Regresión Lineal

Introducción

Dr. Mauricio Toledo-Acosta

Diplomado Ciencia de Datos con Python

Table of Contents

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.

Diferentes tipos de regresión

- Linear Regression
  - Simple Linear Regression
  - Multiple Linear Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Ridge Regression, Lasso Regression
- Support Vector Regression

Table of Contents

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.

Para estimar y a partir de X usamos la ecuación

$$y = \alpha + \beta X$$

A toy example

Regresión Lineal Simple

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

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

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.

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.

Es decir,

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

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.

Es decir,

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

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

Regresión Lineal Simple

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

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

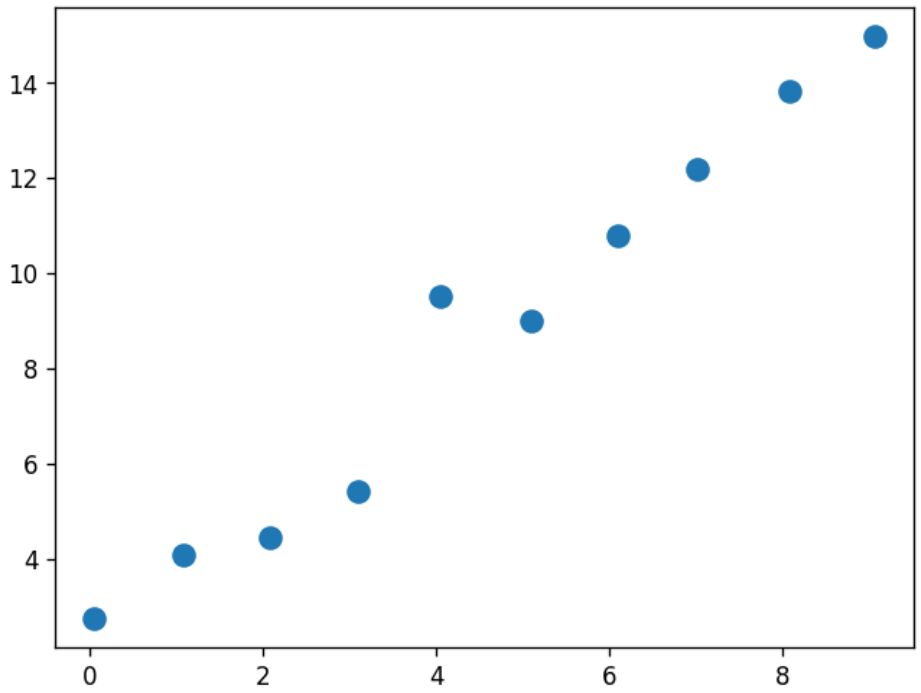
β_0 es la ordenada al origen (intercepto) y β_1 es la pendiente de la línea.

Es decir,

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

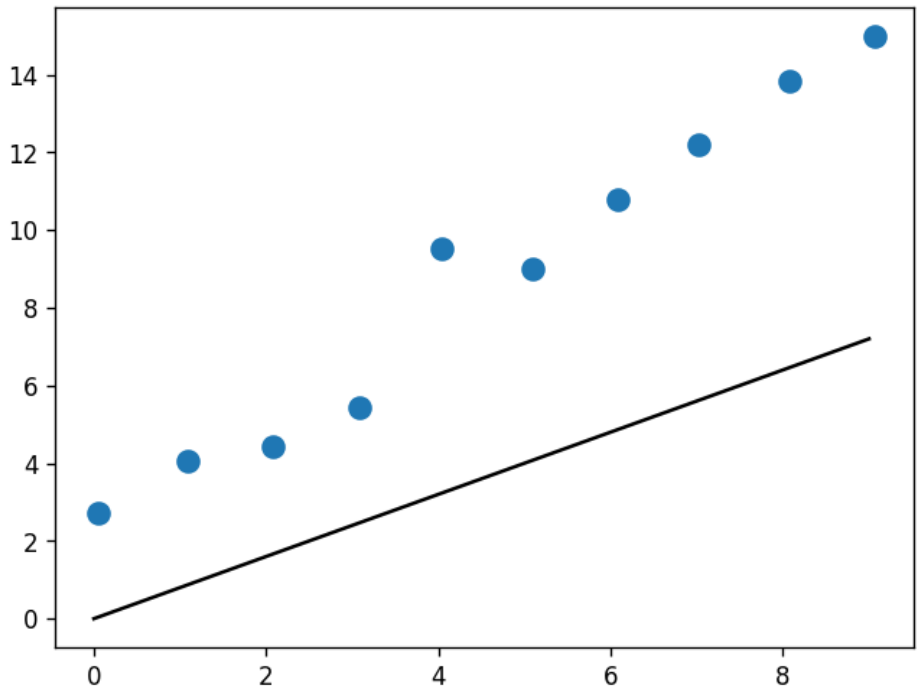
- Si tuvieramos dos puntos, la solución se calcula *fácil*.
- Si tenemos dos puntos, escogemos la *mejor* línea.

