DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de datos

Regresión Lineal

Segundo Cuatrimestre 2025

Mateo Guerrero Schmidt - Pablo Turjanski - Manuela Cerdeiro

Contenido de la clase

- Modelos de regresión
- Modelo de regresión lineal simple
- Regresión múltiple y polinomial
- Ejemplos y precauciones

Regresión

Objetivo:

Estimar la función que determina la relación $X \rightarrow Y$, donde Y es una variable contínua.

Ejemplo: peso - altura, metros cubiertos - precio.

Podemos querer hacerlo para entender, explicar o predecir.

Clasificación vs. Regresión

En clasificación buscamos explicar una variables que es **categórica**.

- True - False
- Setosa - Versicolor - Virginica
- Sobrevivió - No sobrevivió
- Ceibo - Pindó - Eucaliptus - Jacarandá

En regresión buscamos explicar una variable que es **continua** (puede tomar valores en R o en Z)

- Altura de una persona
- Precio de una propiedad
- Temperatura

Ejemplo - Uso de pesticidas

El daño al material genético por parte de los pesticidas es un tema de preocupación a escala global

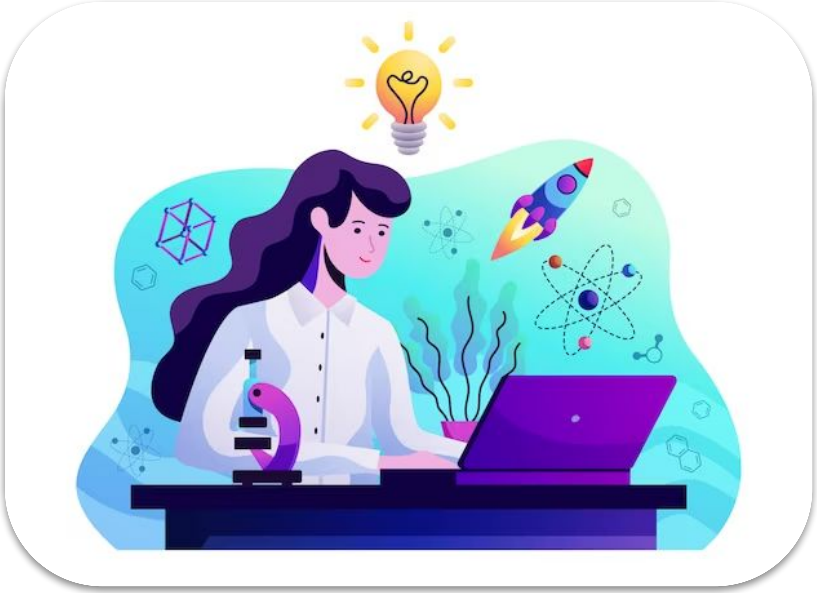
Introducción - Uso de pesticidas

- En particular, el glifosato es un herbicida utilizado para el control de malezas, inhibiendo el crecimiento de las plantas

- Las formulaciones comerciales (ej Roundup®) incluyen mezclas para mejorar la eficacia de la acción del herbicida

El estudio

Se llevó a cabo un estudio para evaluar los efectos genotóxicos de Roundup® ("RU") en embriones de *Caiman latirostris* con el fin de evaluar el riesgo potencial asociado a la exposición sufrida en el medio natural de esta especie



El estudio

El estudio plantea un experimento con el principal objetivo de relacionar el daño en el material genético de embriones de yacaré con la dosis de RU

Obtener una función (modelo) que relacione:

- *la concentración de herbicida (X)*
- *el índice de daño al material genético (Y)*

… y analizar los parámetros de dicho modelo

El experimento

- En condiciones de laboratorio, se expuso a 11 huevos de yacaré a distintas concentraciones de RU entre 0 y 1.750 ug/huevo

  →Huevos asignados al azar a las concentraciones de RU

  →Dosis de RU fijada por el investigador

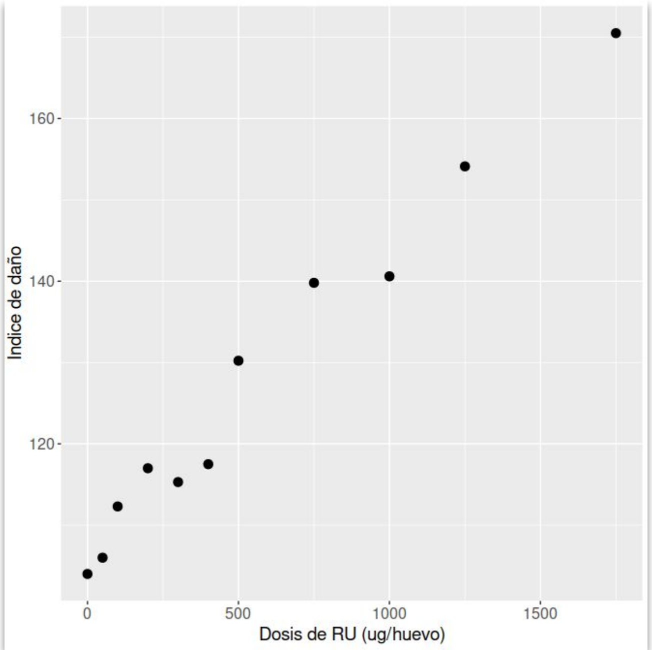
- Al momento de la eclosión se tomaron muestras de sangre y se calculó el daño en el ADN mediante un índice de daño ("DI" debido a su traducción al inglés, Damage Index)

Los datos

¿Qué observan de los datos?

Filter

	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	115.30
6	400	117.50
7	500	130.22
8	750	139.80
9	1000	140.60
10	1250	154.12
11	1750	170.50



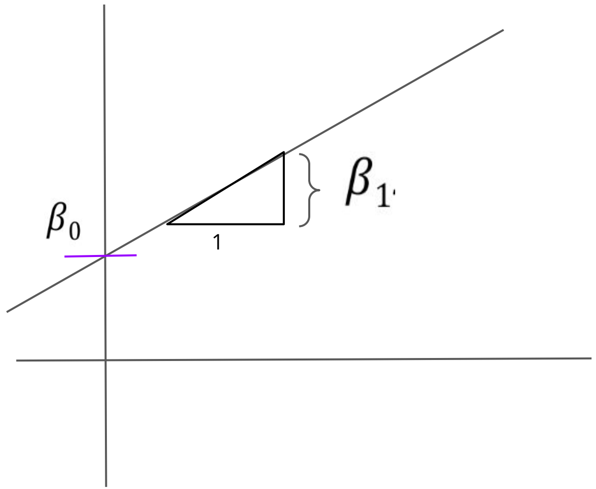
Repaso: función lineal

Es una de las funciones más simples para describir la relación entre dos variables

$$Y = \beta_0 + \beta_1 . X$$

Modelo matemático

β_0, β_1 *son los parámetros del modelo*

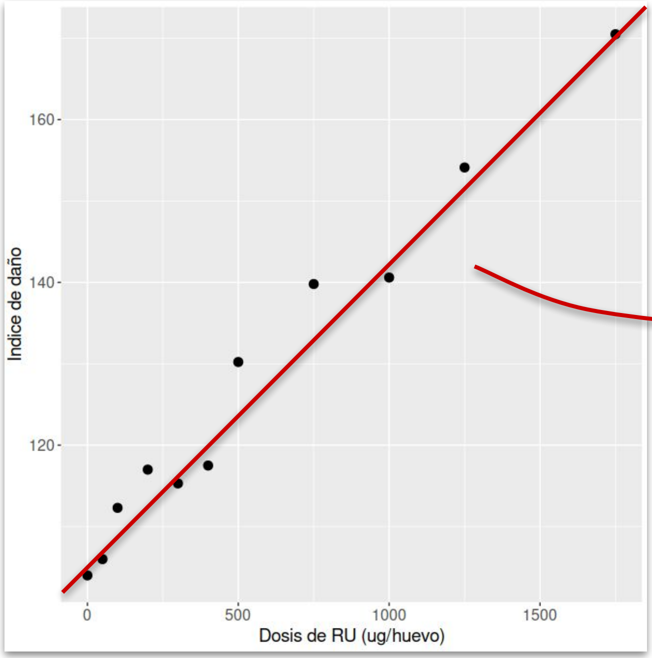


β_0 = ordenada al origen ("intercept") (el punto donde la recta intercepta al eje Y)

β_1 = pendiente de la recta (mide el cambio en Y por cada unidad de cambio en X)

Volviendo al ejemplo: gráfico de dispersión

	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	115.30
6	400	117.50
7	500	130.22
8	750	139.80
9	1000	140.60
10	1250	154.12
11	1750	170.50



¿Es lógico suponer que existe una relación lineal entre ambas variables?

$$DI = \beta_0 + \beta_1 . U$$

Formalizando - Análisis de regresión

- OBJETIVOS (muchos y variados), acá algunos:
  - Describir la relación funcional entre Y y X
  - Determinar cuánta de la variación en Y puede ser explicada por la variación de X y cuánto permanece sin explicar
  - Predecir nuevos valores de Y para valores específicos de X en el dominio estudiado
  - Relación funcional: puede ser de cualquier tipo. En RLS → lineal

Modelo de Regresión Lineal Simple (RLS)

X Y

	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	

- **Modelo de Regresión Lineal Simple**

  $Y_i = \underline{\beta_0 + \beta_1 . x_i} + \varepsilon_i$

  Y_i es la i-ésima observación de la **<u>variable a explicar</u>** Y

  x_i es el i-ésimo valor de la **<u>variable predictora</u>** X

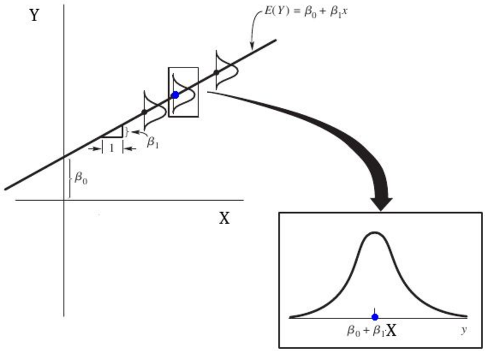
  β_0 y β_1 **<u>parámetros</u>** del modelo: ordenada al origen y pendiente

