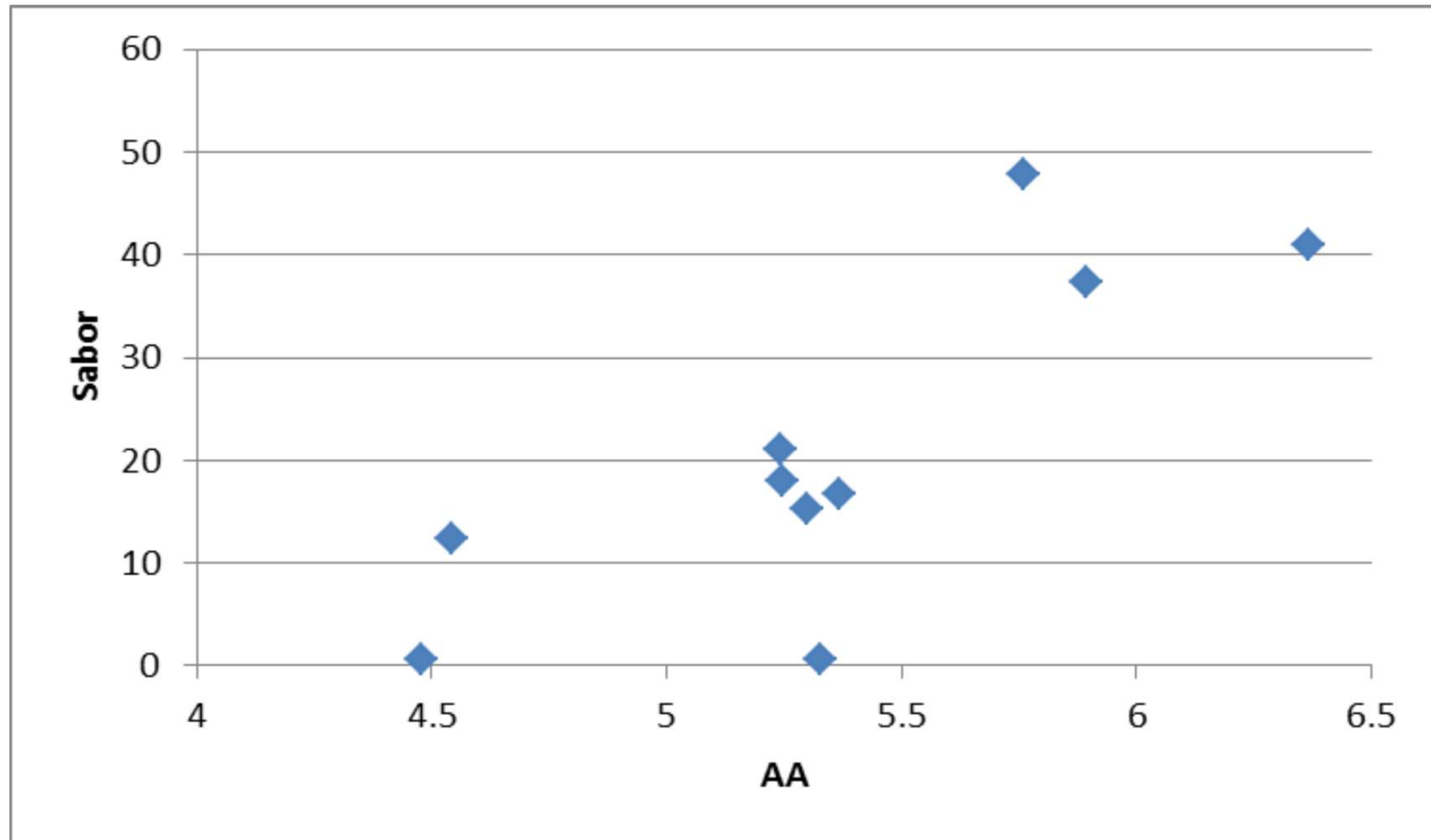
Sabor : puntaje de sabor subjetivo, obtenido combinando los puntajes de varios sujetos.

AA : logaritmo natural de la concentración de ácido acético.

H₂S : logaritmo natural de la concentración de sulfuro de hidrógeno.

AL : concentración de ácido láctico.

Análisis de Regresión Lineal. Ejemplo 1



Modelo estadístico

El modelo poblacional de regresión lineal simple es el siguiente:

$$Y_i = a + \beta X_i + \varepsilon_i$$

Donde a es el estimador de a y b el estimador β .

Estimación del modelo

Los parámetros del modelo son estimados por el método de Mínimos Cuadrados. Este método permite obtener los valores estimados α y β .

La aplicación de este método da los siguientes resultados para la estimación de los parámetros:

$$\hat{\beta} = b = \frac{SP(XY)}{SP(X)} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

$$\hat{\alpha} = a = \bar{Y} - b\bar{X}$$

Análisis de Regresión Lineal. Ejemplo 1 (continuación)

