

Regresión Lineal

Introducción

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python

Table of Contents

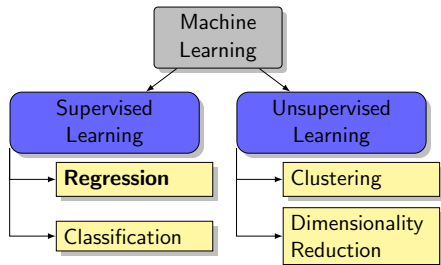
- 1 Introducción
- 2 Regresión Lineal
- 3 Generalizaciones
- 4 Detalles adicionales
 - Linearildad
 - Multicolinealidad entre features

La tarea de la Regresión

Regresión

La **regresión** es una técnica para investigar la *relación entre variables independientes y una variable dependiente* o resultado. Se utiliza como método de modelaje predictivo en el Machine Learning, en el que se emplea un algoritmo para predecir resultados continuos.

Es una de las principales partes del aprendizaje supervisado.



Ejemplos de la tarea de Regresión

Observa las variables independientes (features) y la variable dependiente (etiqueta/salida).

- Predecir el peso de una persona a partir de su altura.

Ejemplos de la tarea de Regresión

Observa las variables independientes (features) y la variable dependiente (etiqueta/salida).

- Predecir el peso de una persona a partir de su altura.
- Predecir la altura de una persona a partir de la longitud del fémur.

Ejemplos de la tarea de Regresión

Observa las variables independientes (features) y la variable dependiente (etiqueta/salida).

- Predecir el peso de una persona a partir de su altura.
- Predecir la altura de una persona a partir de la longitud del fémur.
- Predecir el precio de una casa a partir de su superficie de construcción, habitaciones, ubicación.

Ejemplos de la tarea de Regresión

Observa las variables independientes (features) y la variable dependiente (etiqueta/salida).

- Predecir el peso de una persona a partir de su altura.
- Predecir la altura de una persona a partir de la longitud del fémur.
- Predecir el precio de una casa a partir de su superficie de construcción, habitaciones, ubicación.
- ...

Ejemplo de regresión

Altura	Peso
169.948447	?
173.865754	?
174.661475	?
170.597762	?

Ejemplo de regresión

Altura	Peso
169.948447	82.552060
173.865754	75.674023
174.661475	84.338528
170.597762	87.721204

Ejemplo de regresión

Altura	Peso
169.948447	82.552060
173.865754	75.674023
174.661475	84.338528
170.597762	87.721204

¿Qué otras variables independientes podrían ayudar a la tarea?

Diferentes tipos de regresión

- Regresión Lineal
 - Regresión Lineal Simple
 - Regresión Lineal Multiple
- Regresión Polinomial
- Regresión con regularización: Ridge, Lasso.
- Regresión Logística (Clasificación)
- Quantile Regression
- Support Vector Regression

Table of Contents

1 Introducción

2 Regresión Lineal

3 Generalizaciones

4 Detalles adicionales

- Linearidad
- Multicolinealidad entre features

Regresión Lineal

Regresión Lineal

La **regresión lineal** es un tipo de modelo en el que se supone que la relación entre una variable independiente y una variable dependiente es lineal.

Existen dos tipos de Modelo de Regresión Lineal:

- **Regresión lineal simple:** Un modelo de regresión lineal con una variable independiente y una dependiente.
- **Regresión lineal múltiple:** Un modelo de regresión lineal con más de una variable independiente y una variable dependiente.

Variables

En un modelo de regresión lineal hay dos tipos de variables:

- La **variable de entrada** o predictora es la variable o variables que ayudan a predecir el valor de la variable de salida. Se suele denominar X .
- La **variable de salida** es la variable que queremos predecir. Se suele denominar y .

Variables

En un modelo de regresión lineal hay dos tipos de variables:

- La **variable de entrada** o predictora es la variable o variables que ayudan a predecir el valor de la variable de salida. Se suele denominar X .
- La **variable de salida** es la variable que queremos predecir. Se suele denominar y .

Para estimar y a partir de X usamos la ecuación (modelo)

$$y = \beta_0 + \beta_1 X$$

Variables

En un modelo de regresión lineal hay dos tipos de variables:

- La **variable de entrada** o predictora es la variable o variables que ayudan a predecir el valor de la variable de salida. Se suele denominar X .
- La **variable de salida** es la variable que queremos predecir. Se suele denominar y .