  ε_i es el error aleatorio, variación de Y no explicada por X;

- **Recta de regresión Poblacional**

  $\beta_0 + \beta_1 . x$

Estadístico: permite la incorporación de un componente **ALEATORIO**



Adaptado de Wackerly et al 2010

Modelo de Regresión Lineal Simple (RLS)

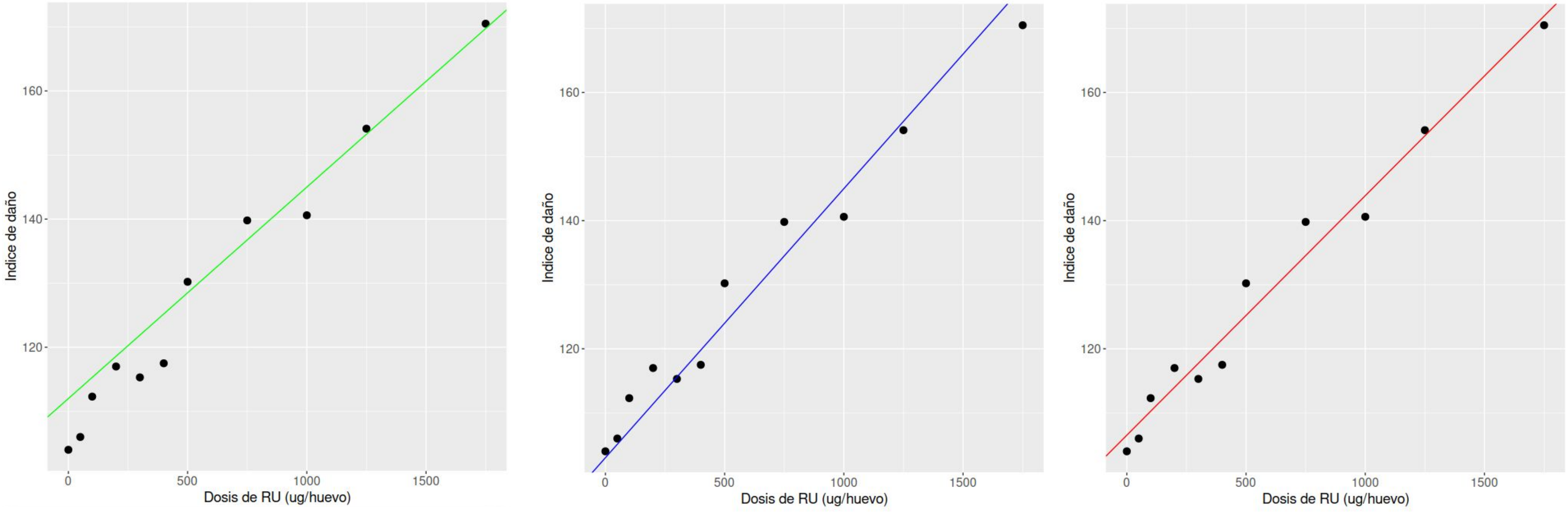
Parámetros

- La función $Y = \beta_0 + \beta_1 . x$ no es observable directamente
- Se estima a través de una muestra como $Y = \hat{\beta}_0 + \hat{\beta}_1 . x$

Estimadores

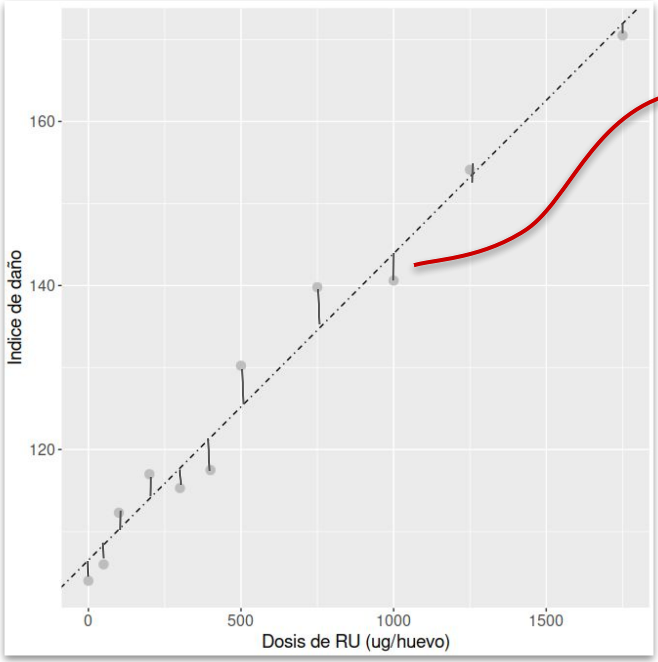
- $\hat{\beta}_0$ y $\hat{\beta}_1$: Estimadores de los parámetros del modelo obtenidos a partir de los pares de datos
- Vamos a ESTIMAR UNA ecuación de la recta a partir de nuestros (11) datos ....
  ¿Cómo? → **la que mejor se <u>ajusta</u> a nuestros datos**

¿Cómo encontramos la recta que mejor se ajusta?



¿La verde? ¿La azul? ¿La roja?

Método de mínimos cuadrados



- **<u>Residuo.</u>** Es la diferencia entre el valor observado (y_i) y el predicho por la recta propuesta $a + b.x_i$

$$y_i - (a + b.x_i)$$

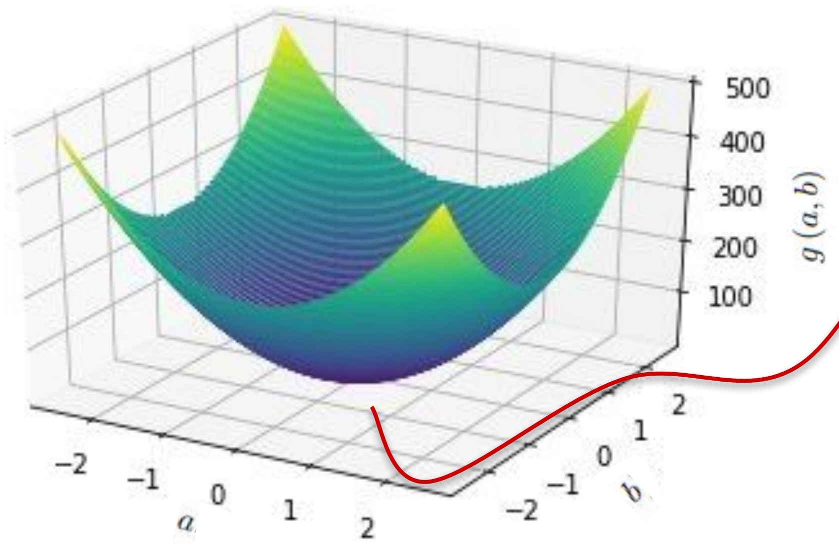
- La "mejor recta" será aquella que minimice la suma de los residuos al cuadrado

¿Por qué?

$$g(a,b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 ,$$

Función que mide el desajuste a la recta

Método de mínimos cuadrados

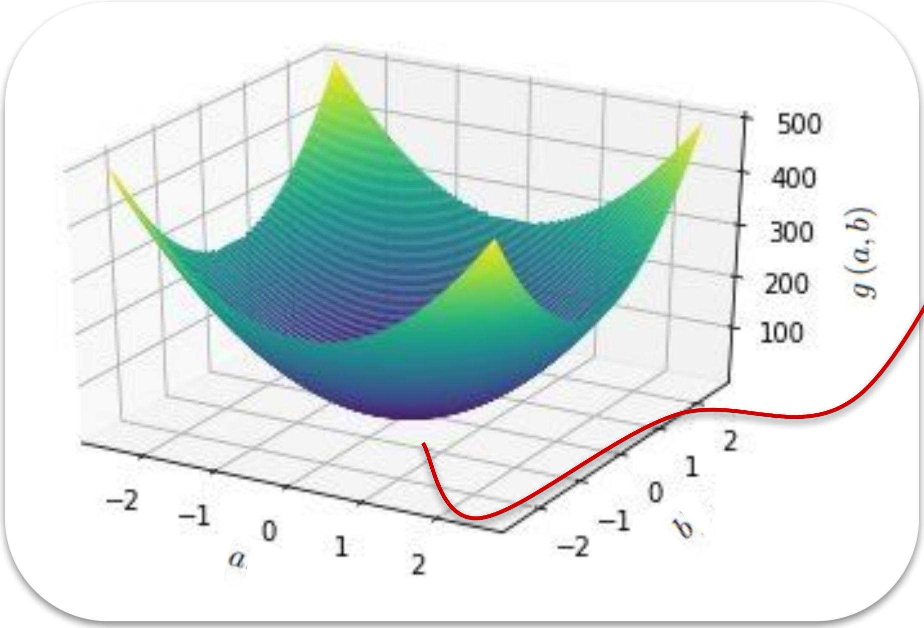


- ✓ Al ser una **<u>función cuadrática</u>** de los parámetros, tiene un mínimo global que además es el **<u>único mínimo local</u>**
- ✓ Lo podemos hallar buscando donde se anula el gradiente: **<u>gradiente *g(a,b)* = 0</u>**.