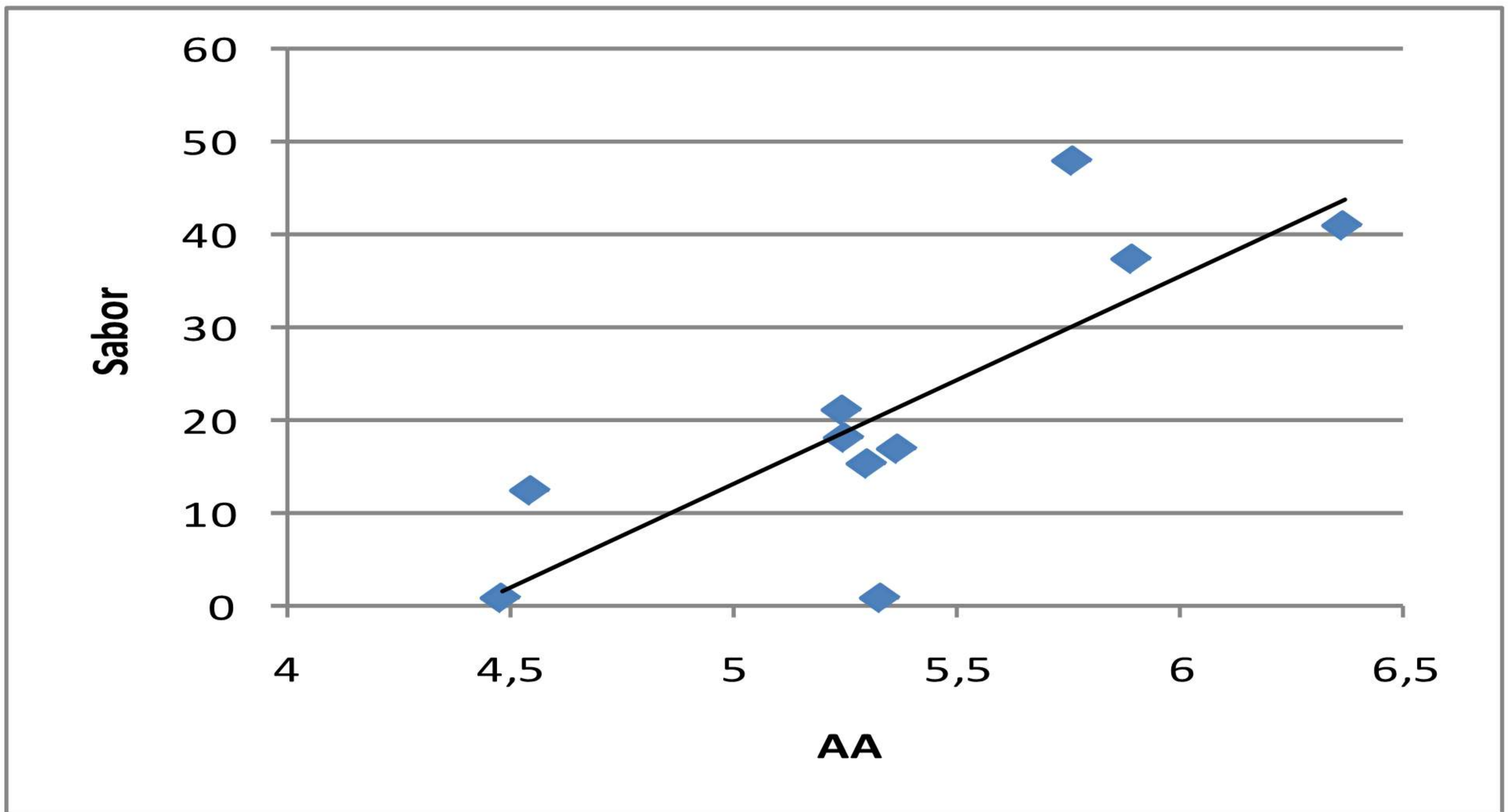
Se va estimar el modelo de regresión que considera a la variable AA como variable predictora. Quedan como ejercicios los análisis de los casos de las variables H₂S y AL.

$$\bar{Y} = 21.08 \quad \bar{X} = 5.3517 \quad \sum X_i^2 = 289.34 \quad \sum Y_i^2 = 6789.06 \quad \sum X_i Y_i = 1193.91$$

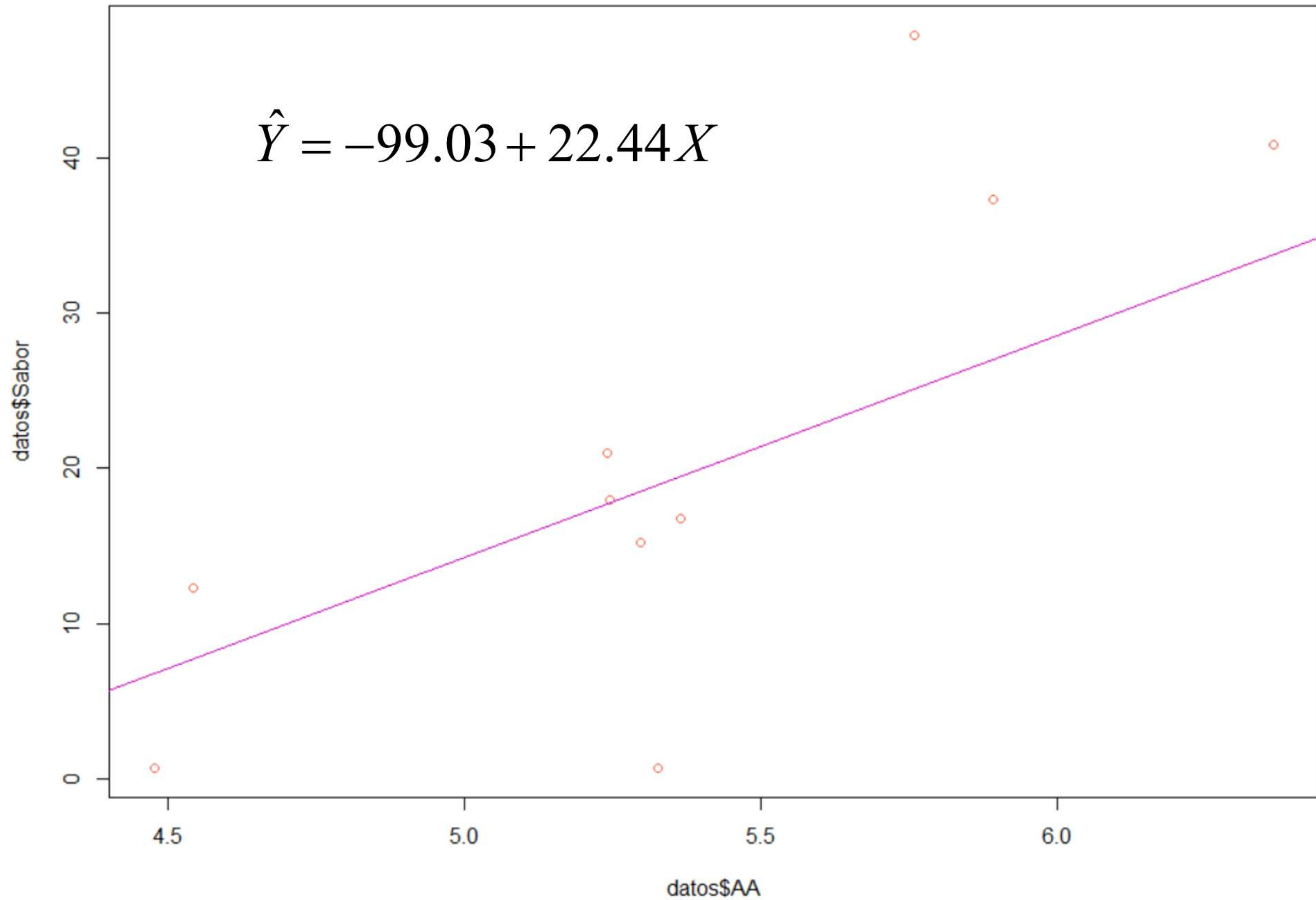
$$b = \frac{1193.91 - 10 * (21.08) * (5.3517)}{289.34 - 10 * (5.3517)^2} = 22.44 \quad a = 21.08 - 22.44 * (5.3517) = -99.03$$

$$\hat{Y} = -99.03 + 22.44X$$

Análisis de Regresión Lineal. Ejemplo 1



Grafica de Dispersion



Análisis de Regresión Lineal. Ejemplo 1

Análisis de Variancia

Hipótesis:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

Cuadro de Análisis de Variancia (cuadro ANVA):

Fuentes de variación	Gl	SC	CM	Fc
Regresión	1	b SP(XY)	SC(Reg)/gl(Reg)	CM(Reg)/CM(Error)
Error	n - 2	SC(Y) - b SP(XY)	SC(Error)/gl(Error)	
Total	n - 1	SC(Y)		

Análisis de Regresión Lineal. Prueba Hipótesis

Para el caso de las variables $Y = \text{sabor}$ y $X = \text{AA}$, se tiene lo siguiente:

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

o dicho literalmente:

H_0 : El sabor del queso no depende de la concentración de ácido acético.