¿Cómo sabemos cuál es la mejor línea?



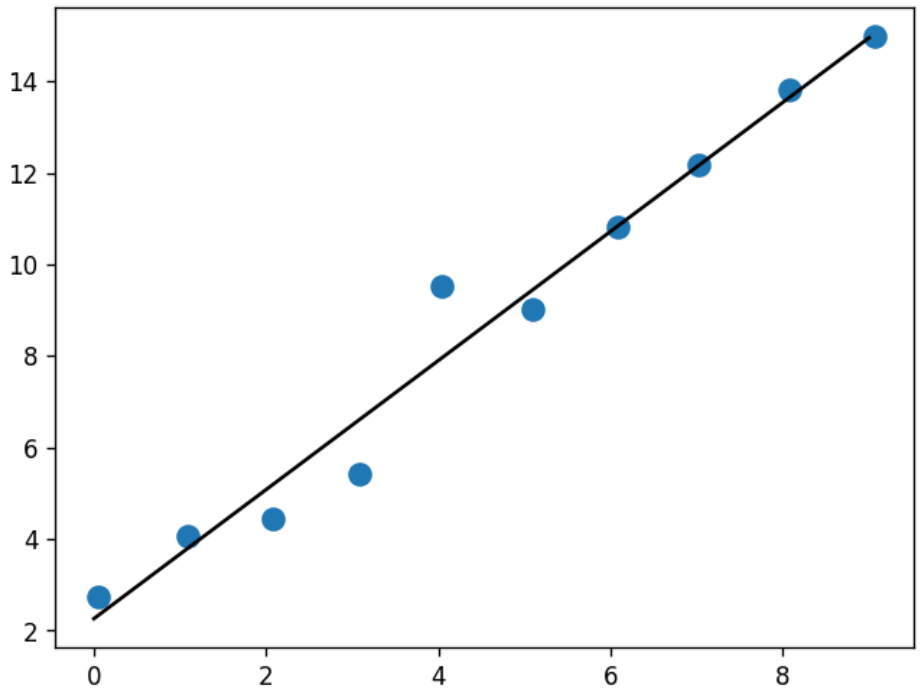
https://www.geogebra.org/m/maeexqmr

¿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

$$e_1 = y_1 - (\beta_0 + \beta_1 x_1)$$
$$e_2 = y_2 - (\beta_0 + \beta_1 x_2)$$
$$\dots$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = |y_1 - (\beta_0 + \beta_1 x_1)|$$
$$e_2 = |y_2 - (\beta_0 + \beta_1 x_2)|$$
$$\dots$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = (y_1 - (\beta_0 + \beta_1 x_1))^2$$
$$e_2 = (y_2 - (\beta_0 + \beta_1 x_2))^2$$
$$\dots$$

¿Cómo sabemos cuál es la mejor línea?

Medimos cada residuo

$$e_1 = (y_1 - (\beta_0 + \beta_1 x_1))^2$$
$$e_2 = (y_2 - (\beta_0 + \beta_1 x_2))^2$$
$$\dots$$