Para estimar y a partir de X usamos la ecuación (modelo)

$$y = \beta_0 + \beta_1 X$$

Math Disclaimer...

Regresión Lineal Simple

Queremos ajustar una línea $y = \beta_0 + \beta_1 x$ a los datos

x_1	y_1
x_2	y_2
\dots	\dots
x_N	y_N

Regresión Lineal Simple

Queremos ajustar una línea $y = \beta_0 + \beta_1 x$ a los datos

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_N & y_N \end{array}$$

β_0 es la ordenada al origen (**intercepto**) y β_1 es la pendiente de la línea (**coeficiente**).

Regresión Lineal Simple

Queremos ajustar una línea $y = \beta_0 + \beta_1 x$ a los datos

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_N & y_N \end{array}$$

β_0 es la ordenada al origen (**intercepto**) y β_1 es la pendiente de la línea (**coeficiente**).

Es decir,

$$\begin{array}{rcl} y_1 & = & \beta_0 + \beta_1 x_1 \\ \dots & & \dots \\ y_N & = & \beta_0 + \beta_1 x_N \end{array}$$

Regresión Lineal Simple

Queremos ajustar una línea $y = \beta_0 + \beta_1 x$ a los datos

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_N & y_N \end{array}$$

β_0 es la ordenada al origen (**intercepto**) y β_1 es la pendiente de la línea (**coeficiente**).

Es decir,

$$\begin{array}{rcl} y_1 & = & \beta_0 + \beta_1 x_1 \\ \dots & & \dots \\ y_N & = & \beta_0 + \beta_1 x_N \end{array}$$

- Si tuviéramos dos puntos, la solución se calcula *fácil*.

Regresión Lineal Simple

Queremos ajustar una línea $y = \beta_0 + \beta_1 x$ a los datos

$$\begin{array}{cc} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_N & y_N \end{array}$$

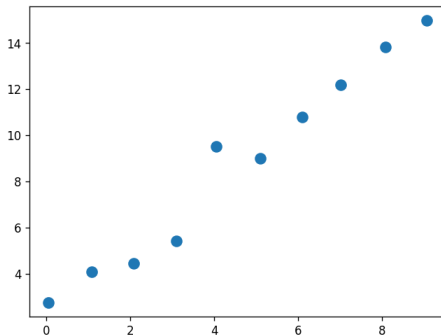
β_0 es la ordenada al origen (**intercepto**) y β_1 es la pendiente de la línea (**coeficiente**).

Es decir,

$$\begin{array}{rcl} y_1 & = & \beta_0 + \beta_1 x_1 \\ \dots & & \dots \\ y_N & = & \beta_0 + \beta_1 x_N \end{array}$$

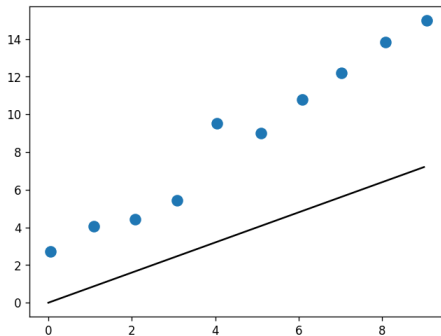
- Si tuviéramos dos puntos, la solución se calcula *fácil*.
- Si tenemos más de dos puntos, escogemos la *mejor* línea.

¿Cómo sabemos cuál es la mejor línea?



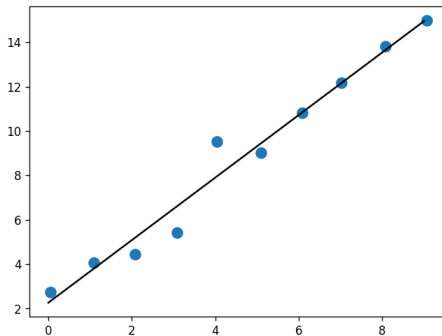
<https://www.geogebra.org/m/maeexqmr>

¿Cómo sabemos cuál es la mejor línea?



<https://www.geogebra.org/m/maeexqmr>

¿Cómo sabemos cuál es la mejor línea?