H_1 : El sabor del queso si depende de la concentración de ácido acético.

Fuentes de variación	gl	SC	CM	Fc
Regresión	1	1476	1476	13.58
Error	8	869	109	
Total	9	2345		

Análisis de Regresión Lineal. Conclusión de la prueba Hipótesis

Conclusión

El valor de tabla para un nivel de significación del 5% es $F_{(0.95,1,8)} = 5.318$.

Como el valor calculado es mayor al valor de tabla se rechaza H_0 .

En conclusión, existe suficiente evidencia estadística para aceptar que el sabor del queso depende de la concentración de ácido acético a través de un modelo lineal.

P from F

F

13.58

DF_n

1

DF_d

8

Compute P


P Value Results

F=13.58 DF_n=1 DF_d=8

The P value equals 0.0062

- $p \text{ value} = 0.006171$
(RStudio)
- $p \text{ value} < \text{"alpha"} (0.05)$
- Se rechaza " H_0 "
- El sabor del queso si
depende de la concentración
de ácido acético.

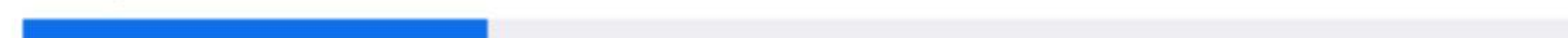
Votación sin título

 2:52 | 1 pregunta | 10 de 10 (100%) participaron

1. qué busco para poder realizar una regresión simple (Opción única) *

10/10 (100%) han respondido

H0: $\beta = 0$ (3/10) 30%

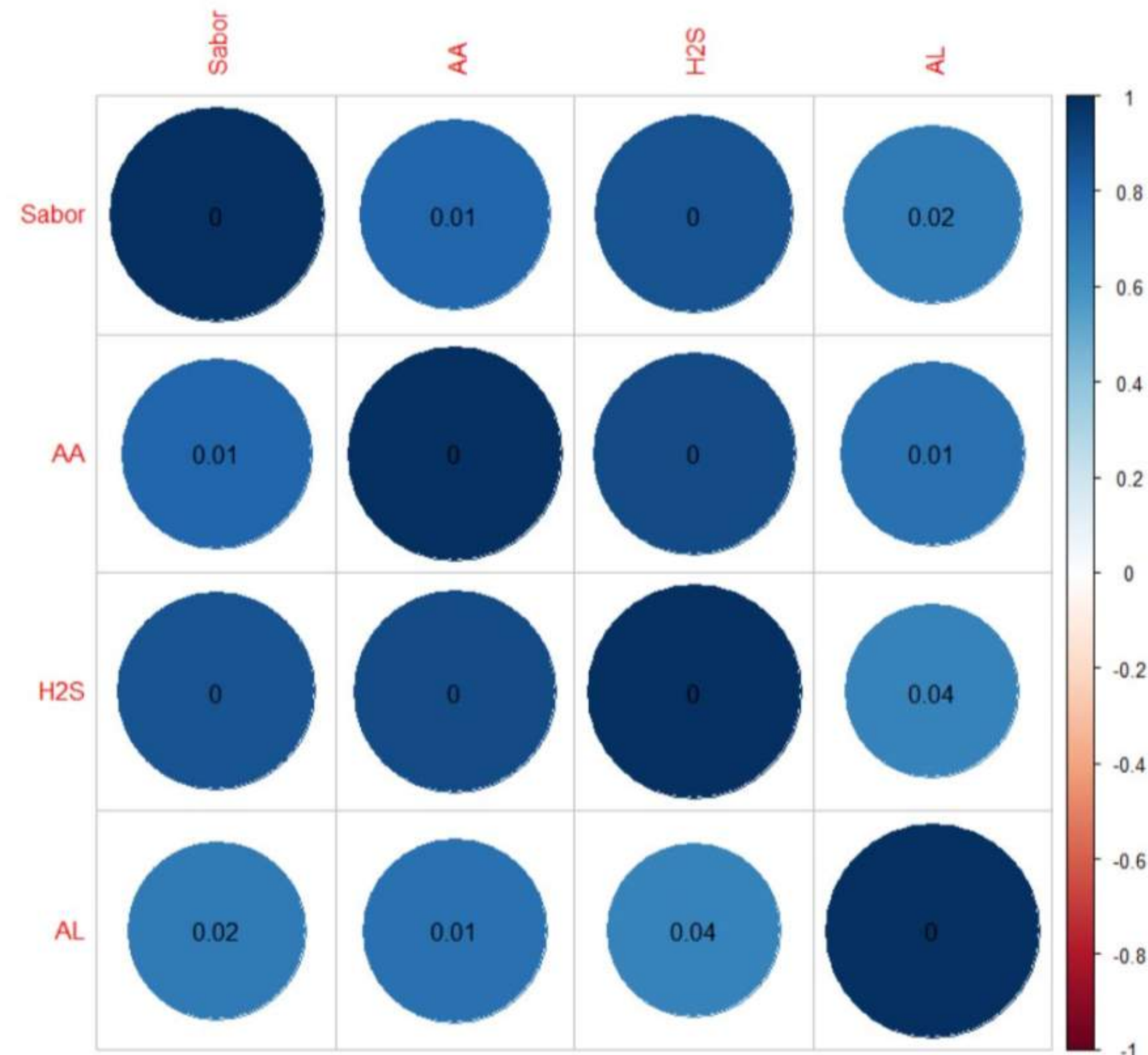


H1: $\beta \neq 0$ (7/10) 70%



Finalizar votación

	Sabor	AA	H2S	AL
Sabor	0.0000	0.006170	0.001337	0.02407
AA	0.00617	0.00000	0.000549	0.01440
H2S	0.00133	0.000549	0.00000	0.03525
AL	0.02407	0.014405	0.035259	0.0000



Análisis de Regresión Lineal. Coeficiente de determinación

El coeficiente de determinación mide el porcentaje de la variabilidad de la respuesta que es explicado por la variable predictora. Su valor va de 0 a 1 y se calcula mediante la siguiente expresión:

$$r^2 = \frac{SC(Reg)}{SC(Total)}$$

$$r^2 = \frac{1476}{2345} = 0.63$$

El 63% de la variabilidad del sabor es explicado por la concentración de ácido acético.

