



DEPARTAMENTO
DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Laboratorio de datos

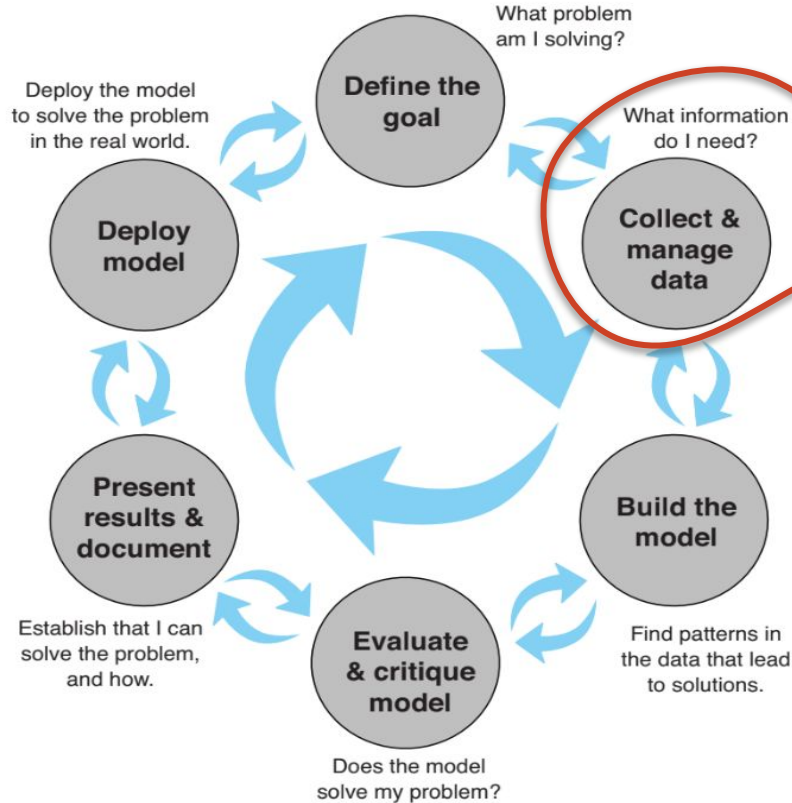
Regresión Lineal

2do Cuatrimestre 2024

Laje, LópezRosenfeld, de Erausquin

(Agradecimientos: Soledad Fernández - Pablo Turkanski -
Manuela Cerdeiro)

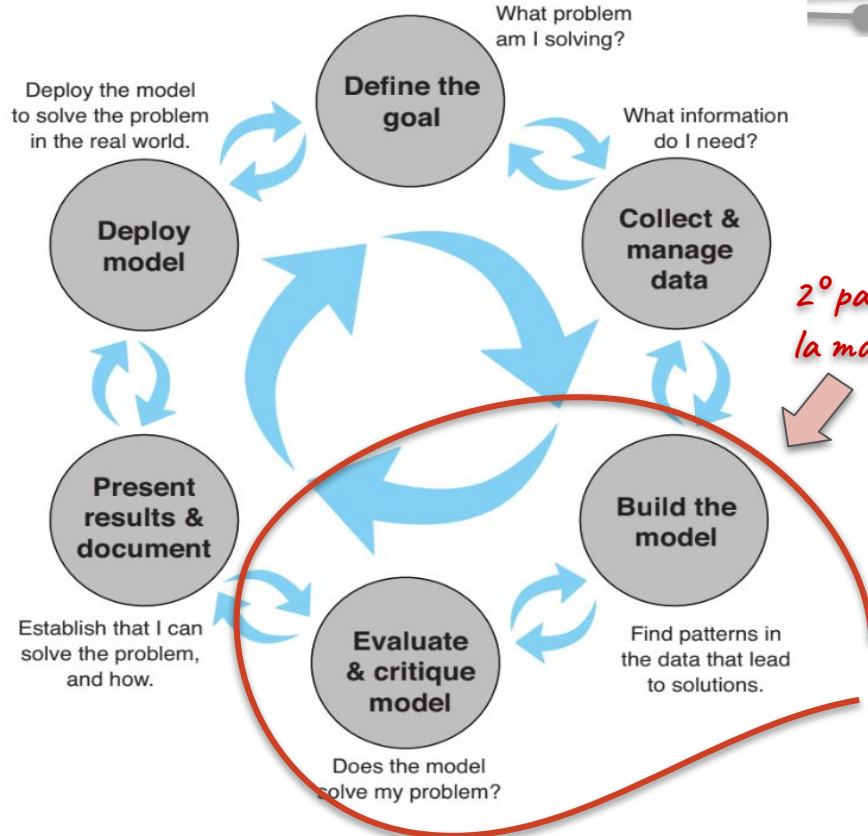
Recorrido de la materia (hasta ahora)



1º parte de la materia

- ✓ Lenguaje de programación (Python)
- ✓ Modelado conceptual de los datos (DER)
- ✓ Representación de los datos (modelo relacional)
- ✓ Formas de consultar los datos (AR/SQL)
- ✓ Recomendaciones para el diseño (Normalización)
- ✓ Calidad de datos
- ✓ Leyes acerca de la Protección de Datos

Recorrido de la materia (hasta ahora)



- ✓ Visualización y Exploración de los datos
- ✓ Intro a Modelado: Clasificación y Regresión
- ✓ Clasificación: Árboles de decisión
- ✓ Evaluación y selección de modelos
- ✓ Regresión y KNN

Regresión

Objetivo:

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

Ejemplo: altura madre - altura hijo.

Podemos querer hacerlo para entender, explicar o predecir.

Clasificación vs. Regresión

En clasificación buscamos explicar una variable 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 \mathbb{R} o en \mathbb{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

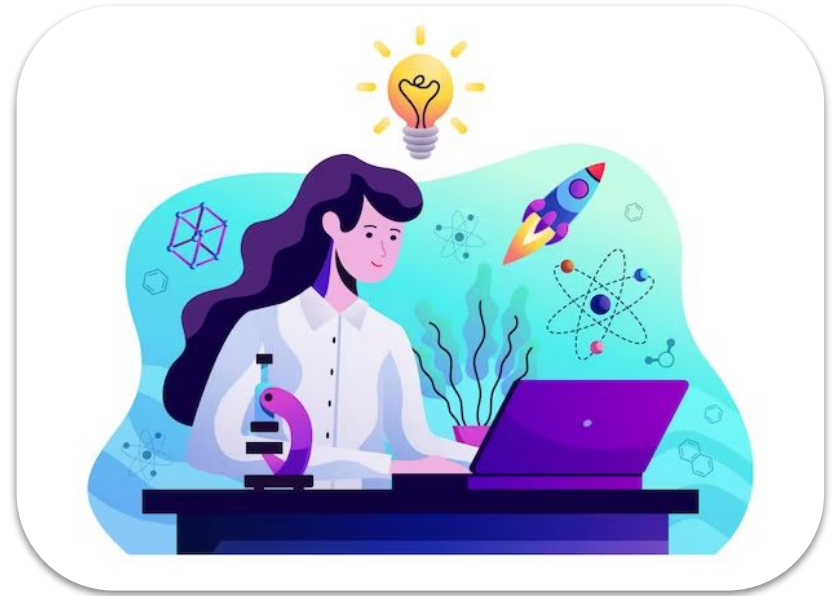


Introducción - Uso de pesticidas

- ✓ 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

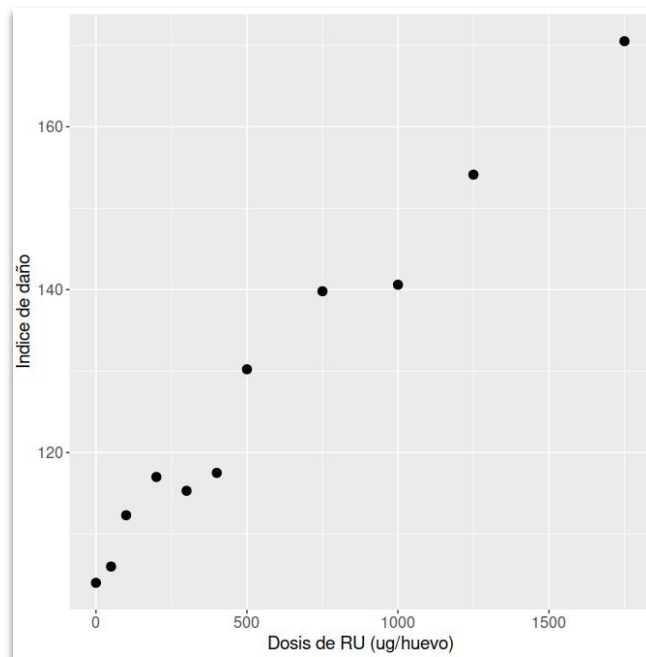
El experimento

- ❑ 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?

	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



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 \rightarrow lineal

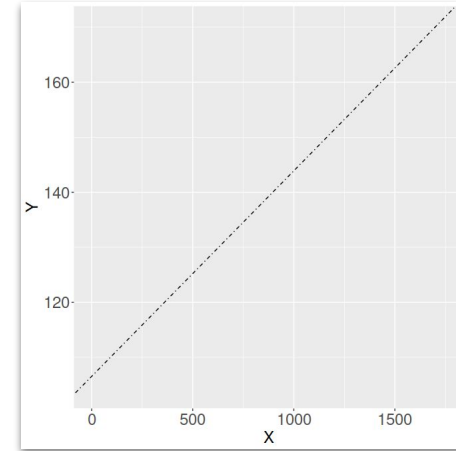
Repaso: función lineal

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

$$Y = a + b.X$$

*Modelo
matemático*

a y b son los parámetros del modelo

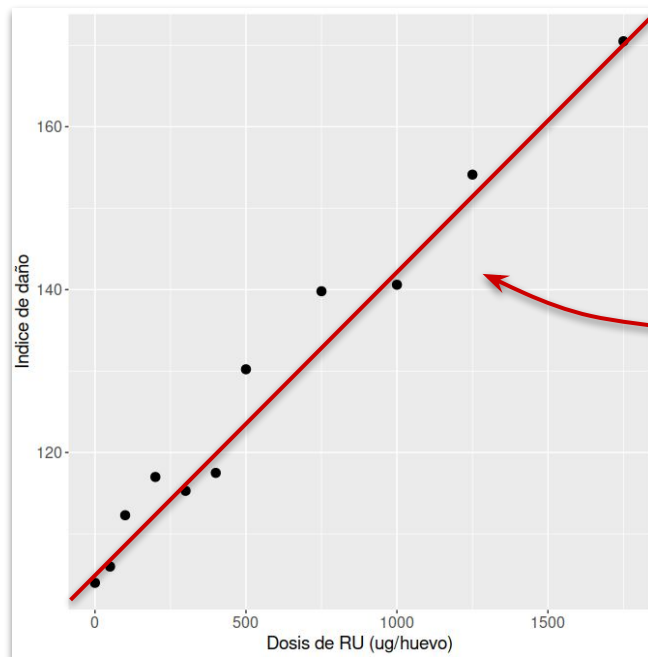


a = ordenada al origen ("intercept") (el punto donde la recta intercepta al eje Y)

b = 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?