<https://www.geogebra.org/m/maeexqmr>

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = y_1 - (\beta_0 + \beta_1 x_1)$$

...

$$e_N = y_N - (\beta_0 + \beta_1 x_N)$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = |y_1 - (\beta_0 + \beta_1 x_1)|$$

...

$$e_N = |y_N - (\beta_0 + \beta_1 x_N)|$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = (y_1 - (\beta_0 + \beta_1 x_1))^2$$

...

$$e_N = (y_N - (\beta_0 + \beta_1 x_N))^2$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = (y_1 - (\beta_0 + \beta_1 x_1))^2$$

...

$$e_N = (y_N - (\beta_0 + \beta_1 x_N))^2$$

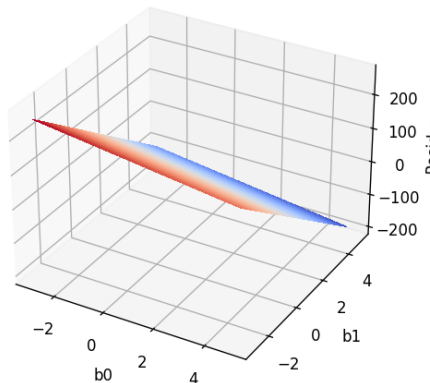
Consideramos la función que suma todos los residuos

$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2$$

Minimizando el error

Considerando la función

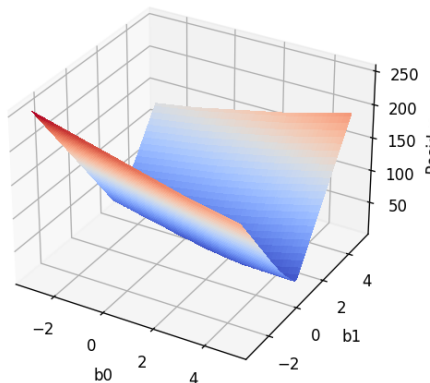
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2$$



Minimizando el error

Considerando la función

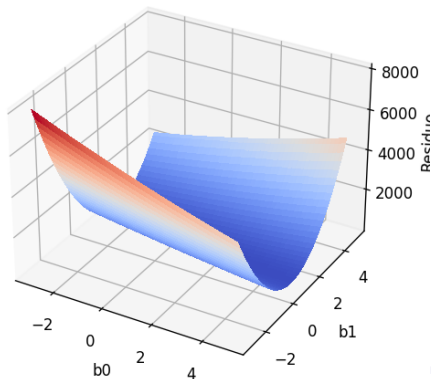
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^N |y_i - (\beta_0 + \beta_1 x_i)|$$



Minimizando el error

Considerando la función

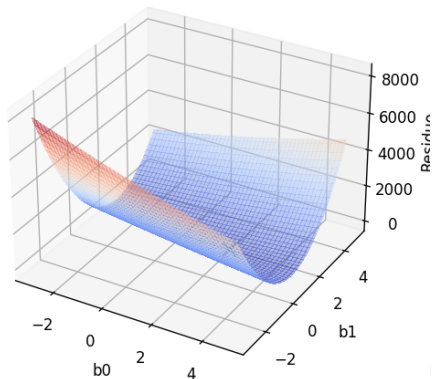
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2$$



Minimizando el error

Considerando la función

$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^N (y_i - (\beta_0 + \beta_1 x_i))^2$$



La solución: Formulación vectorial

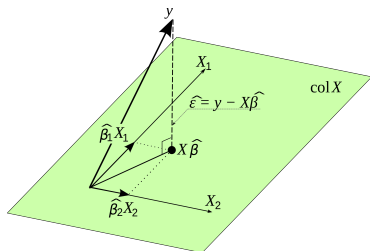
Considerando los residuos tenemos

$$\begin{aligned}y_1 &= \beta_0 + \beta_1 x_1 + e_1 \\y_2 &= \beta_0 + \beta_1 x_2 + e_2 \\&\dots \\y_N &= \beta_0 + \beta_1 x_N + e_N\end{aligned}$$

Lo podemos reescribir como

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

La solución: Formulación vectorial



$$\mathbf{e}^T \mathbf{X} = 0 \in \mathcal{M}_{1 \times 2}$$

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \mathbf{X} = 0$$

$$\mathbf{X}^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 0$$

$$\mathbf{X}^T \mathbf{y} - \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}} = 0$$

$$\mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \hat{\boldsymbol{\beta}}$$

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \hat{\boldsymbol{\beta}}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

<https://www.geogebra.org/m/g7uxxvkt>

Regresión Lineal Multiple

En el caso general, tenemos m variables independientes (features) x^1, \dots, x^m .