Análisis de Regresión Lineal. Coeficiente de correlación

El coeficiente de correlación es una medida de la asociación existente entre dos variables cuantitativas. Este coeficiente toma valores desde -1 hasta 1. Para interpretar un coeficiente de correlación tenga en cuenta lo siguiente:


El coeficiente de correlación es la raíz cuadrada del coeficiente de determinación con el signo de b (pendiente estimada).

$$r = \sqrt{0.63} = 0.79$$

$r = 0.79$ indica una elevada correlación positiva.



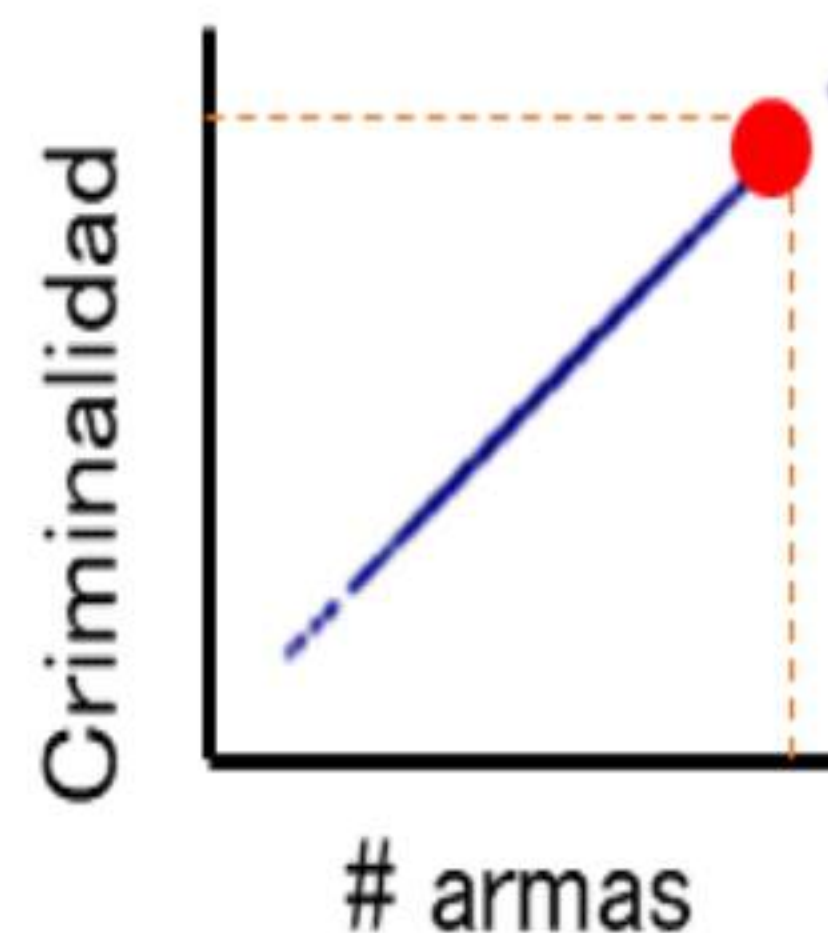
	Sabor	AA	H2S	AL
Sabor	1.00000000	0.7933025	0.8620447	0.7004457
AA	0.7933025	1.00000000	0.8905297	0.7400013
H2S	0.8620447	0.8905297	1.00000000	0.6666809
AL	0.7004457	0.7400013	0.6666809	1.00000000

- 
- Absolute R values ranged from:
 - 0.91 to 1.00 for very high correlation;
 - from 0.71 to 0.90 for high correlation,
 - from 0.51 to 0.70 for moderate correlation,
 - from 0.31 to 0.50 for low correlation, and
 - from 0.00 to 0.30 for absence of correlation (Fang et al., 2019).

3 claves

para interpretar una correlación

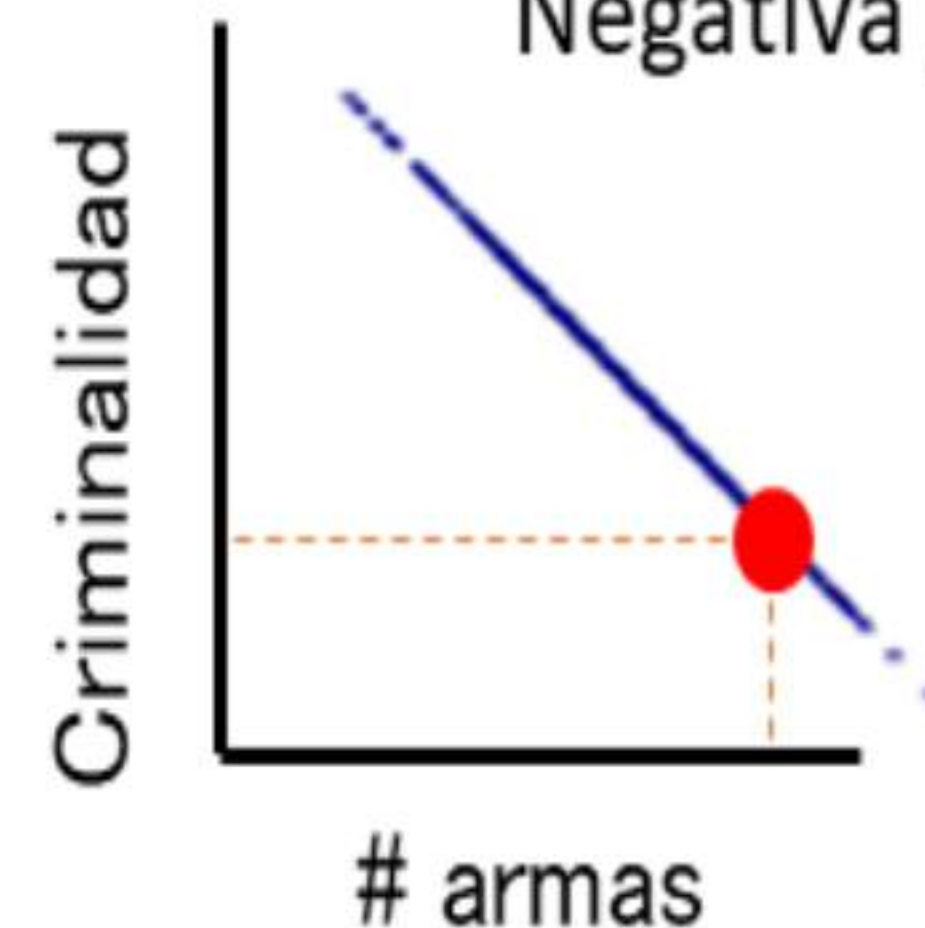
Positiva [+ 1]