$$g(a,b) = \sum_{i=1}^{n} (Y_i - (a + bX_i))^2 ,$$

Función que mide el desajuste a la recta

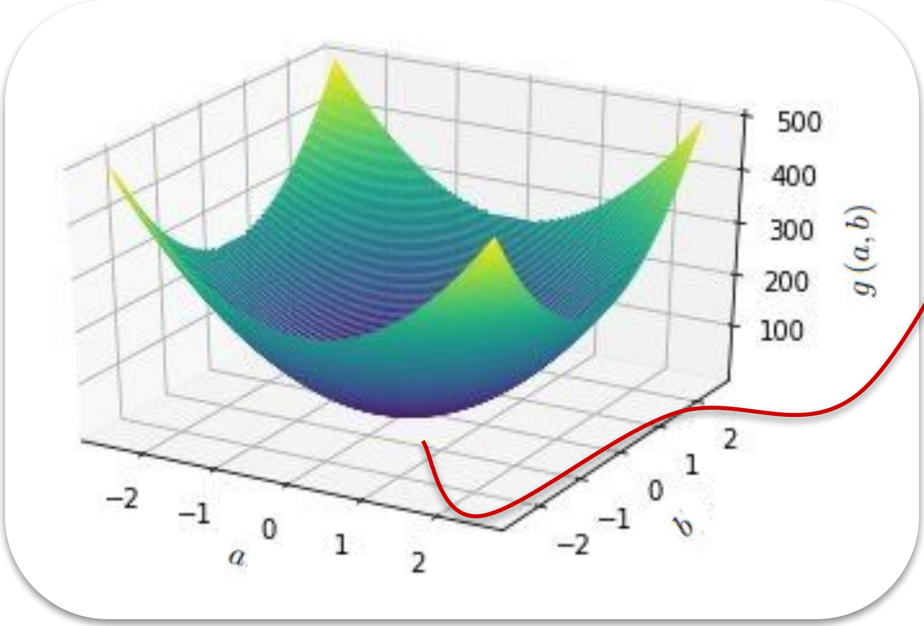
Método de mínimos cuadrados



✓ Entonces, buscamos los parámetros que minimizan g(a,b) derivándola respecto de uno de los parámetros, y luego respecto del otro. Igualamos ambas ecuaciones a cero y obtenemos:

$$\begin{cases} -2\sum_{i=1}^{n}\left(y_i - a - b\,x_i\right) = 0 \\ -2\sum_{i=1}^{n}\left(y_i - a - b\,x_i\right)x_i = 0 \end{cases}$$

Método de mínimos cuadrados



- ✓ Ahora despejamos los parámetros y obtenemos:

$\hat{\beta}_1$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

$\hat{\beta}_0$

$$a = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - b\,\bar{x}$$

Nuestro ejemplo

Misma ecuación anterior, pero reescrita

Insumos

- n pares de observaciones (X_i, Y_i)
- $\bar{X}$ = promedio de las X_i
- $\bar{Y}$ = promedio de las Y_i

	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} =$?
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} =$?

Nuestro ejemplo

Misma ecuación anterior, pero reescrita

Insumos

- *n* pares de observaciones (X_i, Y_i)
- $\bar{X}$ = promedio de las X_i
- $\bar{Y}$ = promedio de las Y_i

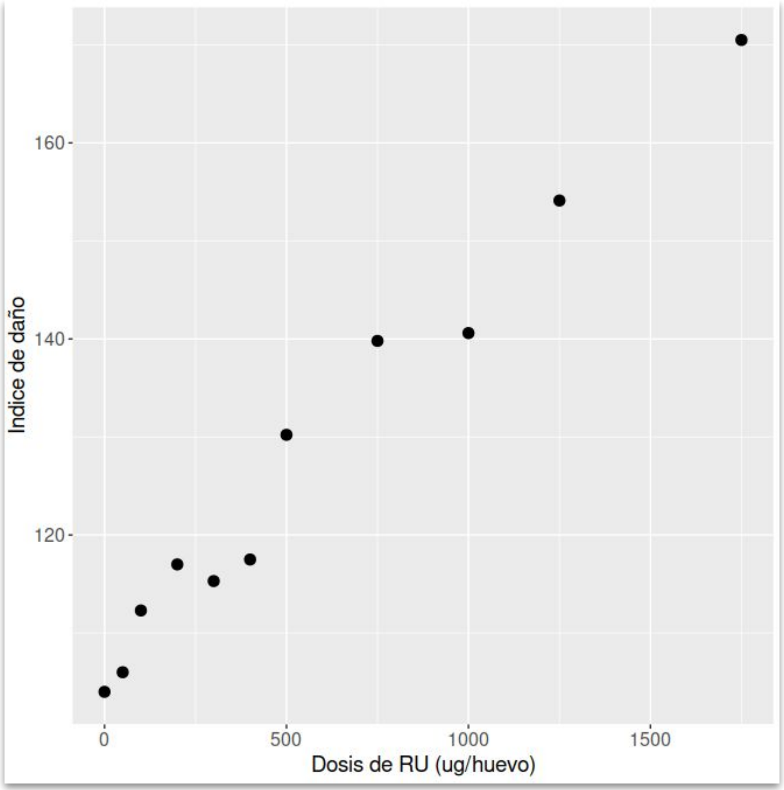
	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	

- $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ = 0.037
- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \, . \, \bar{X}$ = 106.5

La recta ajustada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

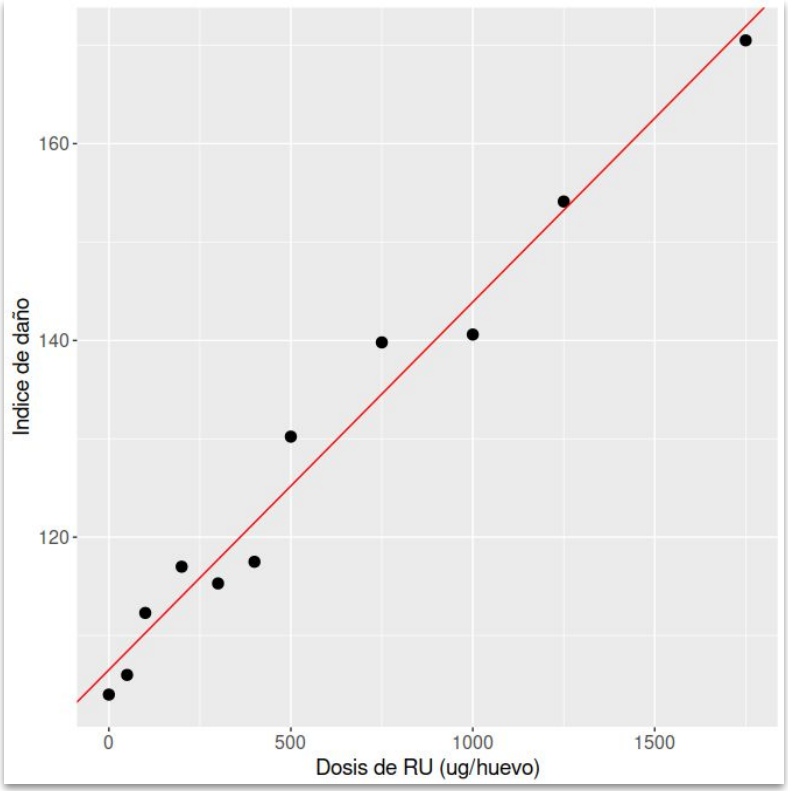
$$\hat{y} = 106.5 + 0.037 * DosisRU$$



La recta ajustada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$

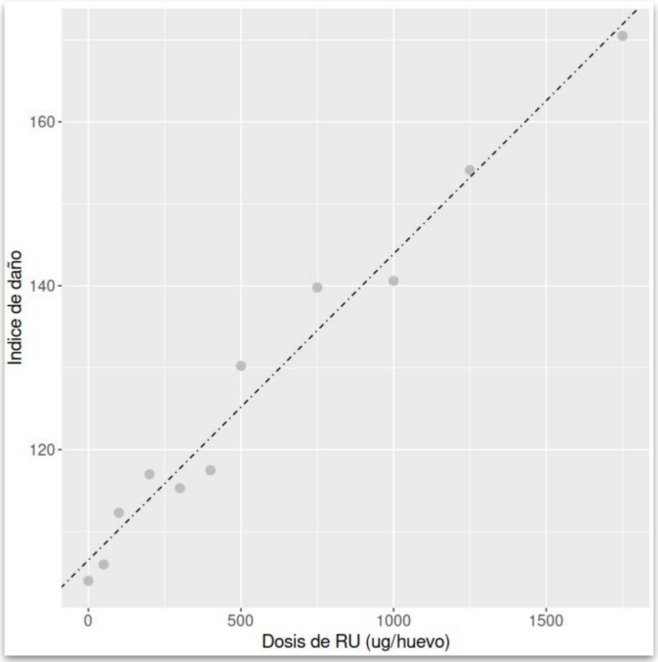


Estimación de la recta

```
# -*- coding: utf-8 -*-
"""
Materia     : Laboratorio de datos - FCEyN - UBA
Clase       : Clase Regresion Lineal
Detalle     : Modelo de Regresion Lineal Simple
Autores     : Maria Soledad Fernandez y Pablo Turjanski
Modificacion: 2023-10-13
"""

# Importamos bibliotecas
import pandas as pd
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
#%%
#################################################################
########  DEFINICION DE FUNCIONES AUXILIARES
#####################################
```

	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	115.30
6	400	117.50
7	500	130.22
8	750	139.80
9	1000	140.60
10	1250	154.12
11	1750	170.50

Volviendo al ejemplo



Los datos (observaciones) ¿están exactamente sobre la recta?