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \dots & x_1^m \\ 1 & x_2^1 & \dots & x_2^m \\ \dots & \dots & \dots & \dots \\ 1 & x_N^1 & \dots & x_N^m \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_m \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$$

Y la solución OLS es, otra vez,

$$\hat{\beta} = (X^T X)^{-1} X^T \mathbf{y}$$

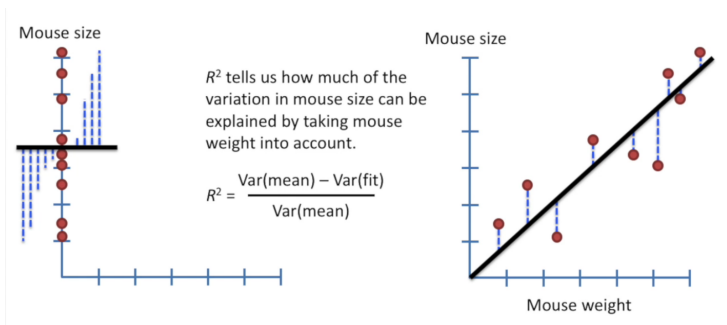
Interpretación de los coeficientes

- El signo de un coeficiente indica si existe una correlación positiva o negativa entre cada variable independiente y la variable dependiente.
- La magnitud del coeficiente indica cuánto cambia la media de la variable dependiente si se produce un cambio de una unidad en la variable independiente y se mantienen constantes las demás variables del modelo. Esto permite evaluar el efecto de cada variable de forma aislada de las demás.

Estos coeficientes son estimaciones de los *verdaderos* (los coeficientes de toda la población).

El coeficiente R^2

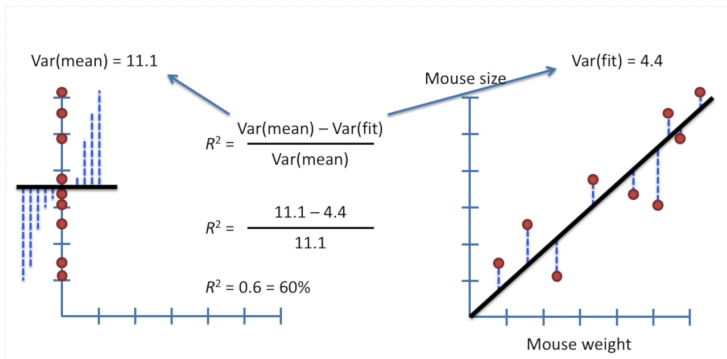
Una pregunta relevante se refiere a la incertidumbre acerca de los parámetros $\hat{\beta}$, ya que estos son variables aleatorias. El coeficiente R^2 explica la varianza de los datos que es explicada por el efecto de las variables dependientes.



https://www.youtube.com/watch?v=nk2CQITm_eo

El coeficiente R^2

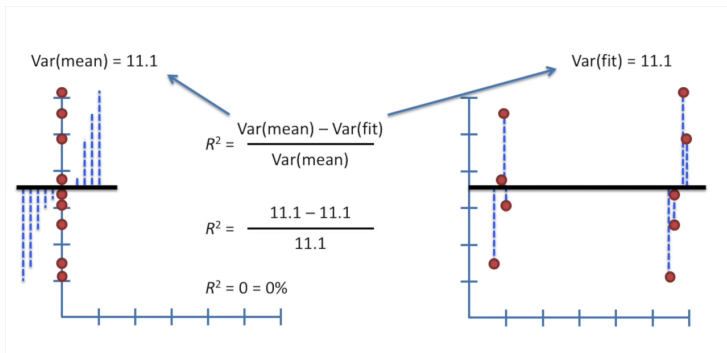
Una pregunta relevante se refiere a la incertidumbre acerca de los parámetros $\hat{\beta}$, ya que estos son variables aleatorias. El coeficiente R^2 explica la varianza de los datos que es explicada por el efecto de las variables dependientes.



https://www.youtube.com/watch?v=nk2CQITm_eo

El coeficiente R^2

Una pregunta relevante se refiere a la incertidumbre acerca de los parámetros $\hat{\beta}$, ya que estos son variables aleatorias. El coeficiente R^2 explica la varianza de los datos que es explicada por el efecto de las variables dependientes.



https://www.youtube.com/watch?v=nk2CQITm_eo

Resumiendo

- La regresión lineal descubre la relación lineal entre varias variables predictoras y (una) variable dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.