1

Dirección

Negativa [-- 1]



Hay personas que piensan que mientras más armas exista en una ciudad hay más criminalidad [**Riesgo**]

Otras personas piensan que mientras más armas exista menos criminalidad puede haber en la ciudad [**Protección**]

3 Significacia

Menor a 0.05



Existe una relación

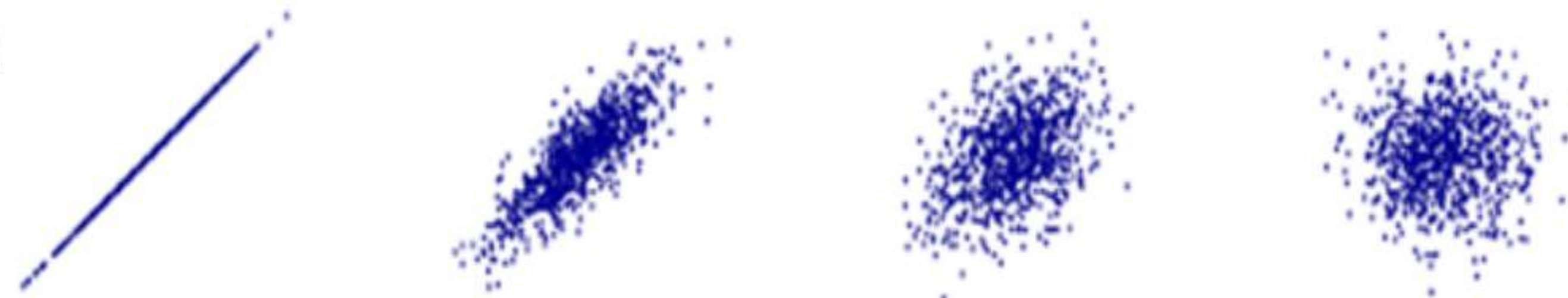
Mayor a 0.05



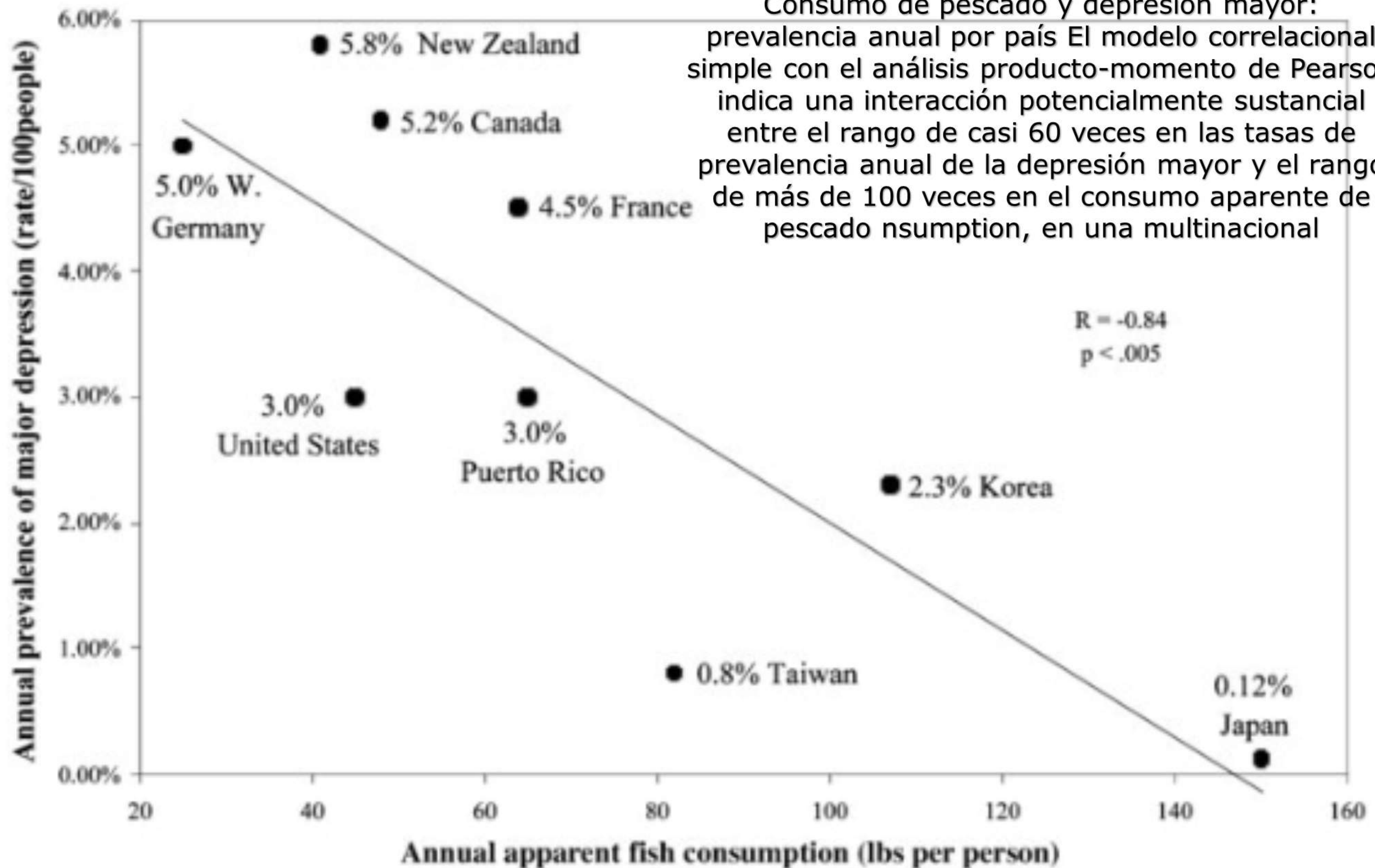
Son independientes/
No hay relación

Fuerza 2

Mientras más cerca a 1 indicaría que hay más posibilidad de que las armas influyan en la criminalidad