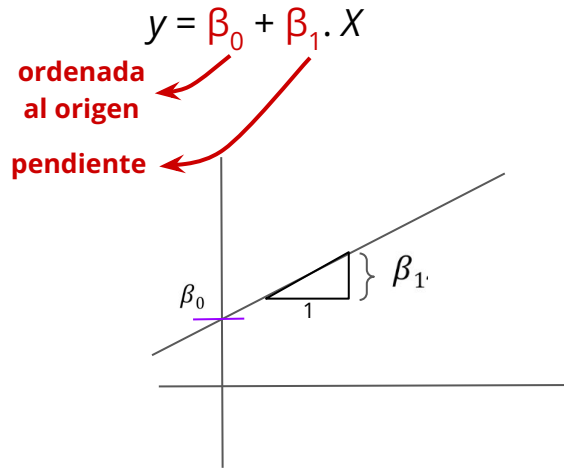
$$DI = a + b \cdot RU$$



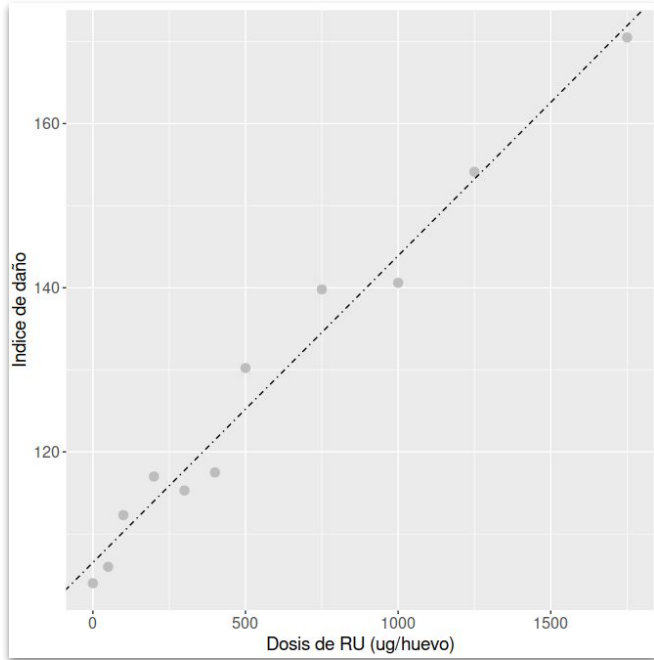
Modelo de Regresión Lineal Simple (RLS)

Sabemos que el comportamiento de la función lineal es **determinístico** ...

"El comportamiento de Y puede ser descrito por una función SIN ERROR"



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?

Modelo de Regresión Lineal Simple (RLS)

❑ Modelo de Regresión Lineal Simple

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i$$

Y_i es la i-ésima observación de la **variable a explicar** Y

x_i es el i-ésimo valor de la **variable predictora** X

β_0 y β_1 **parámetros** del modelo: ordenada al origen y pendiente

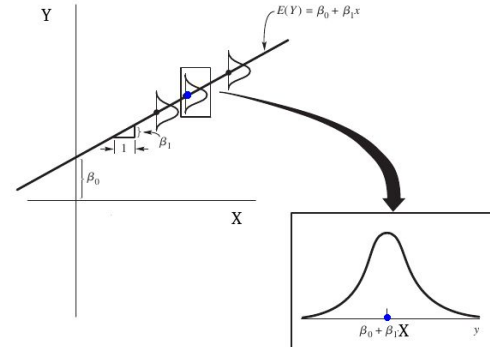
ε_i es el error aleatorio, variación de Y no explicada por X;

❑ Recta de regresión Poblacional

$$\beta_0 + \beta_1 \cdot x$$

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

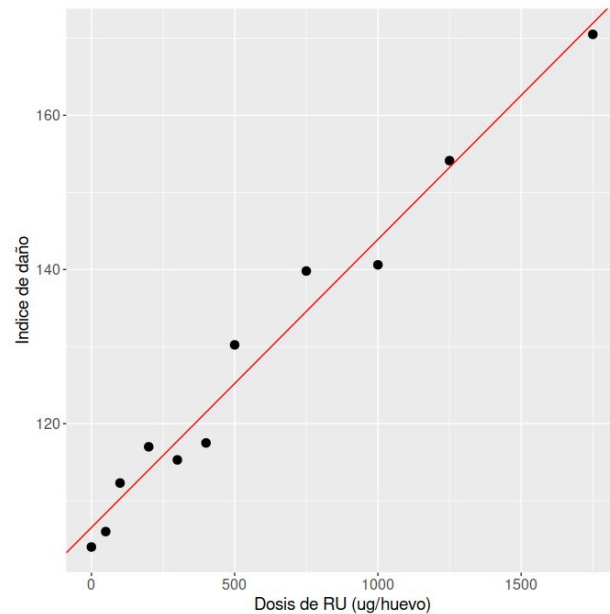
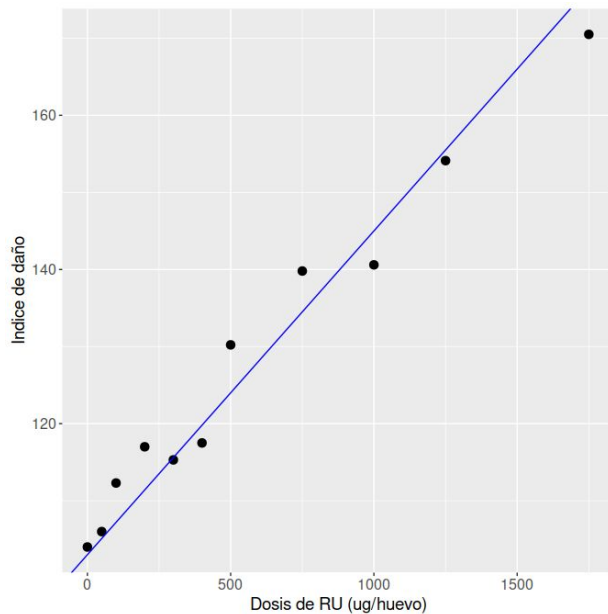
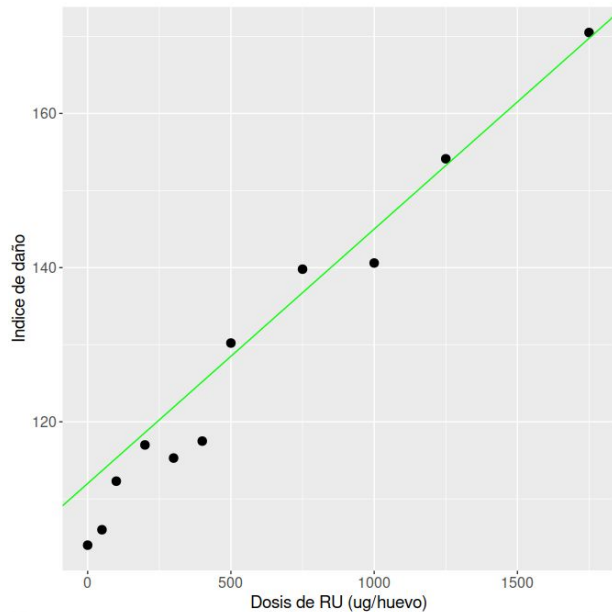
Estadístico: permite la incorporación de un componente **ALEATORIO**



Modelo de Regresión Lineal Simple (RLS)

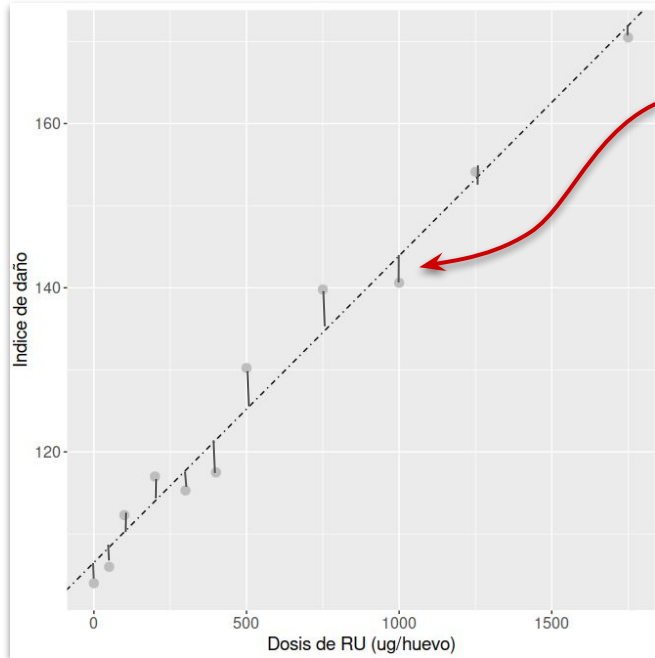
- ❑ La función $Y = \beta_0 + \beta_1 \cdot x$ no es observable directamente **Parámetros**
- ❑ Se estima a través de una muestra como $Y = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$ **Estimadores**
- ❑ $\hat{\beta}_0$ y $\hat{\beta}_1$: Estimadores de los parámetros del modelo obtenidos a partir de los pares de datos
- ❑ Vamos a ESTIMAR UNA ecuación de la recta a partir de nuestros (11) datos
¿Cómo? → **la que mejor se ajusta a nuestros datos**

¿Cómo encontramos la recta que mejor se ajusta?