Resumiendo

- La regresión lineal descubre la relación lineal entre varias variables predictoras y (una) variable dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.
- Con estos coeficientes podemos realizar predicciones de nuevos valores.

Resumiendo

- La regresión lineal descubre la relación lineal entre varias variables predictoras y (una) variable dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.
- Con estos coeficientes podemos realizar predicciones de nuevos valores.
- Los coeficientes cuantifican la relación de cada variable con la variable de salida.

Resumiendo

- La regresión lineal descubre la relación lineal entre varias variables predictoras y (una) variable dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.
- Con estos coeficientes podemos realizar predicciones de nuevos valores.
- Los coeficientes cuantifican la relación de cada variable con la variable de salida.
- El coeficiente R^2 cuantifica qué tanto el modelo explica la varianza de los datos.

Table of Contents

- 1 Introducción
- 2 Regresión Lineal
- 3 Generalizaciones**
- 4 Detalles adicionales
 - Linearildad
 - Multicolinealidad entre features

Regresión Polinomial

En la regresión lineal queremos ajustar un modelo

$$y = \beta_0 + \beta_1 x$$

a los datos

$$\begin{array}{c|c} x_1 & y_1 \\ x_2 & y_2 \\ \dots & \dots \\ x_N & y_N \end{array}$$

esto se traduce en el sistema

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_N \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

Regresión Polinomial

Ahora, queremos ajustar un polinomio de grado 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

Regresión Polinomial

Ahora, queremos ajustar un polinomio de grado 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

el enfoque es considerar a x^2 como una nueva variable y no tanto como el cuadrado de la primer variable.

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_N & x_N^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

Regresión Polinomial

Ahora, queremos ajustar un polinomio de grado 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

el enfoque es considerar a x^2 como una nueva variable y no tanto como el cuadrado de la primer variable.

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_N & x_N^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

Sigue siendo un problema lineal en los coeficientes β_j .

Regresión Polinomial

Ahora, queremos ajustar un polinomio de grado 2

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

el enfoque es considerar a x^2 como una nueva variable y no tanto como el cuadrado de la primer variable.

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_N & x_N^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \dots \\ e_N \end{pmatrix}$$

Sigue siendo un problema lineal en los coeficientes β_j . Es necesario, entonces, generar la nueva columna de datos x_j^2 antes de realizar la regresión lineal.

Regresión Polinomial Multiple

Si tenemos varias variables independientes

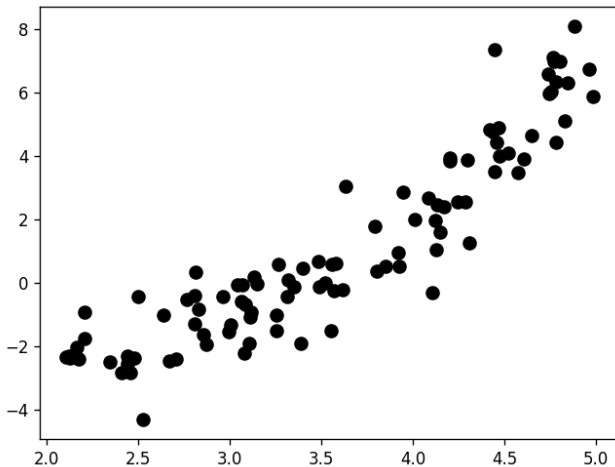
$$\begin{array}{c|c|c} x_1^{(1)} & x_1^{(2)} & y_1 \\ x_2^{(1)} & x_2^{(2)} & y_2 \\ \dots & \dots & \dots \\ x_N^{(1)} & x_N^{(2)} & y_N \end{array}$$

Es necesario generar nuevas columnas para incluir los datos de grado 2:

$$\begin{array}{c|c|c|c|c|c} x_1^{(1)} & x_1^{(2)} & x_1^{(1)2} & x_1^{(2)2} & x_1^{(1)} x_1^{(2)} & y_1 \\ x_2^{(1)} & x_2^{(2)} & x_2^{(1)2} & x_2^{(2)2} & x_2^{(1)} x_2^{(2)} & y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_N^{(1)} & x_N^{(2)} & x_N^{(1)2} & x_N^{(2)2} & x_N^{(1)} x_N^{(2)} & y_N \end{array}$$