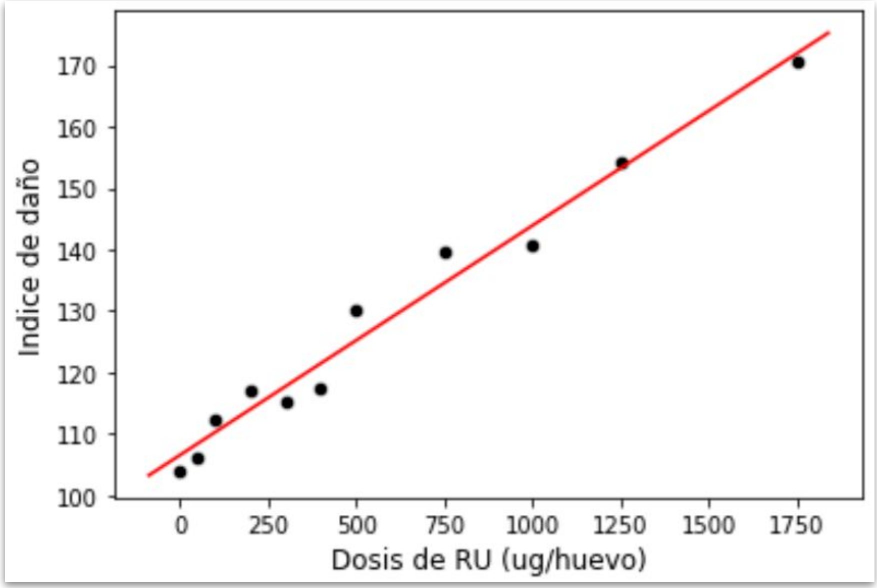
Si repetimos el experimento ¿los puntos se ubicarán exactamente en el mismo lugar?

Entonces, dado un valor de X ¿siempre se va a obtener el mismo valor de Y?

Interpretación de los coeficientes

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$



- **Ordenada al origen (intercept).** Es el valor medio de ID cuando la dosis de RU es 0 (sin herbicida), en este caso 106.5
- **Pendiente.** Por cada unidad adicional de Dosis de RU (ug/huevo), se observa un incremento medio en el índice de daño de 0.037 unidades (de ID...)

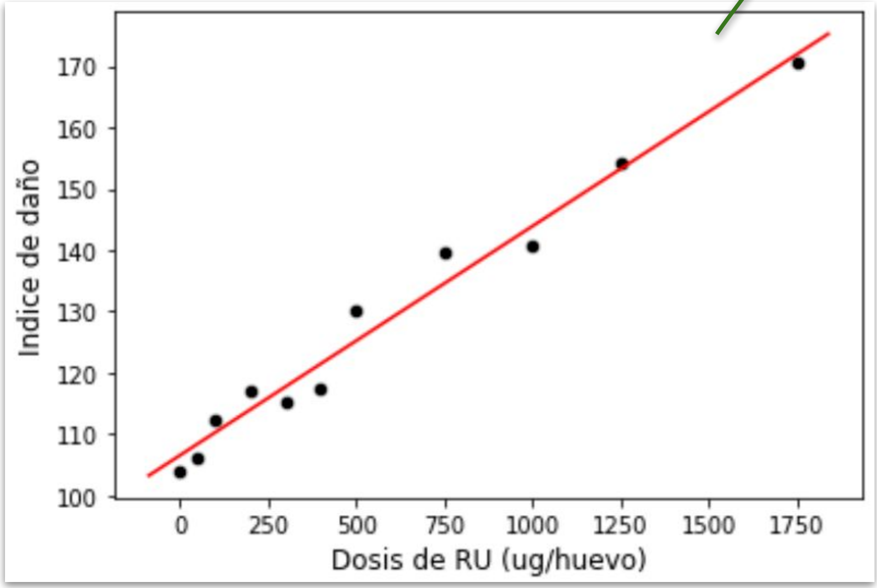
Predicción

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$

Filter

	RU	DI
1	25	
2	600	
3	1500	

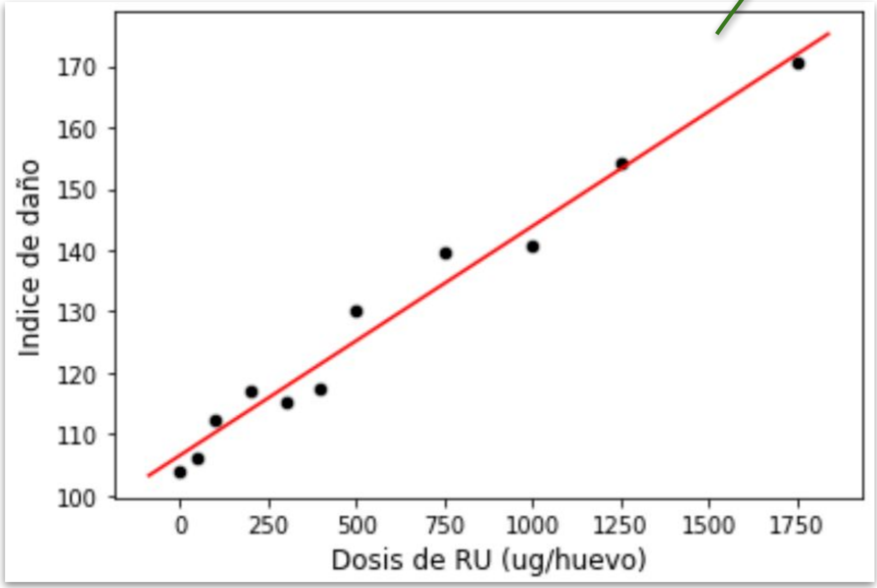


Predicción

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$

	RU	DI
1	25	
2	600	
3	1500	



regresionLinealSimpleRU.py

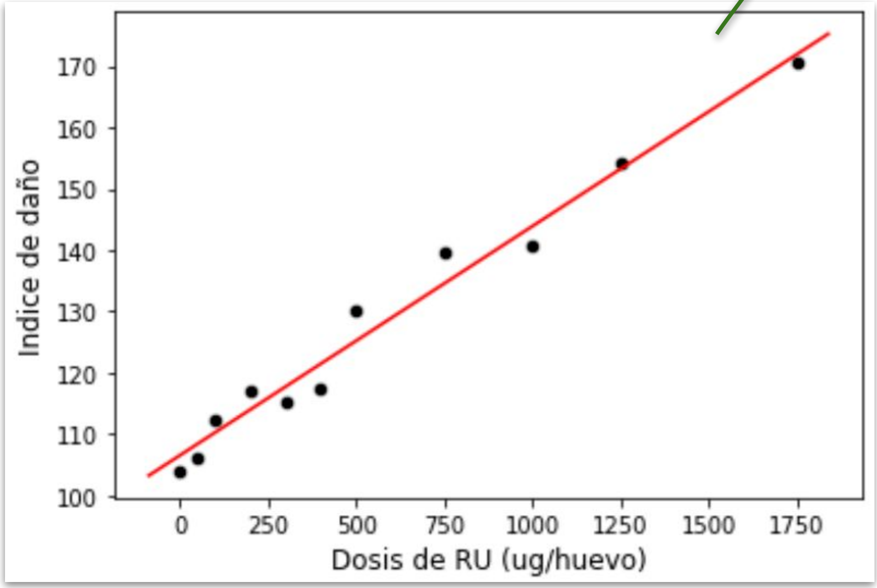
```python
# -*- coding: utf-8 -*-
"""
Materia     : Laboratorio de datos - FCEyN - UBA
Clase       : Clase Regresion Lineal
Detalle     : Modelo de Regresion Lineal Simple
Autores     : Maria Soledad Fernandez y Pablo Turjanski
Modificacion: 2023-10-13
"""

# Importamos bibliotecas
import pandas as pd
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
#%%
############################################################
########  DEFINICION DE FUNCIONES AUXILIARES
####################################
```

Predicción

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$

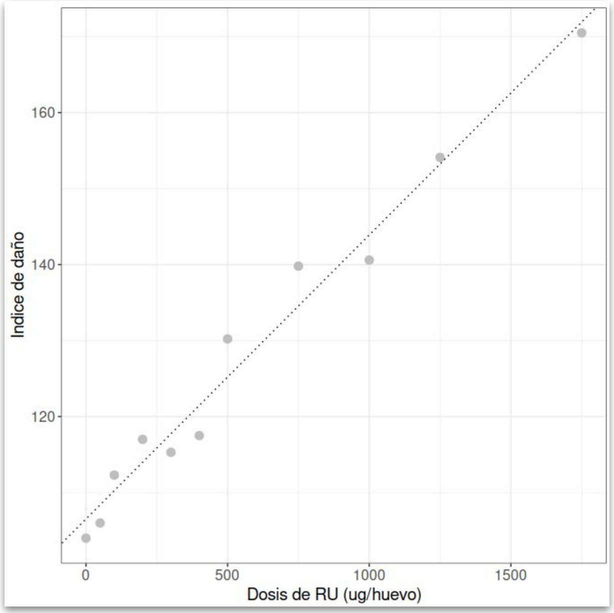
	RU	DI
1	25	107.46
2	600	128.96
3	1500	162.61