¿La verde? ¿La azul? ¿La roja?

Método de mínimos cuadrados



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

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

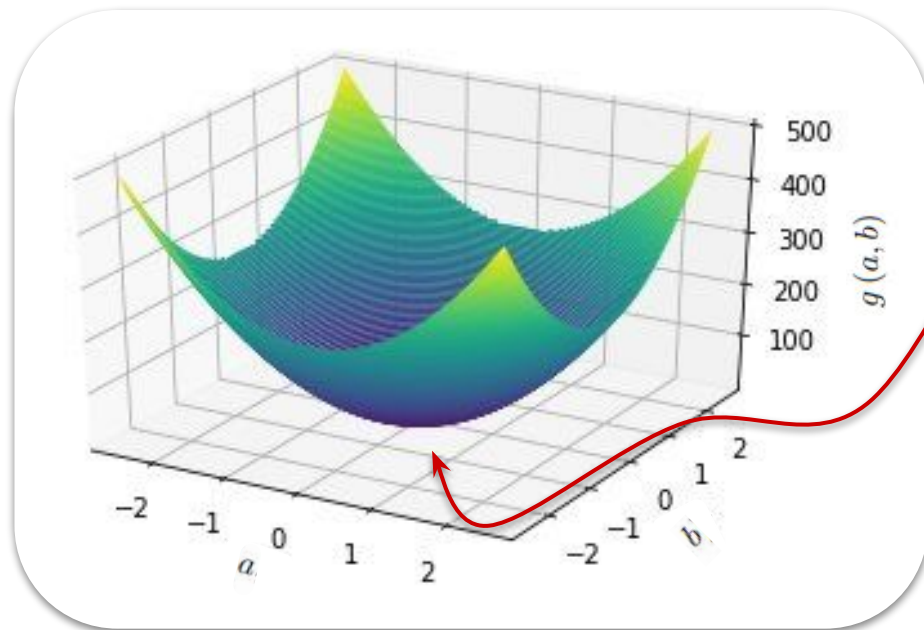
- ❑ La “mejor recta” será aquella que minimice la suma de los residuos al cuadrado

¿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

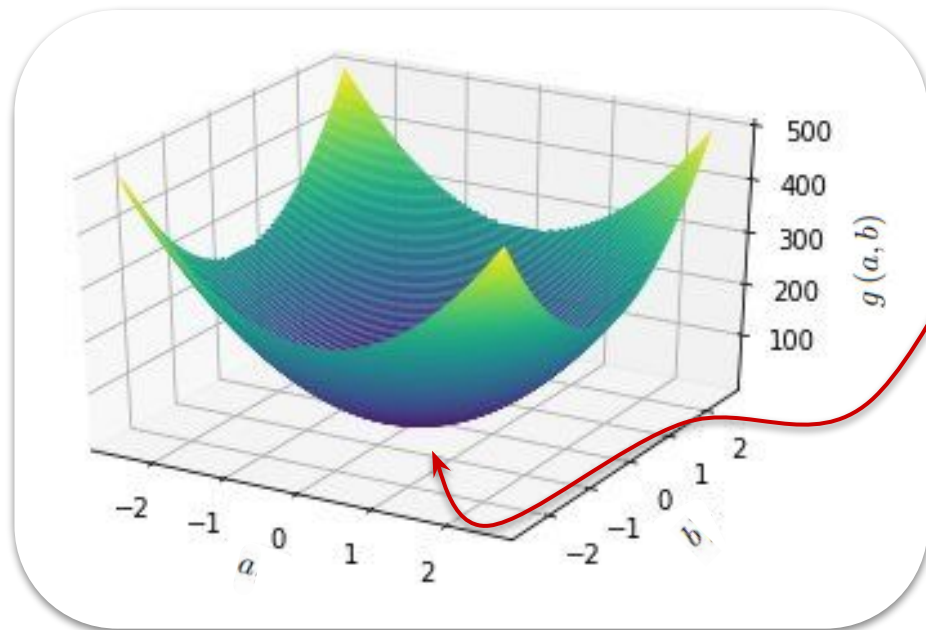


- ✓ Al ser una **función cuadrática** de los parámetros, tiene un mínimo global que además es el **único mínimo local**
- ✓ Lo podemos hallar buscando donde se anula el gradiente:
gradiente $g(a,b) = 0$.

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

Función que mide el desajuste a la recta

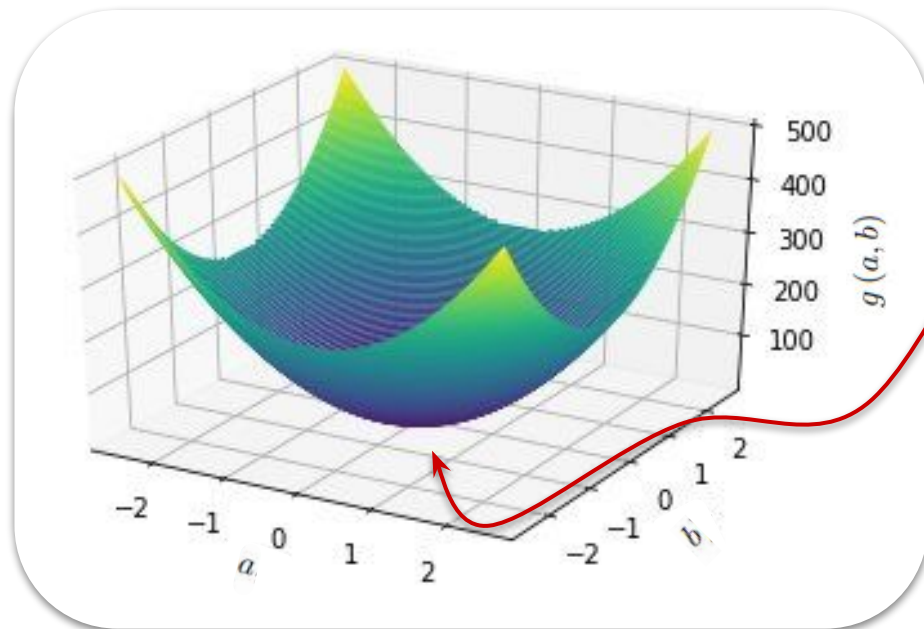
Método de mínimos cuadrados



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:

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

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

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	

Misma ecuación anterior, pero reescrita

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



$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} =$$



Nuestro ejemplo

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	

Misma ecuación anterior, pero reescrita

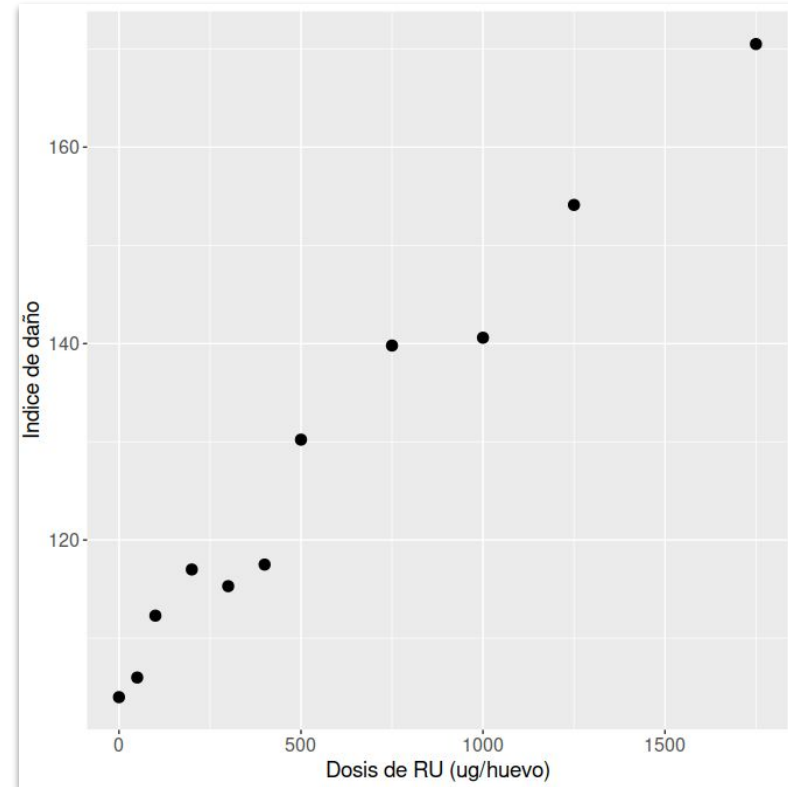
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = 0.037$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \cdot \bar{X} = 106.5$$