Regresión Polinomial: Ejemplo

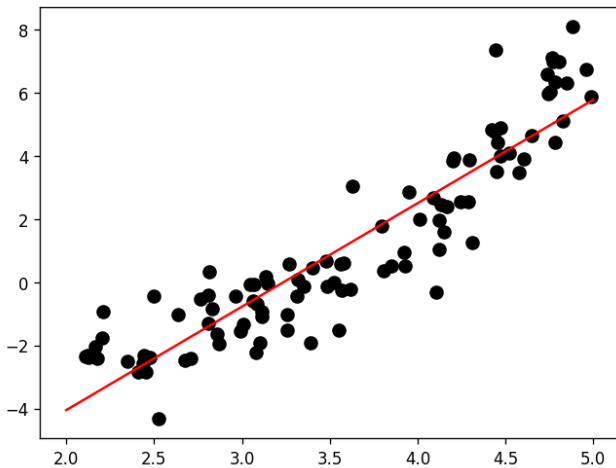
Los datos



Regresión Polinomial: Ejemplo

Regresión lineal

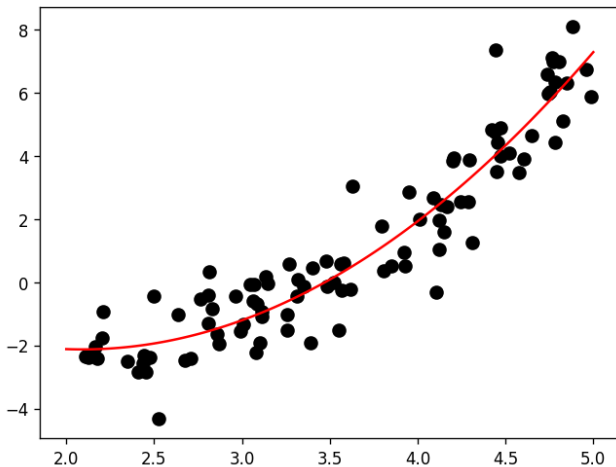
r^2 score=0.8375



Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 2

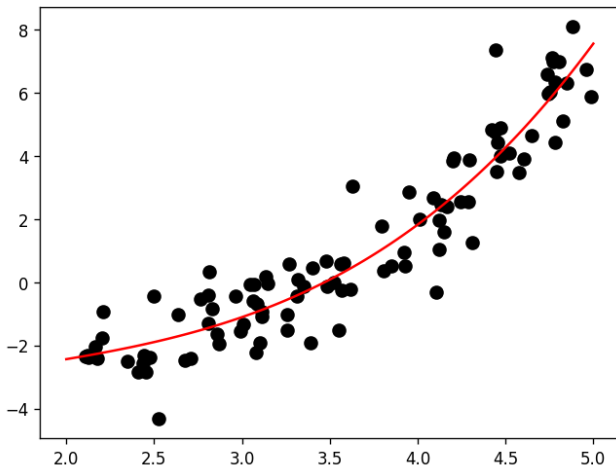
r^2 score=0.8918



Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 3

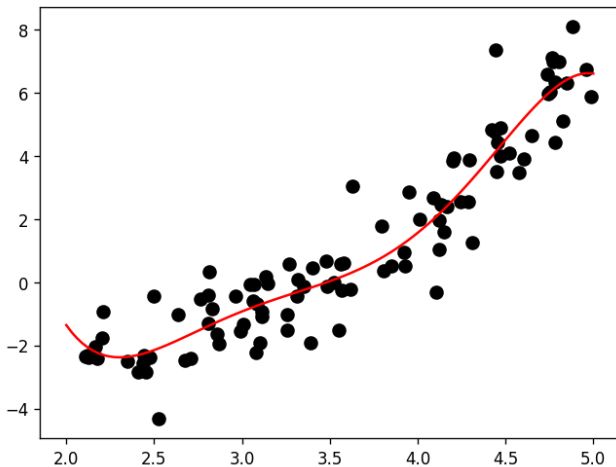
r^2 score=0.8928



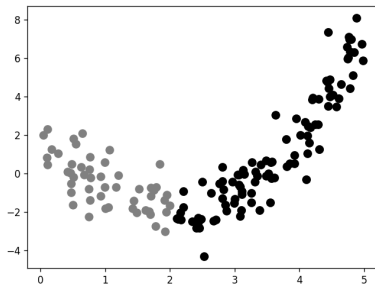
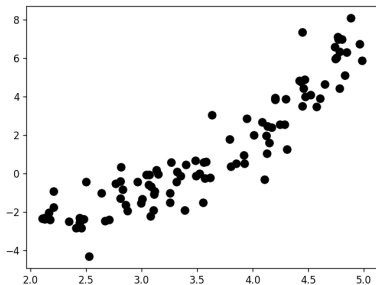
Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 5