regresionLinealSimpleRU.py

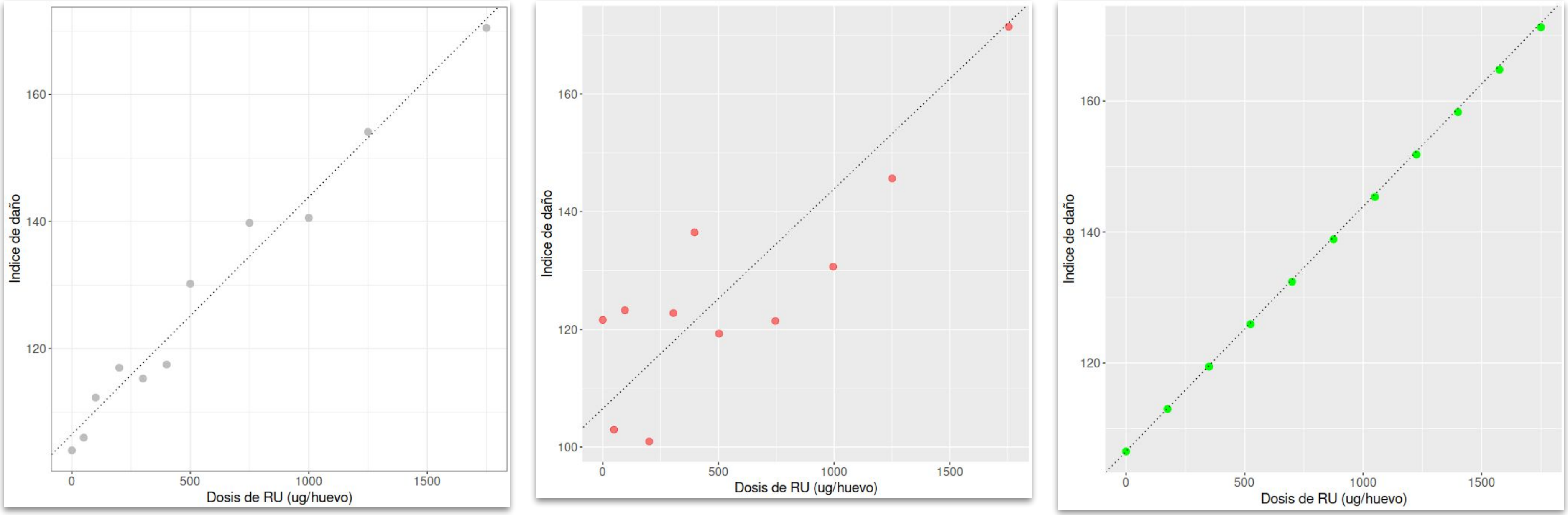
```
# -*- coding: utf-8 -*-
"""
Materia     : Laboratorio de datos - FCEyN - UBA
Clase       : Clase Regresion Lineal
Detalle     : Modelo de Regresion Lineal Simple
Autores     : Maria Soledad Fernandez y Pablo Turjanski
Modificacion: 2023-10-13
"""

# Importamos bibliotecas
import pandas as pd
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
#%%
##########################################################
######## DEFINICION DE FUNCIONES AUXILIARES
```

Varianza del modelo



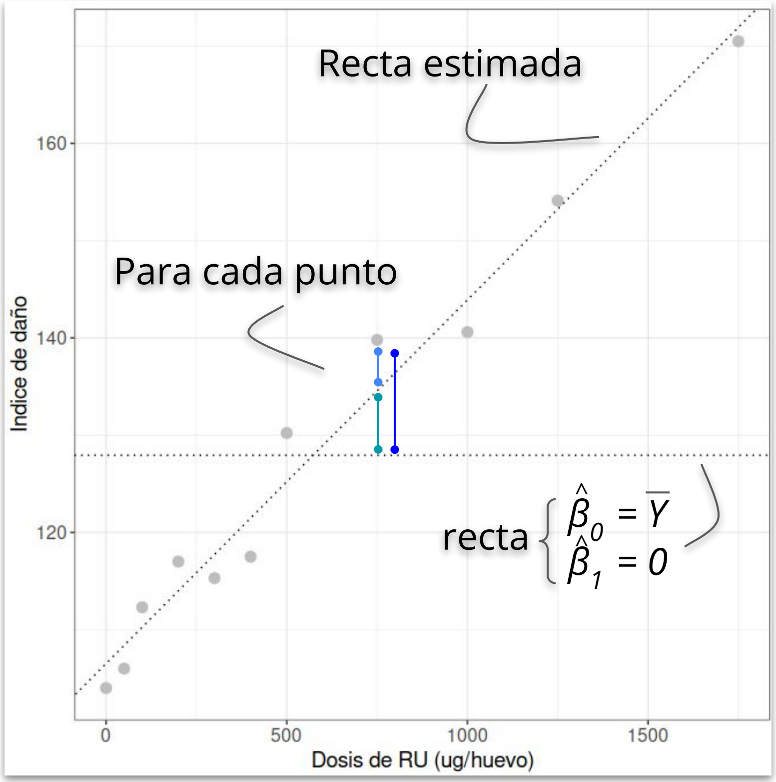
Nuestros datos

Varianza del modelo



Nuestros datos

Otros escenarios



- Si no hay relación entre la dosis de RU y el ID entonces $\begin{cases} \hat{\beta}_0 = \overline{Y} \\ \hat{\beta}_1 = 0 \end{cases}$

- La variabilidad TOTAL del modelo puede separarse en EXPLICADA y NO EXPLICADA

variabilidad total	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$
variabilidad no explicada	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
variabilidad explicada	$\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$

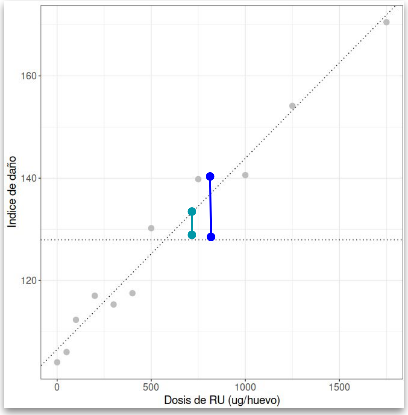
- Coeficiente de determinación (R^2): Mide la proporción de variabilidad de la variable respuesta explicada por variaciones en x, es decir por el modelo de regresión

$$R^2 = \frac{SC\,explic}{SC\,total}$$

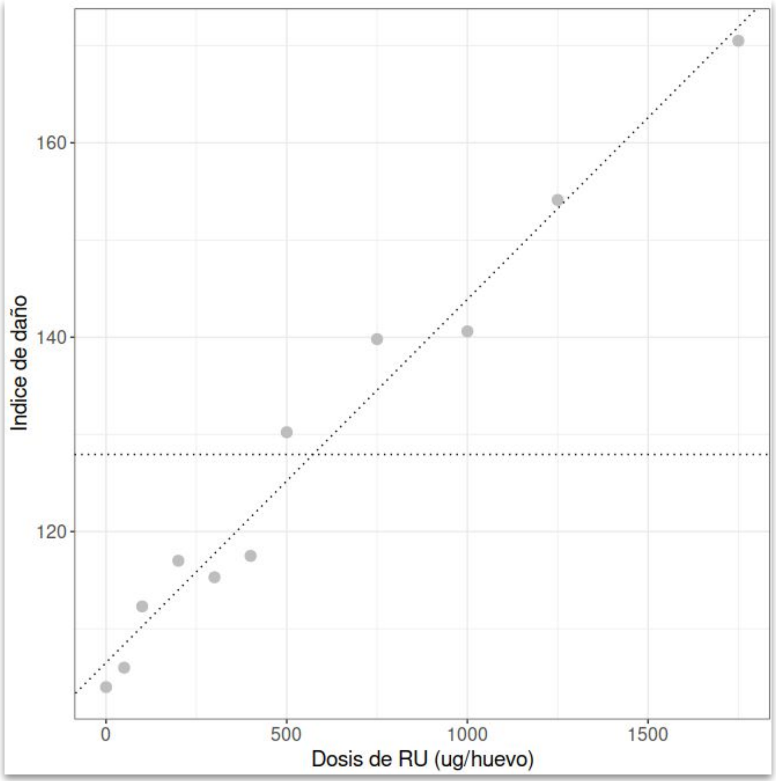
Coeficiente de determinación R^2

- Es una medida de la capacidad predictiva del modelo (de RLS)
- Mide la "proporción de la variabilidad en Y explicada por el modelo" (de RLS)
- No depende de las unidades de medición
- Toma valores entre 0 y 1: $0 \le R^2 \le 1$
- A mayor R^2: más cercanos los puntos a la recta,
- A mayor R^2, mayor "fuerza" para predecir (dentro del rango)

$$R^2 = \frac{SC\,explic}{SC\,total}$$



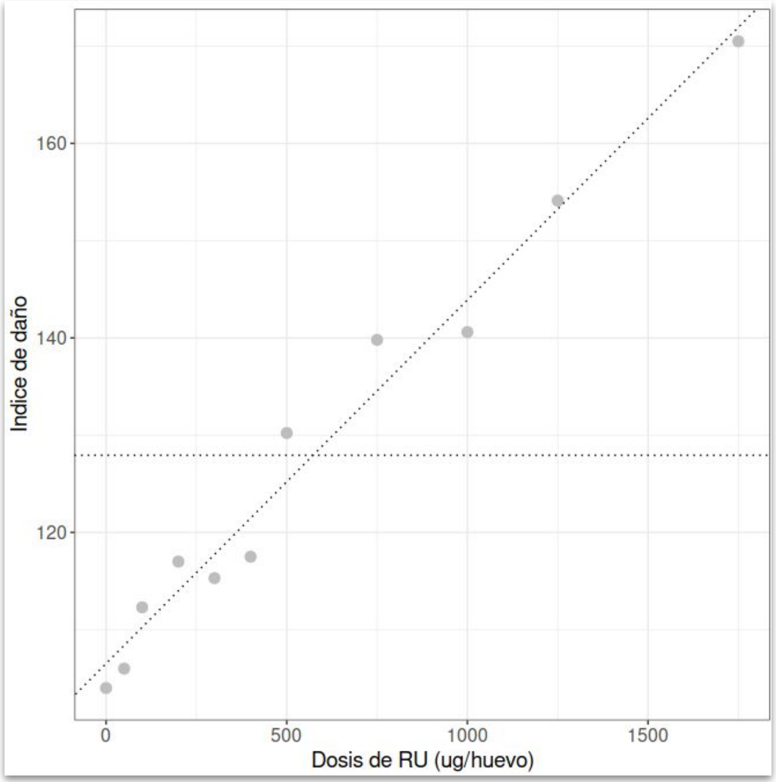
Resumiendo, en nuestro experimento



Ecuación estimada = $\hat{y} = 106.5 + 0.037 * DosisRU$

$R^2 = 0.97$

Resumiendo, en nuestro experimento



Ecuación estimada = $\hat{y} = 106.5 + 0.037 * DosisRU$

R^2 = 0.97

Si repetimos el experimento, ¿obtendremos los mismos valores de $\hat{\beta}_0$ y $\hat{\beta}_1$? y R^2?