La recta ajustada

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

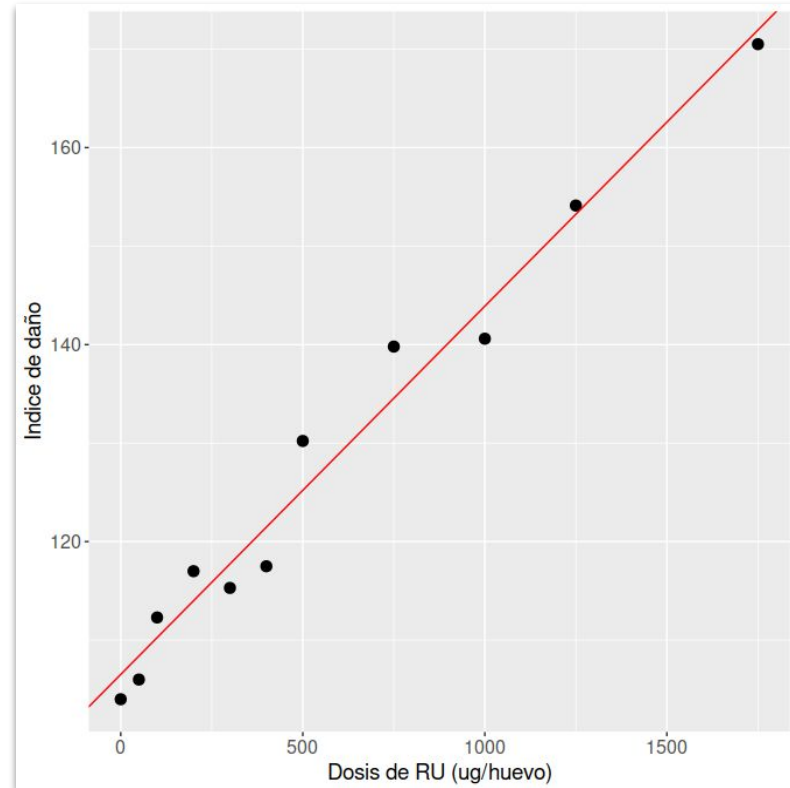
$$\hat{y} = 106.5 + 0.037 * DosisRU$$



La recta ajustada

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

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



Estimación de la recta

```
regresionLinealSimpleRU.py X
1  -*- coding: utf-8 -*-
2  """
3  Materia      : Laboratorio de datos - FCEyN - UBA
4  Clase       : Clase Regresion Lineal
5  Detalle     : Modelo de Regresion Lineal Simple
6  Autores     : Maria Soledad Fernandez y Pablo Turjanski
7  Modificacion: 2023-10-13
8  """
9
10 # Importamos bibliotecas
11 import pandas as pd
12 import numpy as np
13 from sklearn import linear_model
14 import matplotlib.pyplot as plt
15 import seaborn as sns
16
17 #####
18 ##### DEFINICION DE FUNCIONES AUXILIARES
19 #####
```

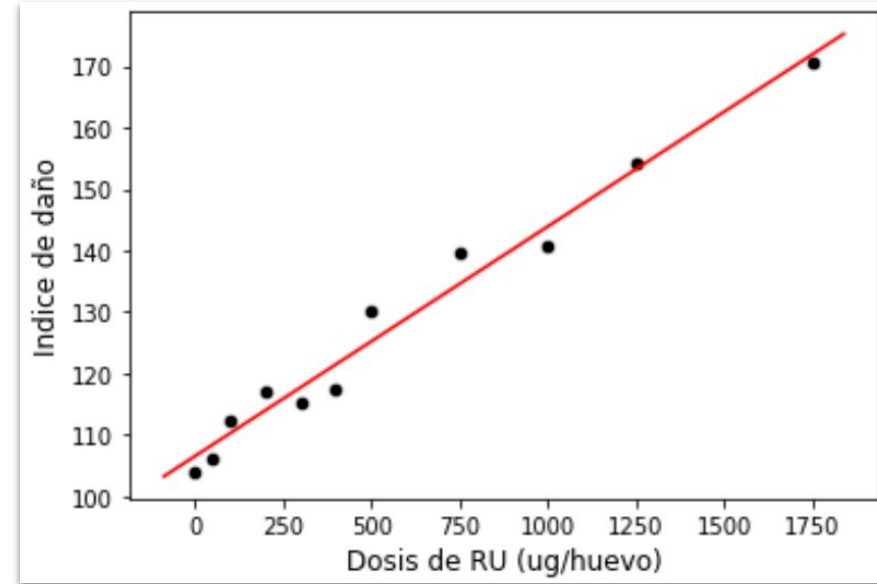
	RU	DI
1	0	104.00
2	50	106.00
3	100	112.30
4	200	117.00
5	300	115.30
6	400	117.50
7	500	130.22
8	750	139.80
9	1000	140.60
10	1250	154.12
11	1750	170.50

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...)

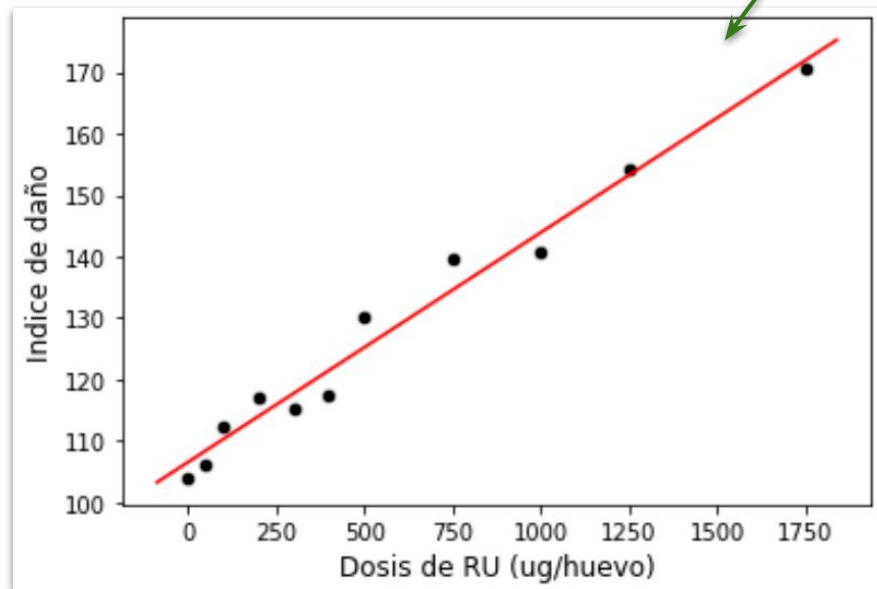


Predicción

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

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

	RU	DI
1	25	
2	600	
3	1500	



Predicción

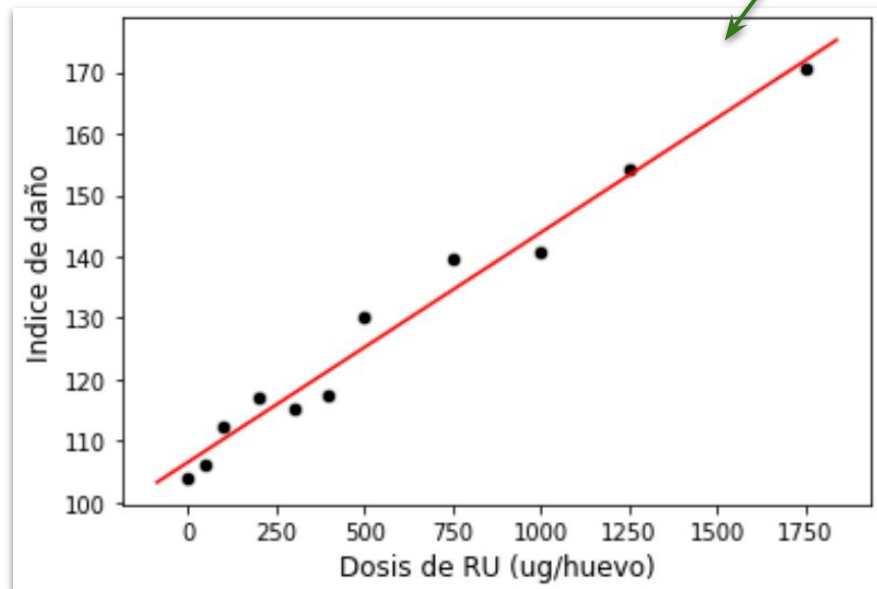
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

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

```
1  # -*- coding: utf-8 -*-
2  """
3  Materia   : Laboratorio de datos - FCEyM - UBA
4  Clase     : Clase Regresión Lineal
5  Detalle   : Modelo de Regresión Lineal Simple
6  Autores   : María Soledad Fernández y Pablo Turfanski
7  Modificación: 2023-10-13
8  """
9
10 # Importamos bibliotecas
11 import pandas as pd
12 import numpy as np
13 from sklearn import linear_model
14 import matplotlib.pyplot as plt
15 import seaborn as sns
16
17 ##### DEFINICION DE FUNCIONES AUXILIARES
18 """
```



	RU	DI
1	25	
2	600	
3	1500	



Predicción

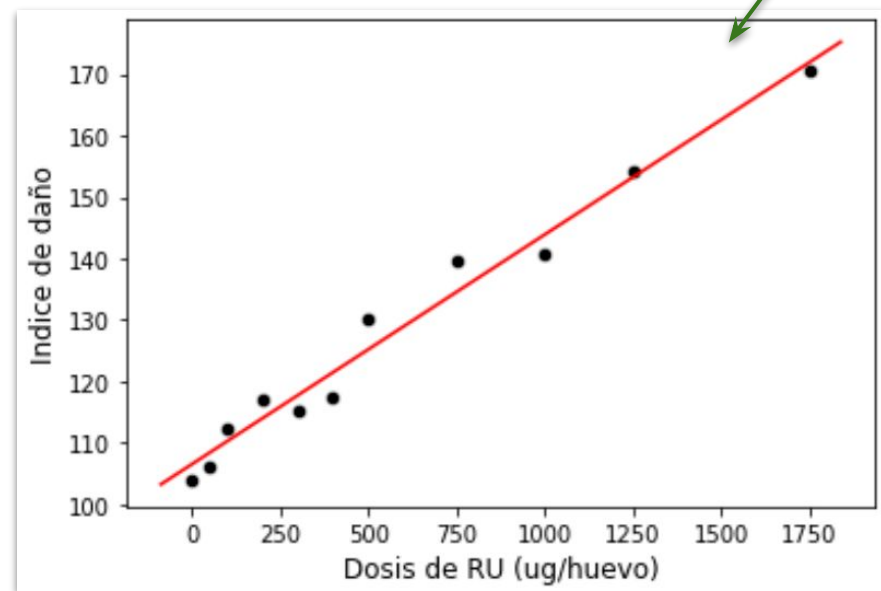
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * DosisRU$$