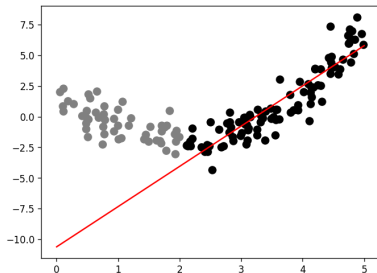
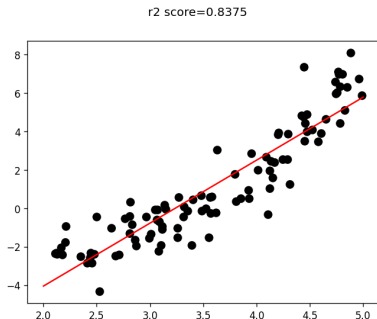
r^2 score=0.8976



Regresión Polinomial: Ejemplo con nuevos datos

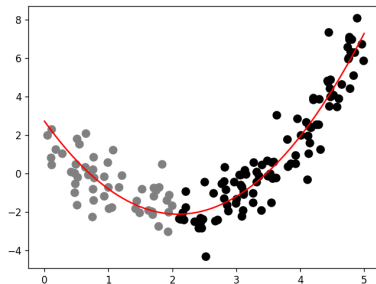
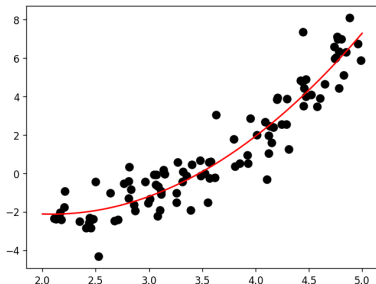


Regresión Polinomial: Ejemplo con nuevos datos



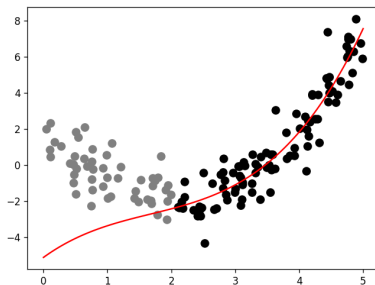
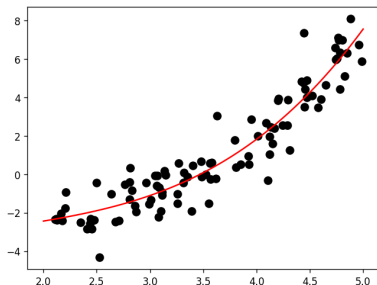
Regresión Polinomial: Ejemplo con nuevos datos

r2 score=0.8918

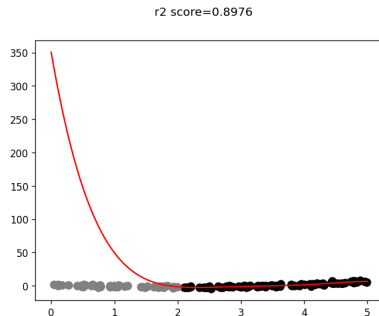
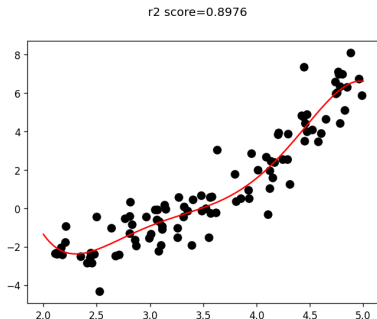


Regresión Polinomial: Ejemplo con nuevos datos

r2 score=0.8928



Regresión Polinomial: Ejemplo con nuevos datos



Conclusión

No es bueno usar un modelo más sencillo o más complejo de lo necesario ya que no son capaces de generalizar nuevos datos.

Conclusión

No es bueno usar un modelo más sencillo o más complejo de lo necesario ya que no son capaces de generalizar nuevos datos.