Error cuadrático medio (MSE *por mean squared error*)

El error cuadrático medio (de <u>**cualquier**</u> modelo de regresión) mide el promedio de los errores al cuadrado, es decir:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Y_i son los valores observados (reales)
$\hat{Y}_i$ son los valores estimados por el modelo.

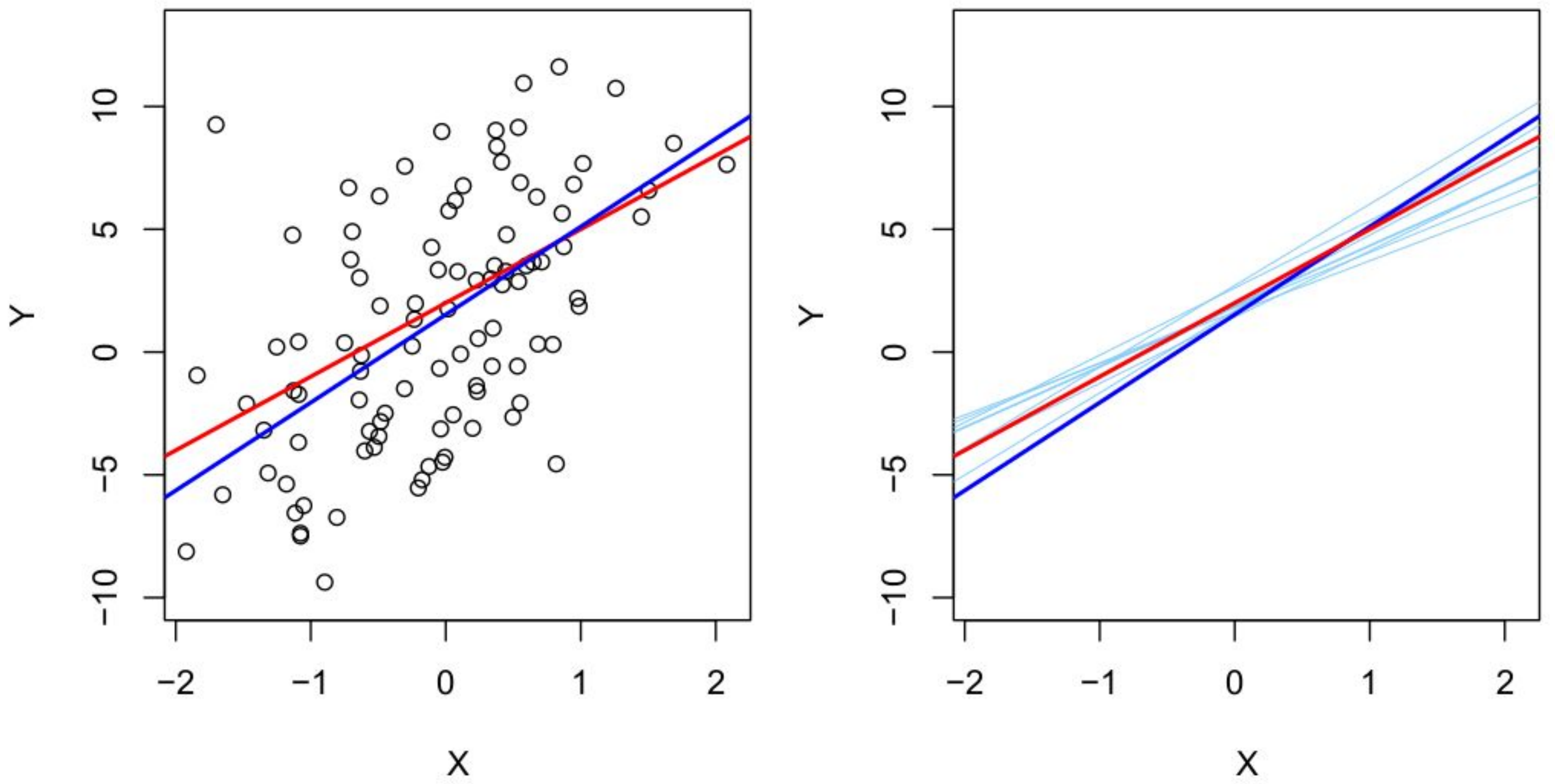
Ejemplo - Regresión y variabilidad de los estimadores

Imaginemos que a cada uno de nosotros nos encargan la realización de un experimento similar al que vimos.

Para ello vamos a:

1. Obtener una muestra. En https://msfernandez.shinyapps.io/applabodatos/ seguir las instrucciones para obtener una muestra. (*)
2. Realizar un gráfico de dispersión de ID en función de la concentración de RU.
3. Estimar la recta de regresión. Interpretar los coeficientes. Escribir la ecuación estimada del modelo.
4. Calcular e interpretar el coeficiente de determinación R^2
5. En la planilla compartida escribir el valor estimado para β_0 , β_1 , y el R^2, para cada una de las muestras obtenidas
6. Comparar los resultados obtenidos en las distintas muestras.

(*) el modelo poblacional (a partir del cual se obtienen los datos) es desconocido para ustedes.



Y
X
−10
−5
0
5
10
−2
−1
0
1
2

Ejemplo adaptado de Poletta y cols (2009)

ELSEVIER

Mutation Research/Genetic Toxicology and Environmental Mutagenesis

Volume 672, Issue 2, 31 January 2009, Pages 95-102

Genotoxicity of the herbicide formulation Roundup® (glyphosate) in broad-snouted caiman (*Caiman latirostris*) evidenced by the Comet assay and the Micronucleus test

G.L. Poletta [a b c], A. Larriera [a d], E. Kleinsorge [b], M.D. Mudry [c]

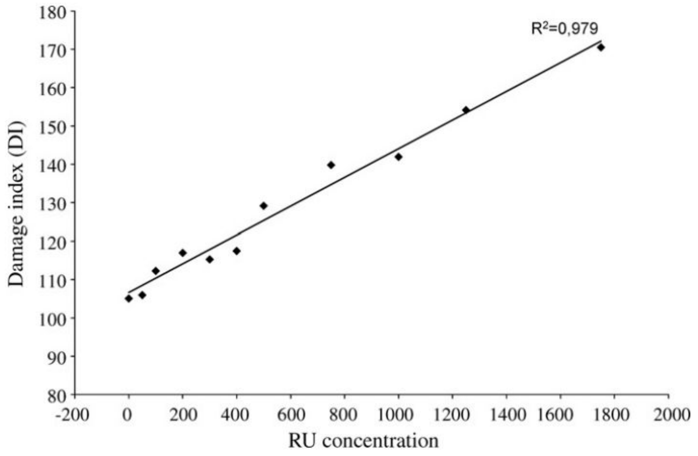


Fig. 3. RU concentration dependent effect for E_1 and E_2 DI mean data. $R^2 = 0.979$, $p < 0.001$.

https://www.sciencedirect.com/science/article/pii/S138357180800301X

Auto MPG

- Dataset de características de autos y consumo de combustible
- Provisto en forma libre por UC Irvine (https://archive.ics.uci.edu/dataset/9/auto+mpg)
- Trataremos de encontrar que hace que aumente el consumo de combustible

Ejercicios

- Graficar las relaciones entre mpg, weight, displacement y acceleration.
  - ¿Qué variable parece más relacionada con mpg?
- Ajustar un modelo de regresión lineal simple para predecir mpg a partir de una sola variable (por ejemplo, weight).
  - Calcular los coeficientes de la recta, R^2 y el MSE.
  - Hacer un gráfico con los datos y la línea de regresión.
- Repetir el modelo con otra variable (por ejemplo, displacement o acceleration).
  - Comparar los resultados. ¿Cuál modelo explica mejor el consumo?
- Si aumenta el peso del auto en 100 unidades, ¿cuánto cambia el consumo estimado?