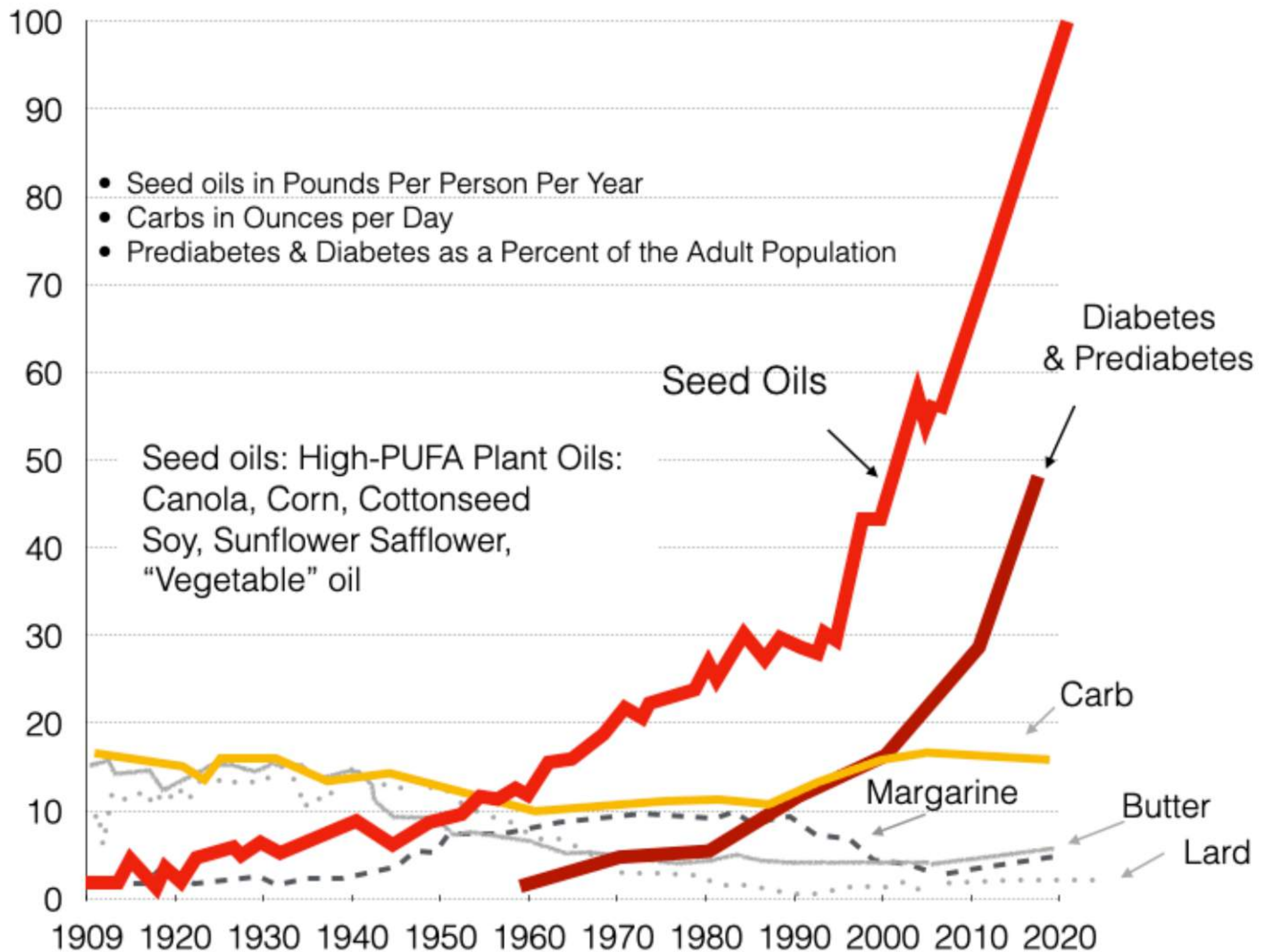
Consumo de pescado y depresión mayor:
prevalencia anual por país El modelo correlacional simple con el análisis producto-momento de Pearson indica una interacción potencialmente sustancial entre el rango de casi 60 veces en las tasas de prevalencia anual de la depresión mayor y el rango de más de 100 veces en el consumo aparente de pescado nsumption, en una multinacional



The Excessive Seed Oil Hypothesis: The More Seed Oil We Eat, the More Disease We Get

(It's Not Carb & Sugar or Animal Fat)

© DrCate.com



- T2D first observed in 1938.
- Statistics not tracked prior to 1958.
- T2D is the granddaddy of all metabolic disease

Fig. 1
IS FACEBOOK DRIVING
THE GREEK DEBT CRISIS?