Al fenómeno de tener un rendimiento muy bueno en los datos de entrenamiento y un rendimiento considerablemente peor en los datos nuevos de prueba se le llama **overfitting**.

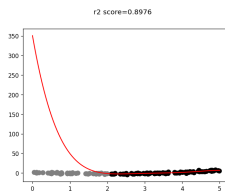
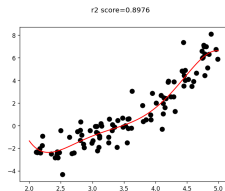
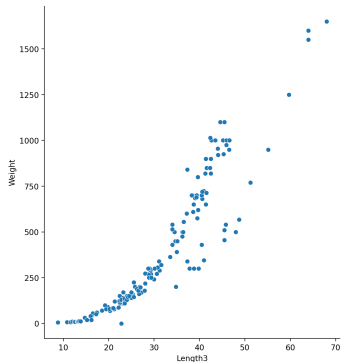
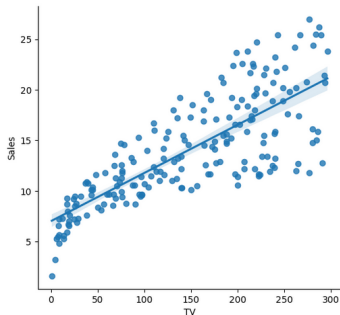


Table of Contents

- 1 Introducción
- 2 Regresión Lineal
- 3 Generalizaciones
- 4 Detalles adicionales
 - Linearidad
 - Multicolinealidad entre features

Linealidad entre variables independientes y dependientes

Ejemplos de relación *lineal* clara:



Linealidad entre variables independientes y dependientes

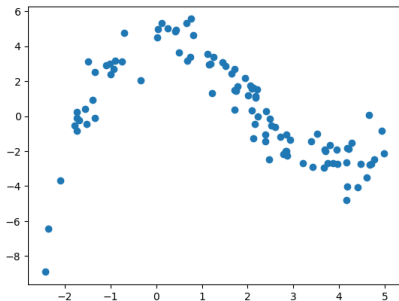
Ejemplos de relación *lineal* clara:

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

	Weight	Length1	Length2	Length3	Height	Width
Weight	1.000000	0.915712	0.918618	0.923044	0.724345	0.886507
Length1	0.915712	1.000000	0.999517	0.992031	0.625378	0.867050
Length2	0.918618	0.999517	1.000000	0.994103	0.640441	0.873547
Length3	0.923044	0.992031	0.994103	1.000000	0.703409	0.878520
Height	0.724345	0.625378	0.640441	0.703409	1.000000	0.792881
Width	0.886507	0.867050	0.873547	0.878520	0.792881	1.000000

Linealidad entre variables independientes y dependientes

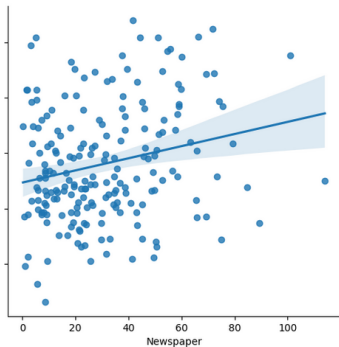
Cuidado con la correlación como herramienta para cuantificar una relación *lineal* entre las variables independientes y dependientes.



	x	y
x	1.000000	-0.426465
y	-0.426465	1.000000

Linealidad entre variables independientes y dependientes

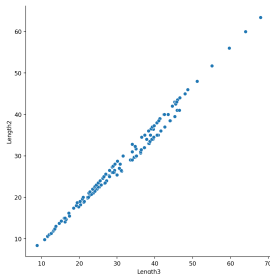
Ejemplo de **no** relación lineal:



	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

Multicolinealidad entre features

- La multicolinealidad perfecta ocurre cuando hay variables predictivas con una relación lineal exacta. En este caso, la matriz $X^T X$ no se puede invertir.
- La multicolinealidad imperfecta se refiere a una situación en la que las variables predictivas tienen una relación lineal casi exacta.



Un caso donde ocurre la multicolinealidad

Al generar features con la codificación *one-hot* se pueden generar features colineales. Por ejemplo, si tenemos la variable `gender`, que puede tener dos valores (`female` y `male`)

gender
F
M
F

female	male
1	0
0	1
0	1

female
1
0
0

Las variables `female` y `male` son colineales ya que

$$\text{female} = 1 - \text{male}$$