Algunos comentarios

- Estudio experimental: posibilidad de establecer relación causal
- No extrapolar
- Hay que tener cuidado con observaciones atípicas e influyentes

Predicción

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

$$\hat{y} = 106.5 + 0.037 * DosisRU$$

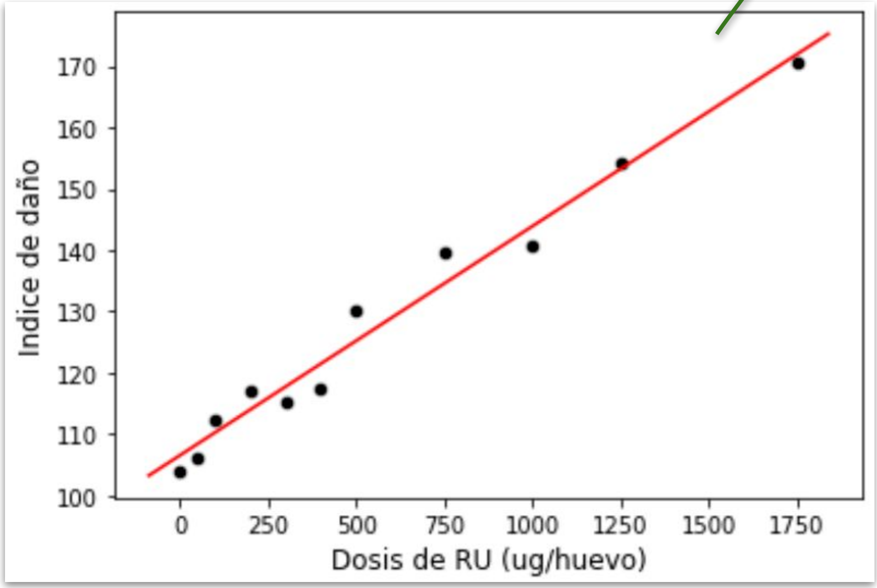
¡Podemos también usar software!
(veamos con python el ejemplo subido al campus)

regresionLinealSimpleRU.py

```python
# -*- coding: utf-8 -*-
"""
Materia     : Laboratorio de datos - FCEyN - UBA
Clase       : Clase Regresion Lineal
Detalle     : Modelo de Regresion Lineal Simple
Autores     : Maria Soledad Fernandez y Pablo Turjanski
Modificacion: 2023-10-13
"""

# Importamos bibliotecas
import pandas as pd
import numpy as np
from sklearn import linear_model
import matplotlib.pyplot as plt
import seaborn as sns
#%%
##########################################################
######## DEFINICION DE FUNCIONES AUXILIARES
##########################################################
```

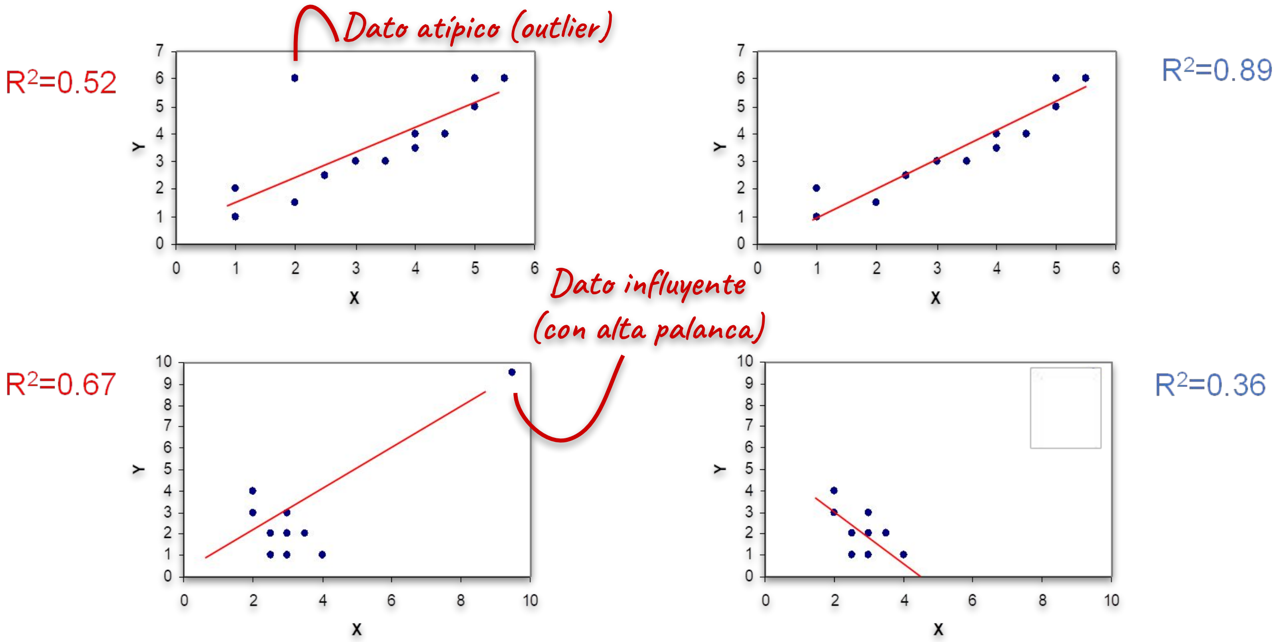
Filter

	RU	DI
1	25	✓
2	150000	✗
3	-150000	✗



Observaciones atípicas e influyentes

- **Atípicas (outliers en Y):** Observaciones con un patrón distinto al resto de los datos, que producen un residuo grande
- **Influyentes (con alta palanca):** Observaciones cuyo valor de X se encuentra alejado del promedio y que tienen mucho peso en las estimaciones de los parámetros. Al ser eliminadas pueden provocan cambios sustanciales en las estimaciones



Dataset “Anscombe”

- Dataset “Anscombe” (1973, Francis Anscombe)
- Muestra la importancia de **graficar para visualizar el efecto de datos atípicos y observaciones influyentes sobre las propiedades estadísticas**

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Para cada dataset:

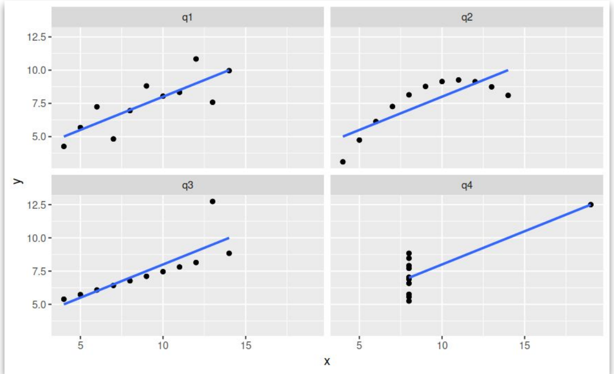
1. Generar el modelo de RLS y reportar
   a. intercept
   b. pendiente
   c. R^2
2. Realizar el gráfico de dispersión y la recta estimada por el RLS

Dataset “Anscombe”

Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

	intercept	pendiente	R2
x1	3.000091	0.5000909	0.6665425
x2	3.000909	0.5000000	0.6662420
x3	3.002455	0.4997273	0.6663240
x4	3.001727	0.4999091	0.6667073



Modelos lineales

- Regresión lineal simple (vimos)
- Regresión lineal múltiple
- Regresión polinomial

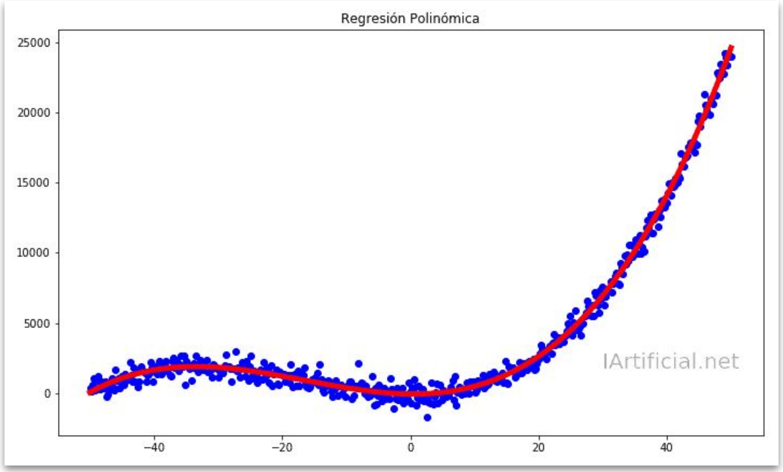
$\beta_0 + \beta_1 . x$

$\beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3$

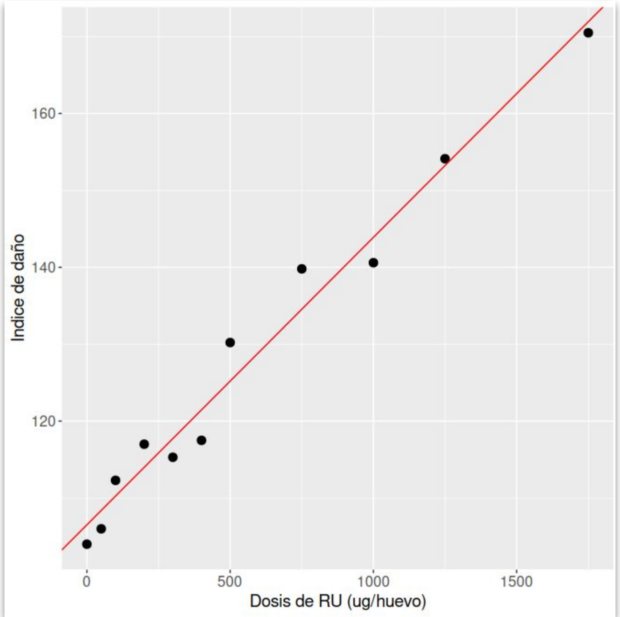
$\beta_0 + \beta_1 . x + \beta_2 . x^2 + \beta_3 . x^3$

Lineal
en
QUÉ

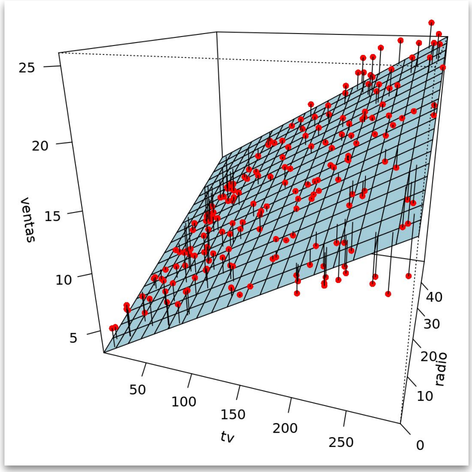
Algunos modelos lineales



$\beta_0 + \beta_1 . x + \beta_2 . x^2 + \beta_3 . x^3$



$\beta_0 + \beta_1 . x$



$\beta_0 + \beta_1 . x_1 + \beta_2 . x_2$

En todos los casos, la función es **lineal** en los **parámetros** del modelo.