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

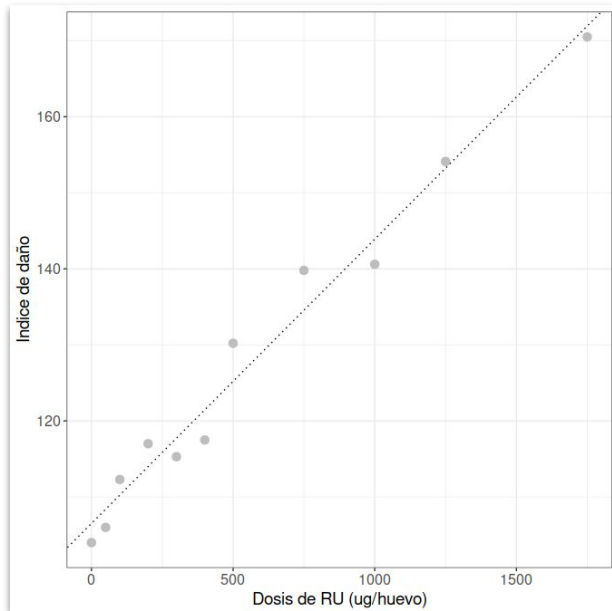
```
1  # -*- coding: utf-8 -*-  
2  """  
3  Materia   : Laboratorio de datos - FCEyM - UBA  
4  Clase    : Clase Regresión Lineal  
5  Detalle  : Modelo de Regresión Lineal Simple  
6  Autores  : María Soledad Fernández y Pablo Turfanski  
7  Modificación: 2023-10-13  
8  """  
9  
10 # Importamos bibliotecas  
11 import pandas as pd  
12 import numpy as np  
13 from sklearn import linear_model  
14 import matplotlib.pyplot as plt  
15 import seaborn as sns  
16  
17  
18 ##### DEFINICION DE FUNCIONES AUXILIARES  
19
```



	RU	DI
1	25	107.46
2	600	128.96
3	1500	162.61

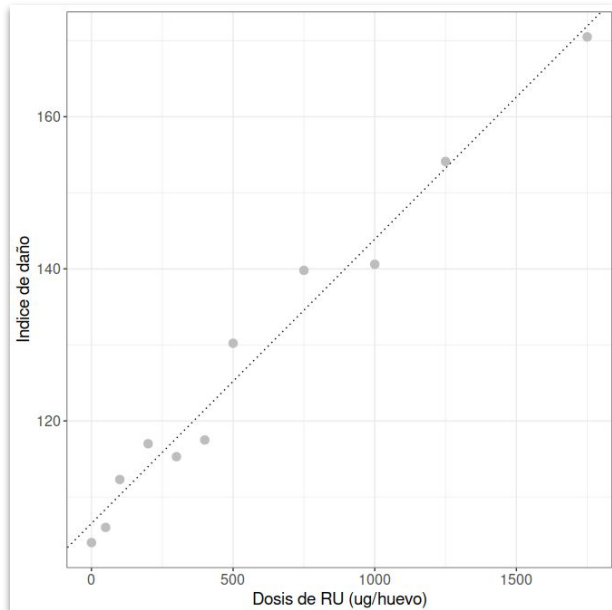


Varianza del modelo

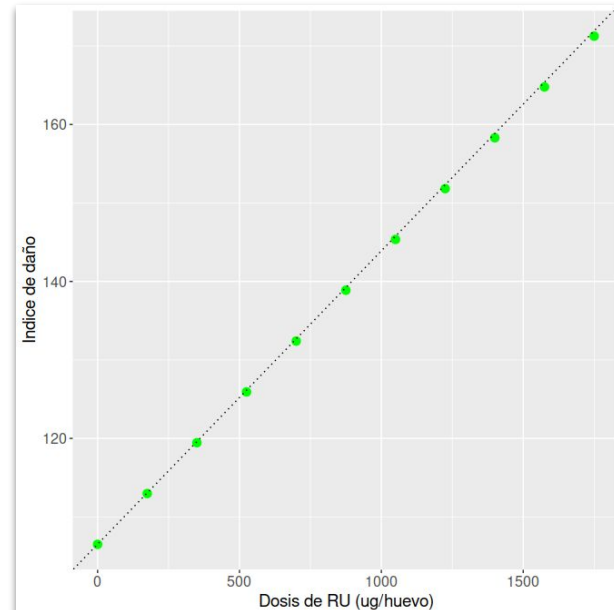
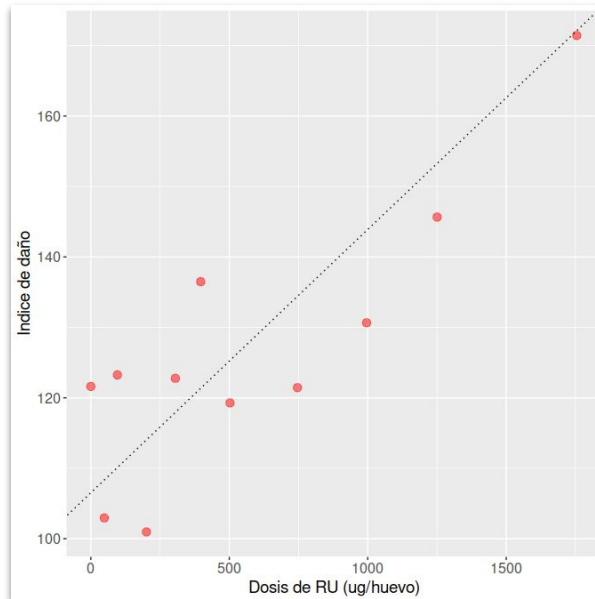


↪ Nuestros datos

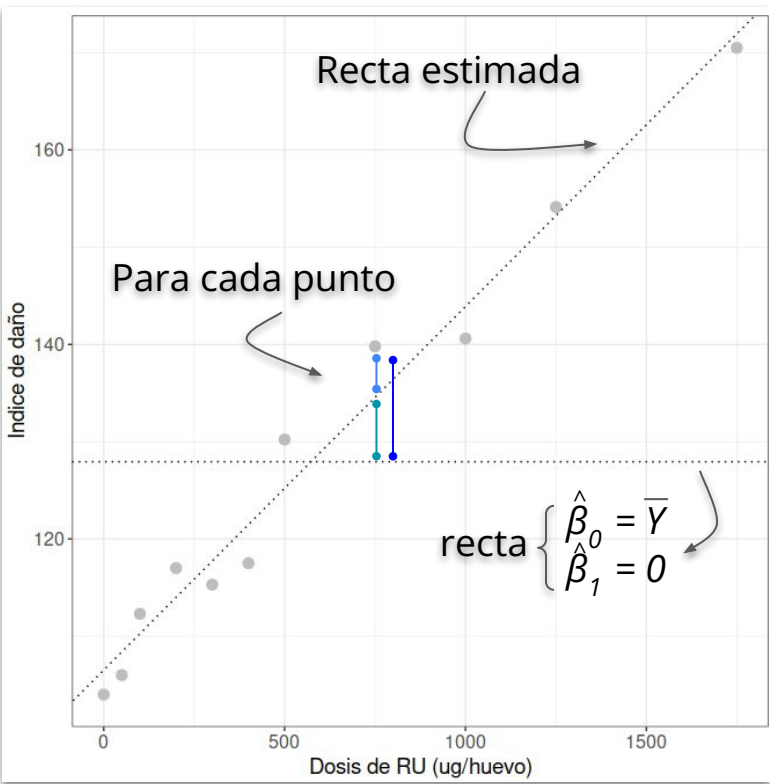
Varianza del modelo



Nuestros datos



Otros escenarios



- ❑ Si no hay relación entre la dosis de RU y el ID entonces $\begin{cases} \hat{\beta}_0 = \bar{Y} \\ \hat{\beta}_1 = 0 \end{cases}$
- ❑ La variabilidad TOTAL del modelo puede separarse en EXPLICADA y NO EXPLICADA

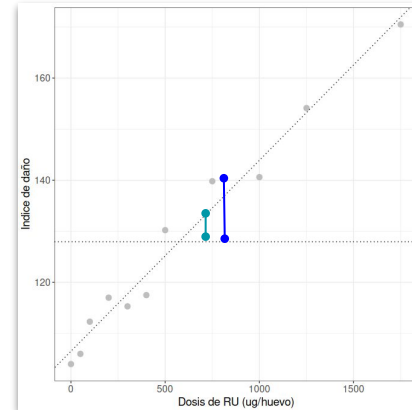
	variabilidad total	$\sum_{i=1}^n (Y_i - \bar{Y})^2$
	variabilidad no explicada	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
	variabilidad explicada	$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- ❑ Coeficiente de determinación (R^2): Mide la proporción de variabilidad de la variable respuesta explicada por variaciones en x, es decir por el modelo de regresión

$$R^2 = \frac{SC_{explic}}{SC_{total}} \quad \text{$$

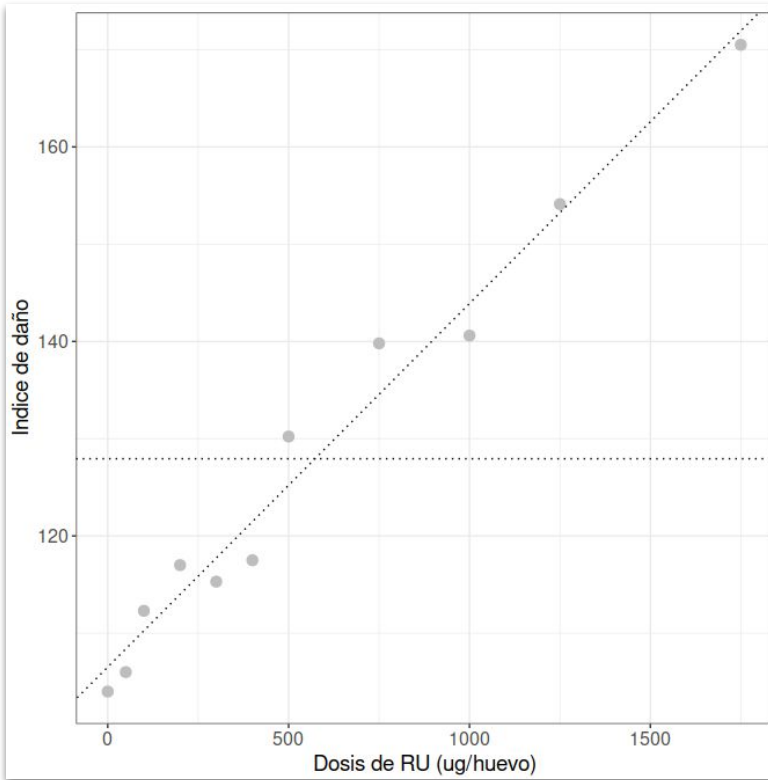
Coeficiente de determinación R^2

- ❑ 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: $0 \leq R^2 \leq 1$
- ❑ A mayor R^2 : más cercanos los puntos a la recta,
- ❑ A mayor R^2 , mayor “fuerza” para predecir (dentro del rango)