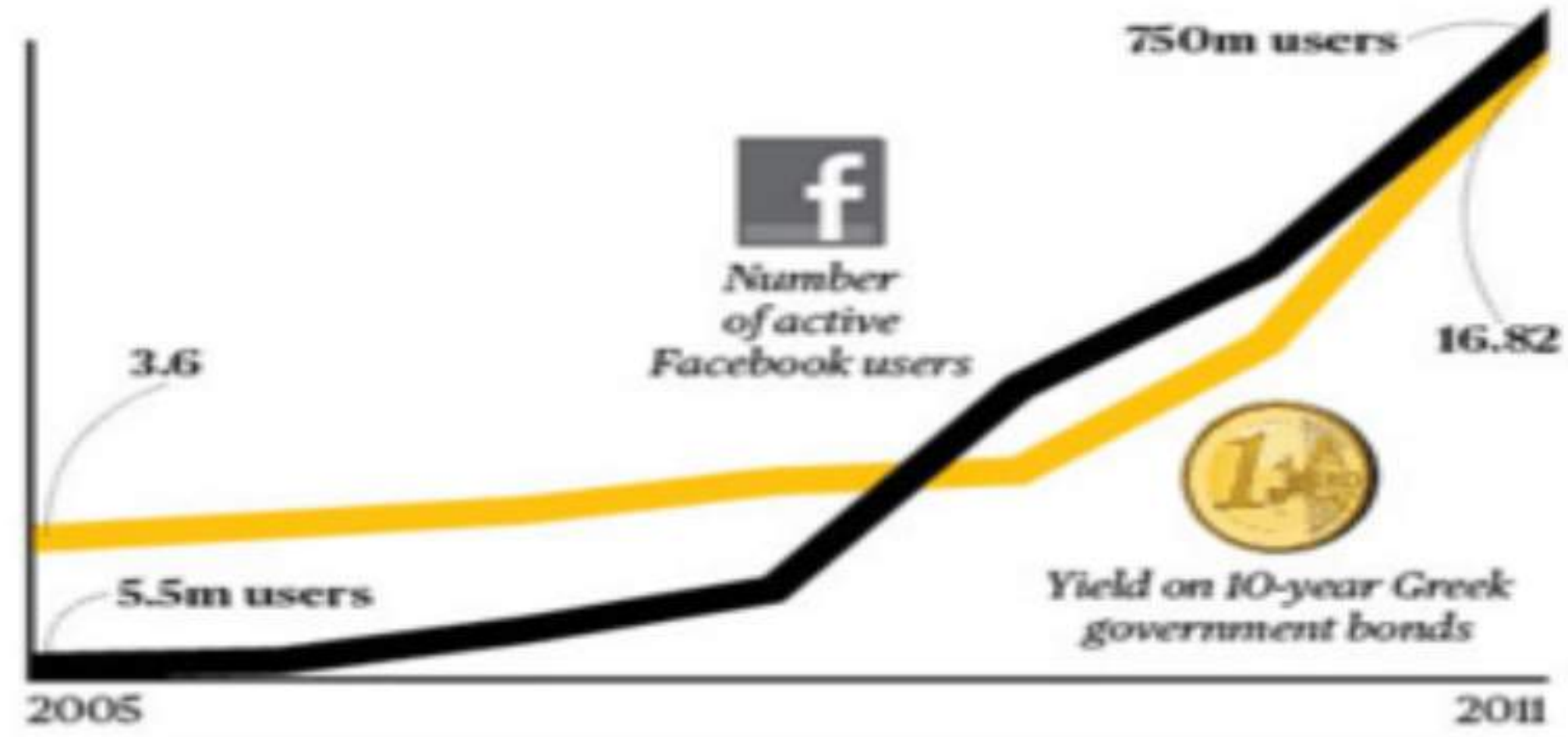


Fig. 2
IS GLOBAL WARMING A HOAX
PROPAGATED BY SCIENTISTS?

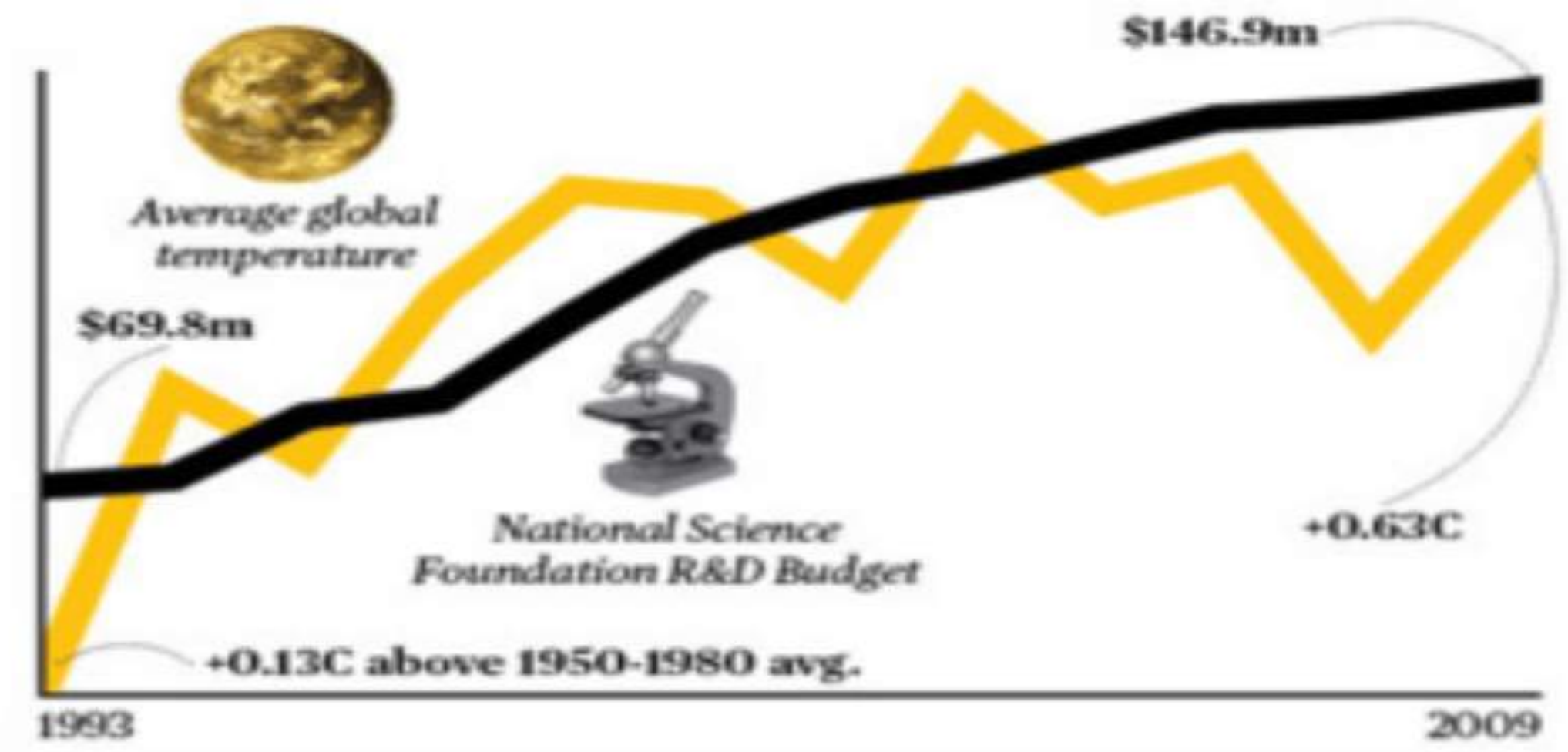


Fig. 3
DID AVAS CAUSE
THE U.S. HOUSING BUBBLE?



Fig. 4
WOULD M. NIGHT SHYAMALAN START MAKING GOOD MOVIES
AGAIN IF PEOPLE BOUGHT MORE NEWSPAPERS?