$$\beta_0 + \beta_1 . x$$

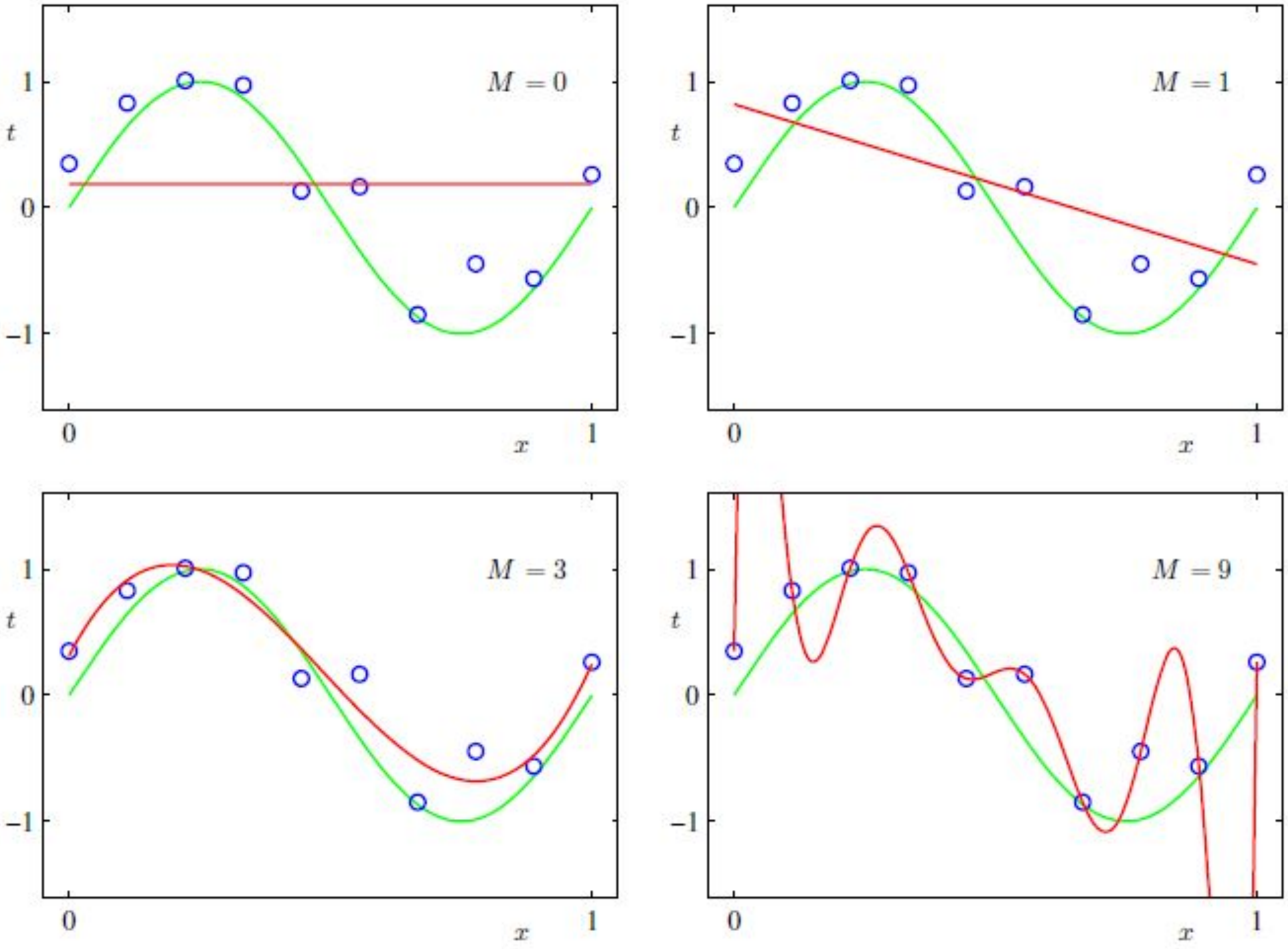
$$\beta_0 + \beta_1 . x_1 + \beta_2 . x_2 + \beta_3 . x_3$$

$$\beta_0 + \beta_1 . x + \beta_2 . x^2 + \beta_3 . x^3$$

Ejercicios - volvemos a los autos

- Ajustar un modelo de regresión lineal para predecir mpg usando dos variables como predictores (por ejemplo: weight y acceleration).
  - Calcular el MSE y los coeficientes.
  - ¿Qué variable tiene más peso en la predicción?
  - ¿El modelo mejora respecto al de una sola variable?
- Elegir una variable (por ejemplo, weight) y ajustar un modelo no lineal usando un polinomio de grado 2 para predecir mpg.
  - Comparar el MSE con el modelo lineal simple.
  - ¿Captura mejor la relación entre las variables?

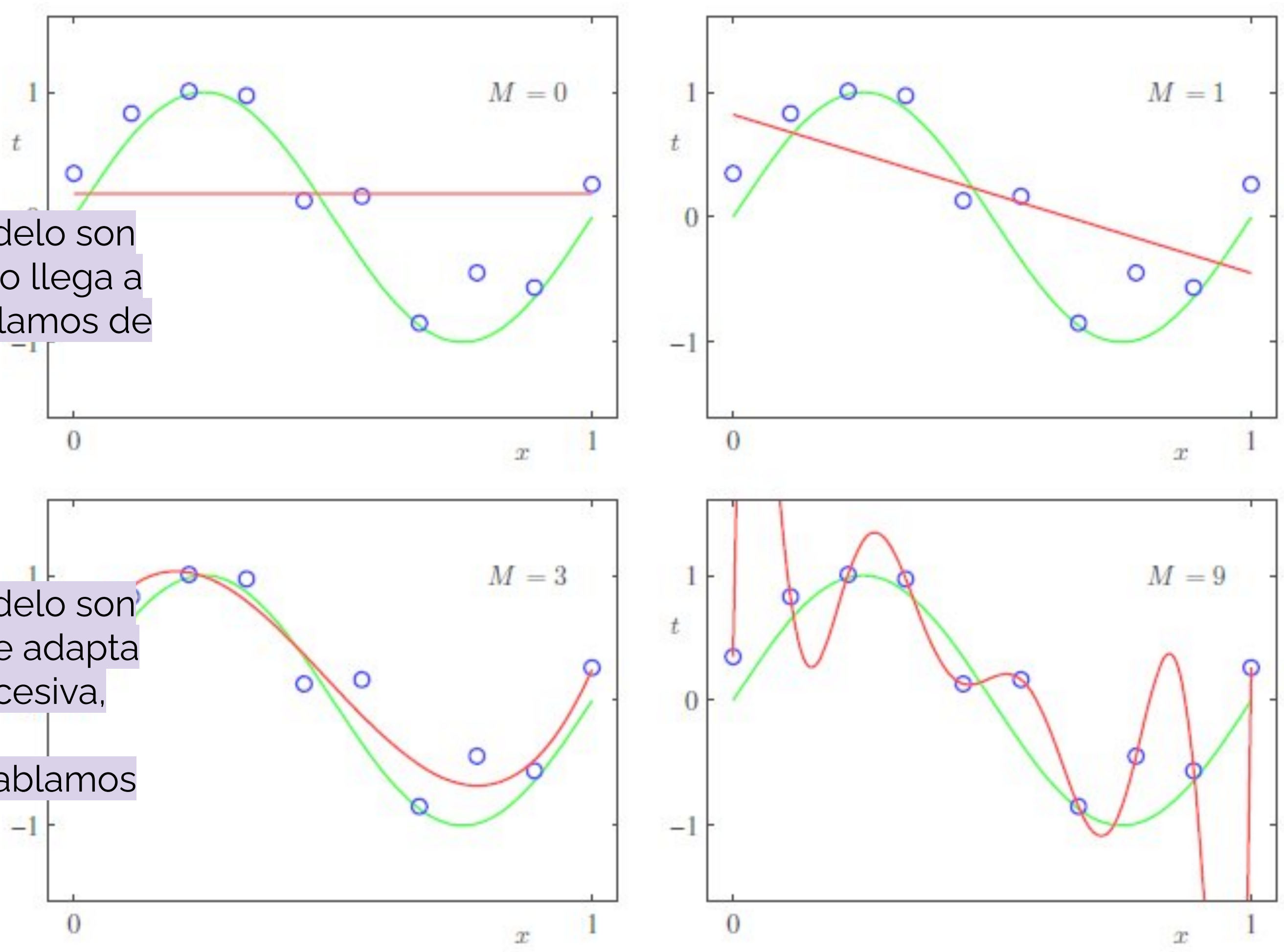
¿Cómo sé si mi modelo realmente **mejora** al agregar parámetros?



(Bishop, "Pattern Recognition and Machine Learning")

Si los parámetros del modelo son insuficientes, el modelo no llega a explicar lo suficiente, hablamos de **sub-ajuste**.

Si los parámetros del modelo son demasiados, el modelo se adapta a los datos de manera excesiva, perdiendo así capacidad explicativa y predictiva, hablamos de **sobre-ajuste**.



(Bishop, "Pattern Recognition and Machine Learning")

Conclusiones

- Vimos varios modelos en los que se pretende explicar o predecir una variable contínua a partir de otras variables.
- Comparamos modelos según una medida de su bondad de ajuste, en este caso el error cuadrático medio o R^2. Hay muchas más.
- Vimos que no siempre más parámetros dan un mejor modelo.
- Vimos que para evaluar cuán bueno es el modelo, hay que ver cómo se desempeña con datos nuevos, distintos a los que usamos para entrenarlo.

Cierre

- Modelos de regresión
- Modelo de regresión lineal simple
- Regresión múltiple y polinomial
- Ejemplos y precauciones

Tarea

Resolver la guía de ejercicios de regresión.

Bibliografía

Libros:

- Introduction to Machine Learning with Python, Müller & Guido
- Machine Learning - Mitchell
- Introduction to Statistical Learning with Applications y Python - James, Witten, Hastie, Tibshirani, Taylor