$$R^2 = \frac{SC_{explic}}{SC_{total}}$$



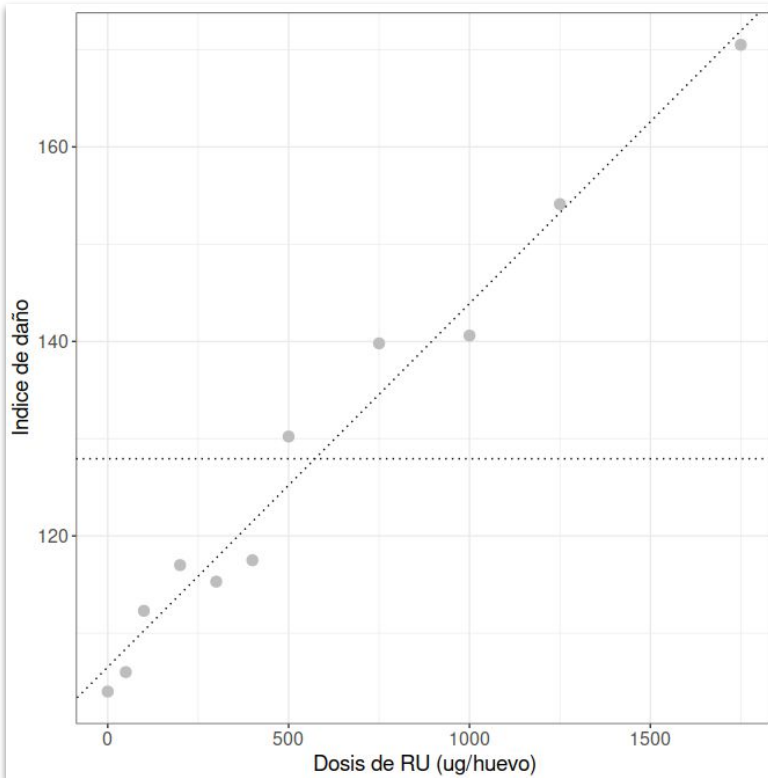
Resumiendo, en nuestro experimento



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 ?



TP - 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 ([link](#)) 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.
7. Construir un histograma con los valores estimados de β_0 y β_1 .
8. Sabiendo que los parámetros poblacionales son $\beta_0=106.5$ y $\beta_1=0.037$ relacionar los conceptos de estimadores y parámetros

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

Ejemplo adaptado de Poletta y cols (2009)





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

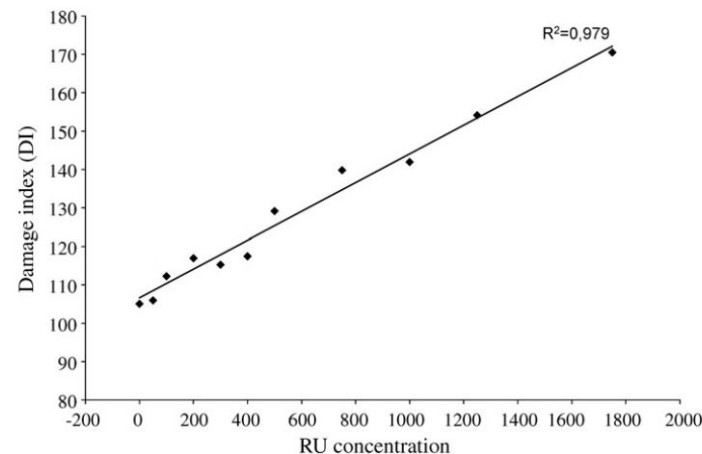


Fig. 3. RU concentration dependent effect for E₁ and E₂ DI mean data. $R^2 = 0.979$, $p < 0.001$.

Algunos comentarios

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

Predicción

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

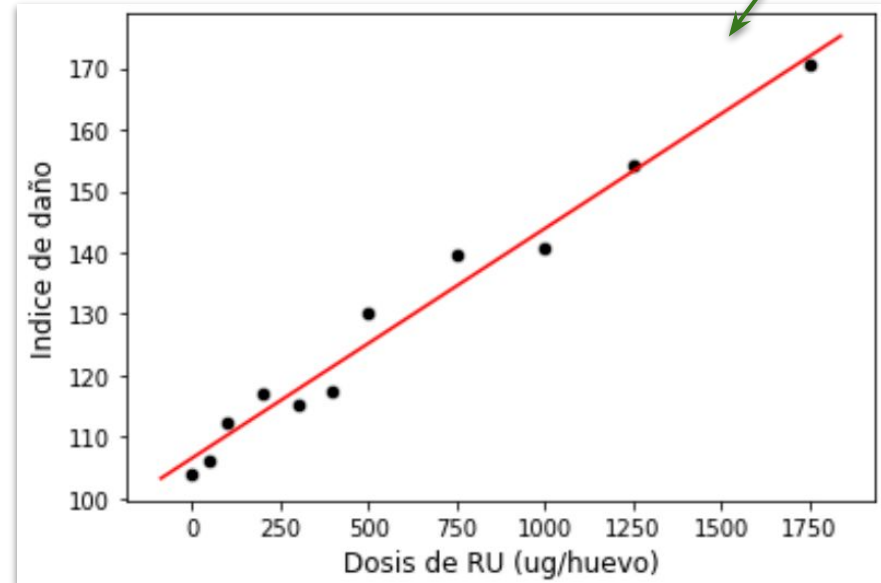
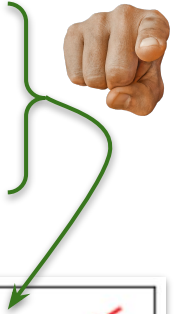
$$\hat{y} = 106.5 + 0.037 * DosisRU$$

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

```
1 # -*- coding: utf-8 -*-
2 """
3 Materia : Laboratorio de datos - FCEyM - UBA
4 Clase : Clase Regresión Lineal
5 Detalle : Modelo de Regresión Lineal Simple
6 Autores : María Soledad Fernández y Pablo Turfanski
7 Modificación: 2023-10-13
8 """
9
10 # Importamos bibliotecas
11 import pandas as pd
12 import numpy as np
13 from sklearn import linear_model
14 import matplotlib.pyplot as plt
15 import seaborn as sns
16
17 ##### DEFINICIÓN DE FUNCIONES AUXILIARES #####
```

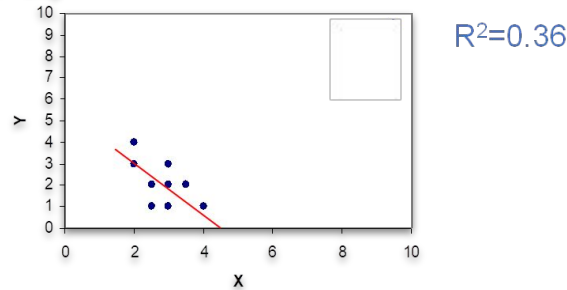
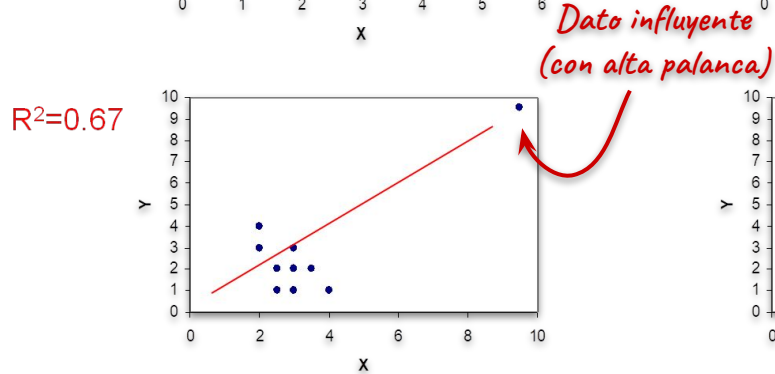
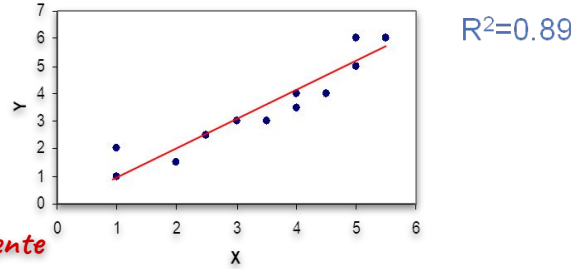
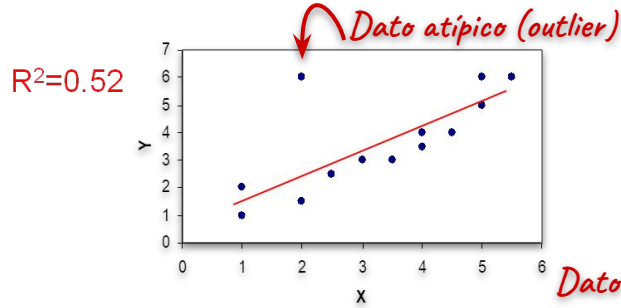