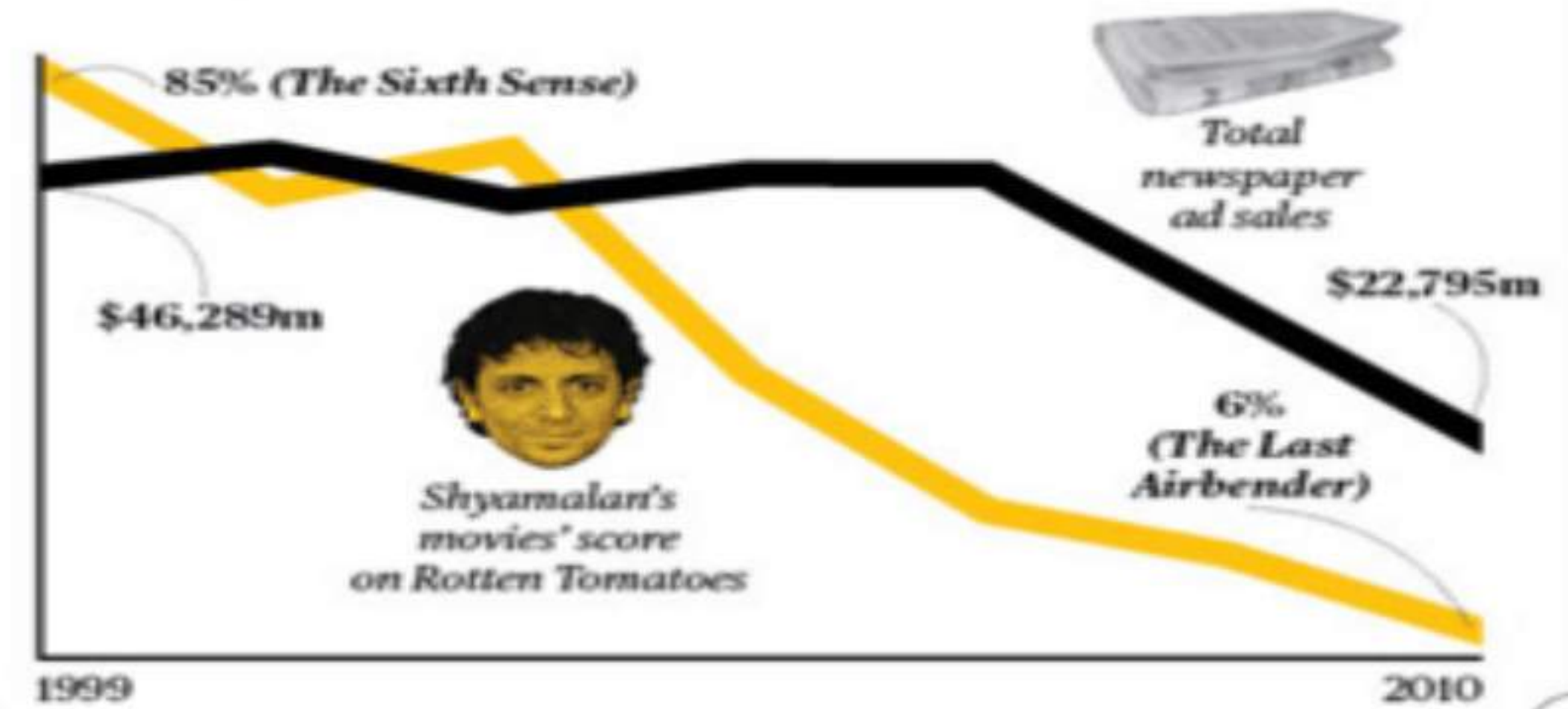


Fig. 5
DID WE TV SABOTAGE MICHELE BACHMANN'S CANDIDACY
BY TAKING STATEN ISLAND CAKES OFF THE AIR?

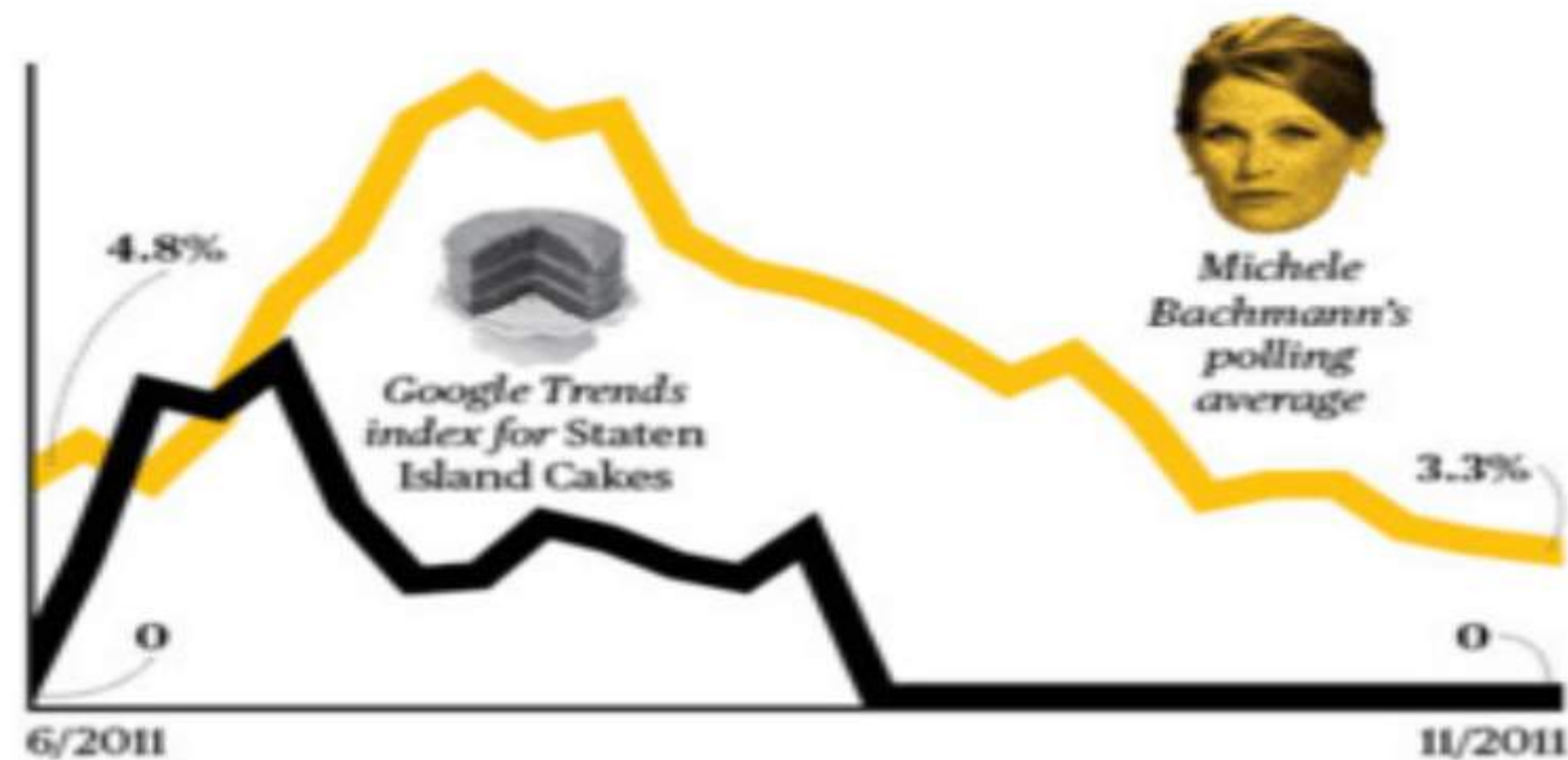


Fig. 6
IS THIS MOUNTAIN RANGE AFFECTING
THE MURDER RATE?