	RU	DI
1	25	✓
2	150000	✗
3	-150000	✗



Observaciones atípica e influyentes

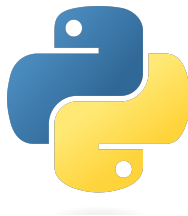
- **Atípicas (outliers en Y):** Observaciones con un patrón distinto al resto de los datos, que producen un residuo grande
- **Influyentes (con alta palanca):** Observaciones cuyo valor de X se encuentra alejado del promedio y que tienen mucho peso en las estimaciones de los parámetros. Al ser eliminadas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

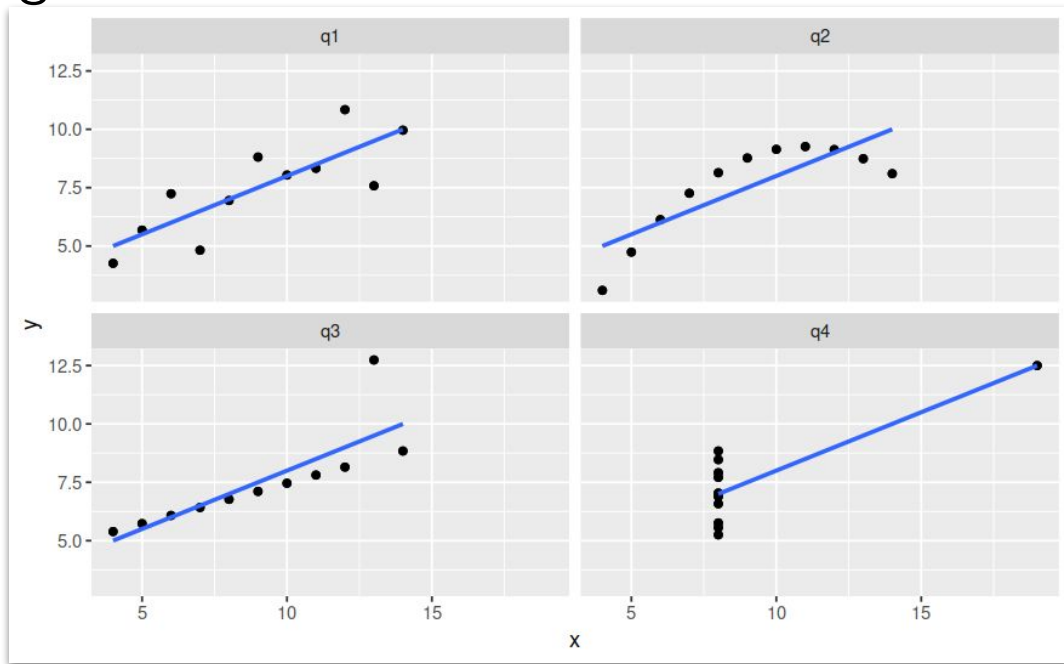
Dataset “Anscombe”

OJO con el análisis exploratorio y el modelado “demasiado a ciegas”

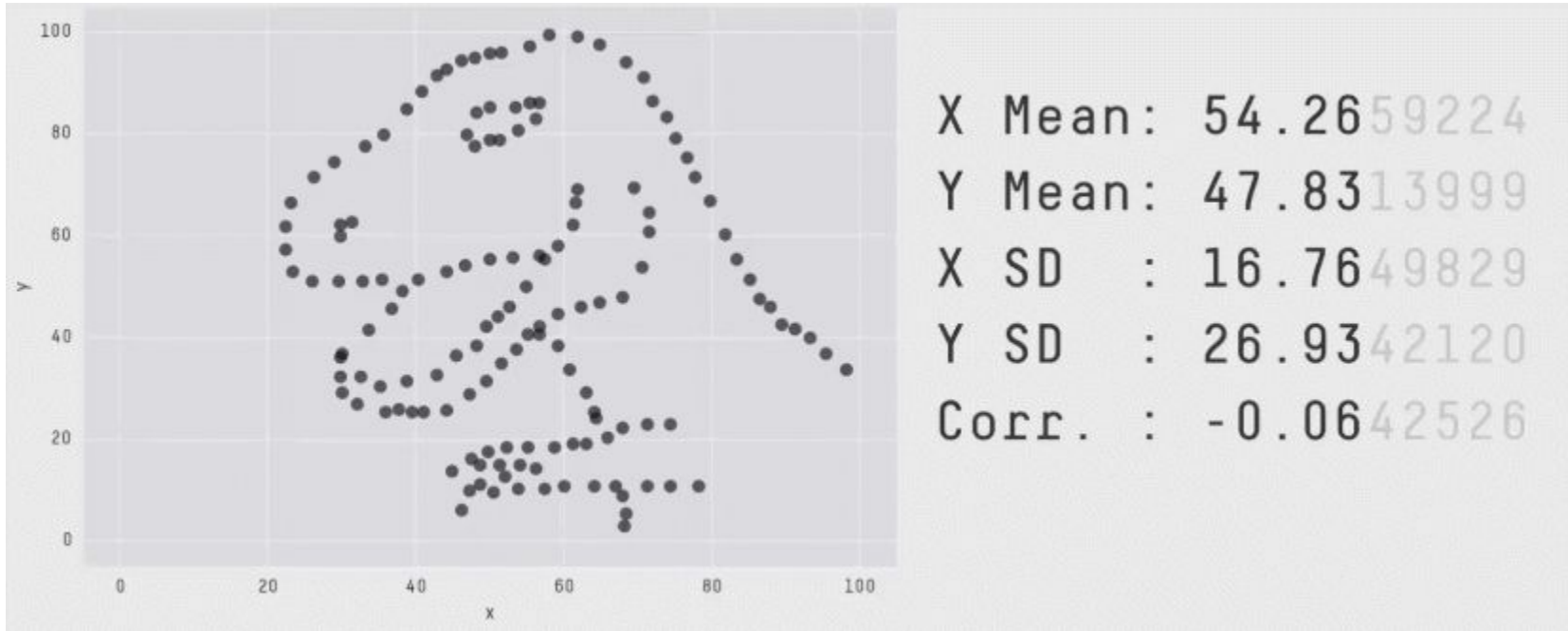
	intercept	pendiente	R2
x1	3.000091	0.5000909	0.6665425
x2	3.000909	0.5000000	0.6662420
x3	3.002455	0.4997273	0.6663240
x4	3.001727	0.4999091	0.6667073

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



Datasaurus Dozen



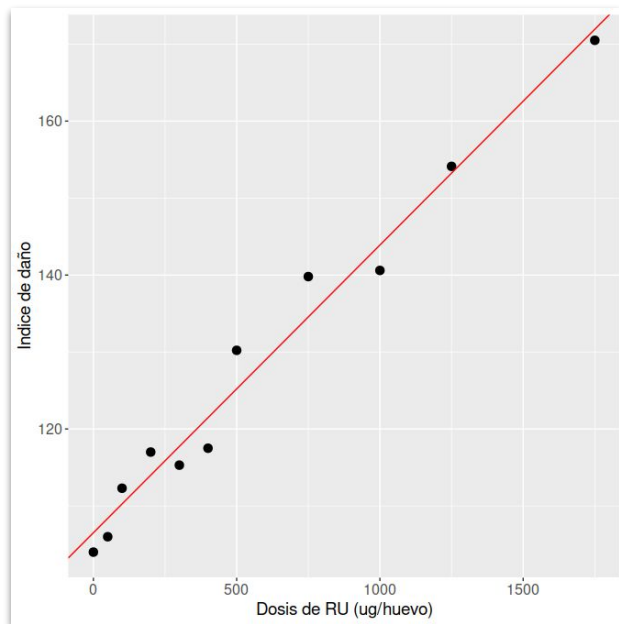
<https://blog.revolutionanalytics.com/2017/05/the-datasaurus-dozen.html>

Modelos lineales

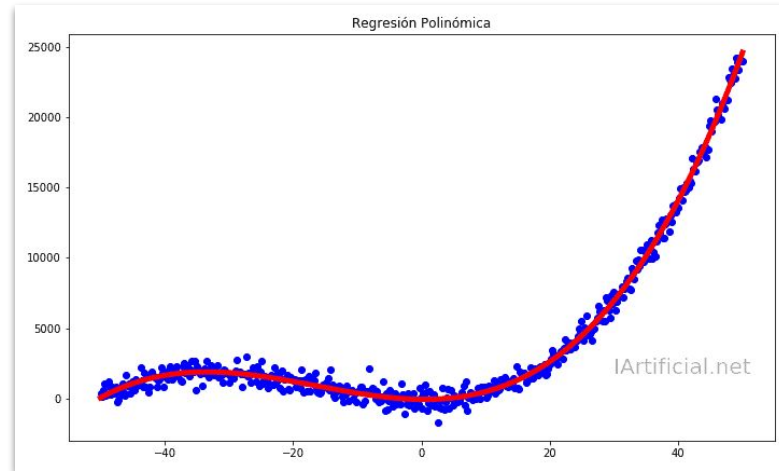
- + Regresión lineal simple (vimos) $\beta_0 + \beta_1 \cdot x$
- + Regresión lineal múltiple $\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$
- + Regresión polinomial $\beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2 + \beta_3 \cdot x^3$



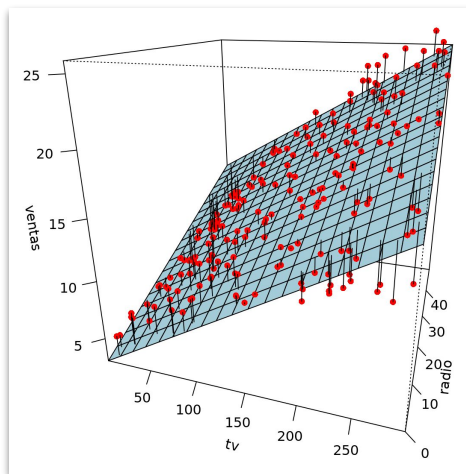
Algunos modelos lineales



$$\beta_0 + \beta_1 \cdot X$$



$$\beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$$



$$\beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$$

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

$$\beta_0 + \beta_1 \cdot x$$

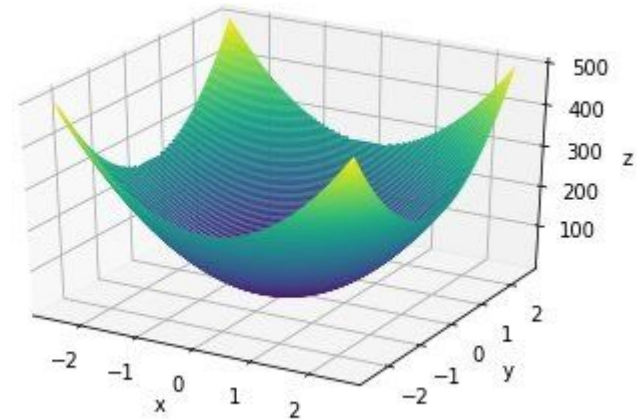
$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

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

¿Cómo medimos cuán bien o mal aproxima nuestra recta?

$$E = \sum r_i^2 = \sum (y_i - (\beta_0 + \beta_1 \cdot x_i))^2$$

Al ser una función cuadrática de los parámetros, tiene un mínimo global que además es el único mínimo local y lo podemos encontrar buscando dónde se anula el gradiente.



Al querer buscar los parámetros que minimicen la suma de los residuos (al cuadrado), obtenemos las ecuaciones que vimos la semana pasada

$$\begin{cases} -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0 \end{cases}$$

que provienen de derivar esa suma respecto de los parámetros.

Entonces los parámetros despejados son:

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

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

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

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

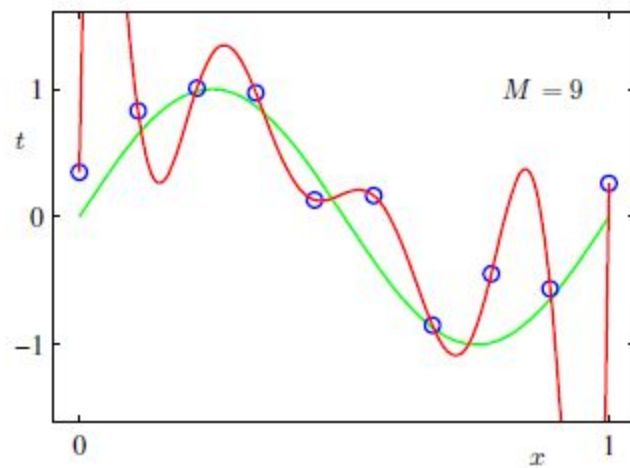
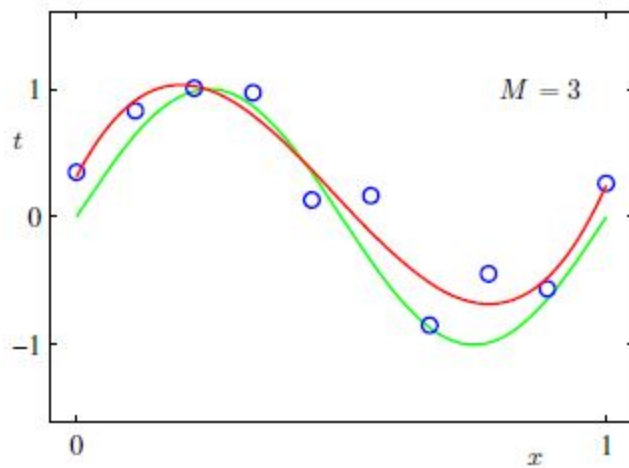
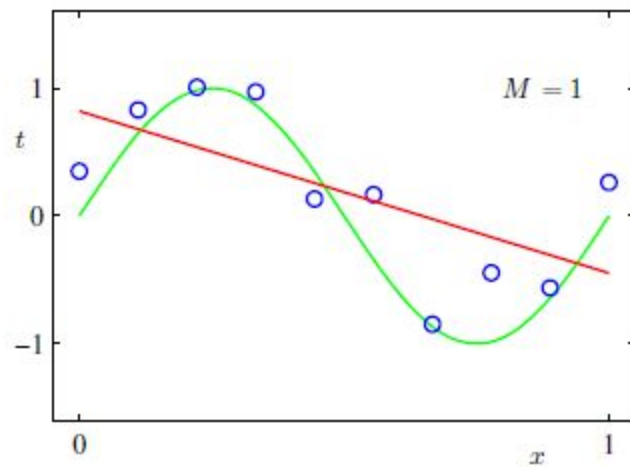
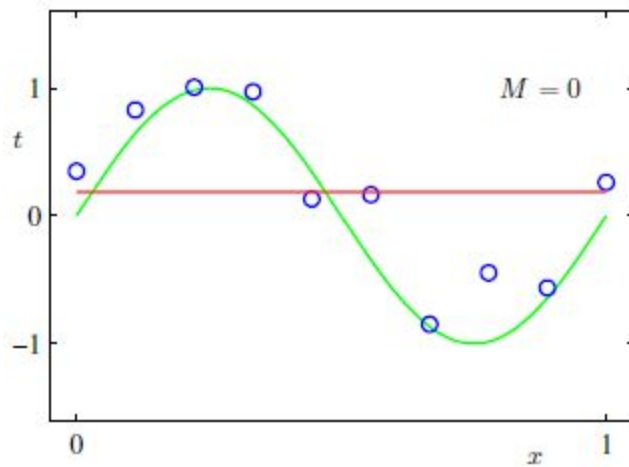
Y_i son los valores observados (reales)

\hat{Y}_i son los valores estimados por el modelo.

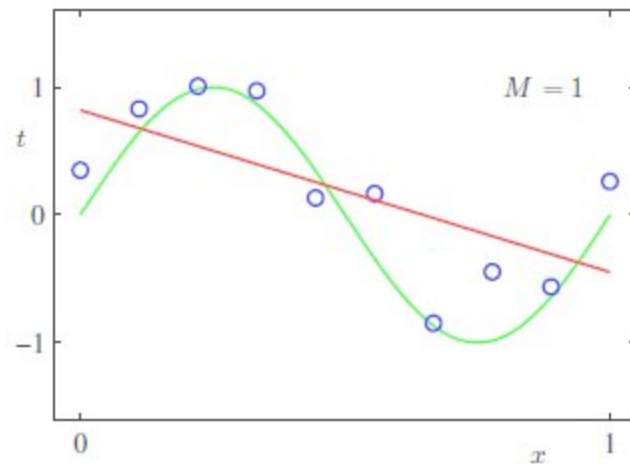
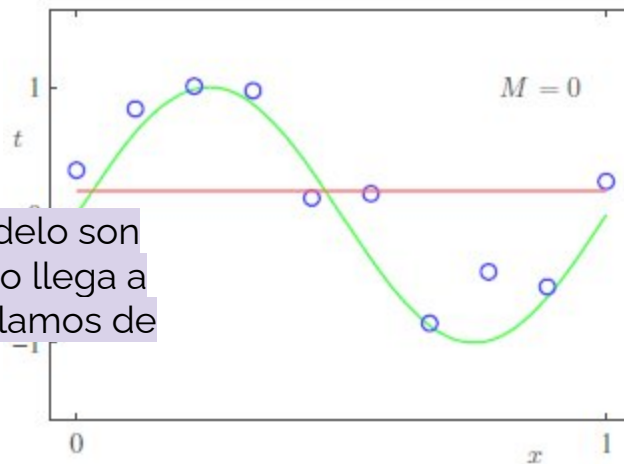
Properati

- + Dataset real de publicaciones digitales de inmuebles
- + Provisto en forma libre por Properati (<https://www.properati.com.ar/>)
- + Trataremos de modelar los precios de las casas de CABA en función de algunas variables relevantes

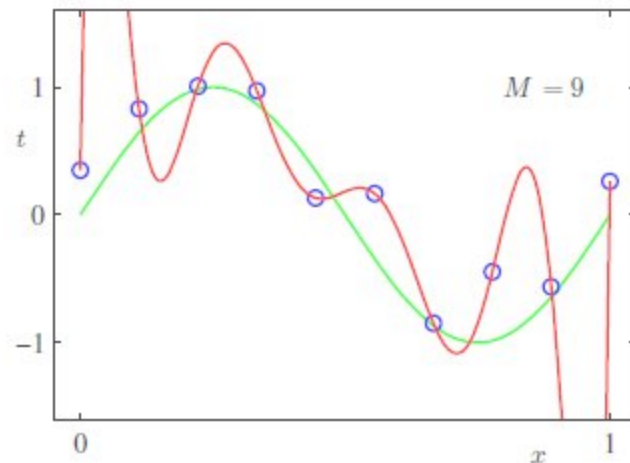
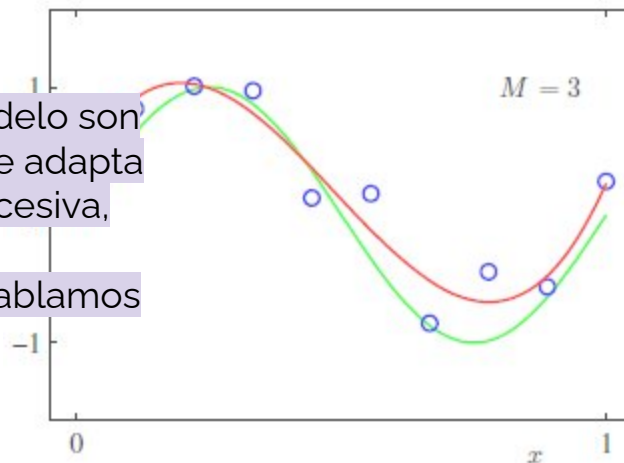
¿Cómo sé si mi modelo realmente **mejora** al agregar parámetros?



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**.



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

Conclusiones

- + Vimos varios modelos en los que se pretende **explicar** o **predecir** una **variable continua** a partir de otras variables.
Modelos paramétricos y no paramétricos
- + 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.