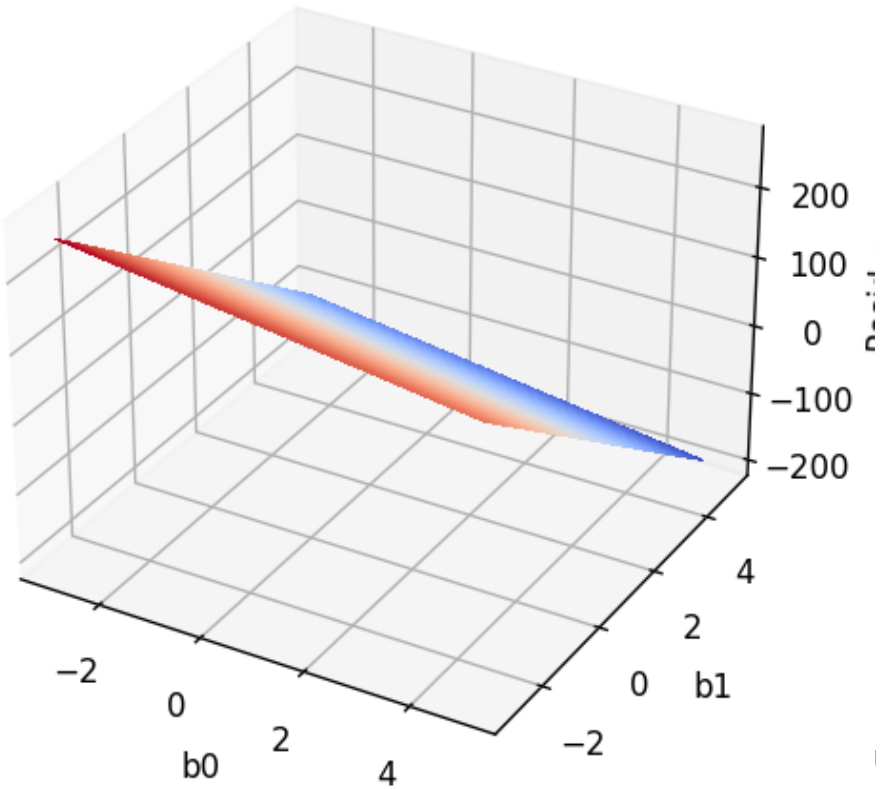
Consideramos la función que suma todos los residuos

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

Minimizando el error

Considerando la función

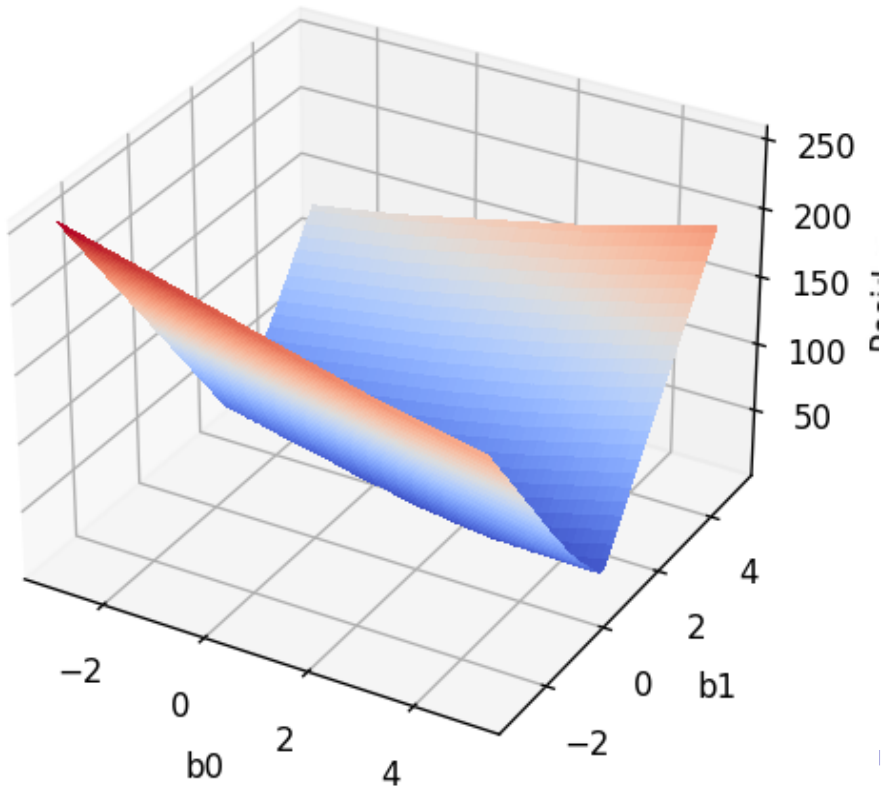
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))$$



Minimizando el error

Considerando la función

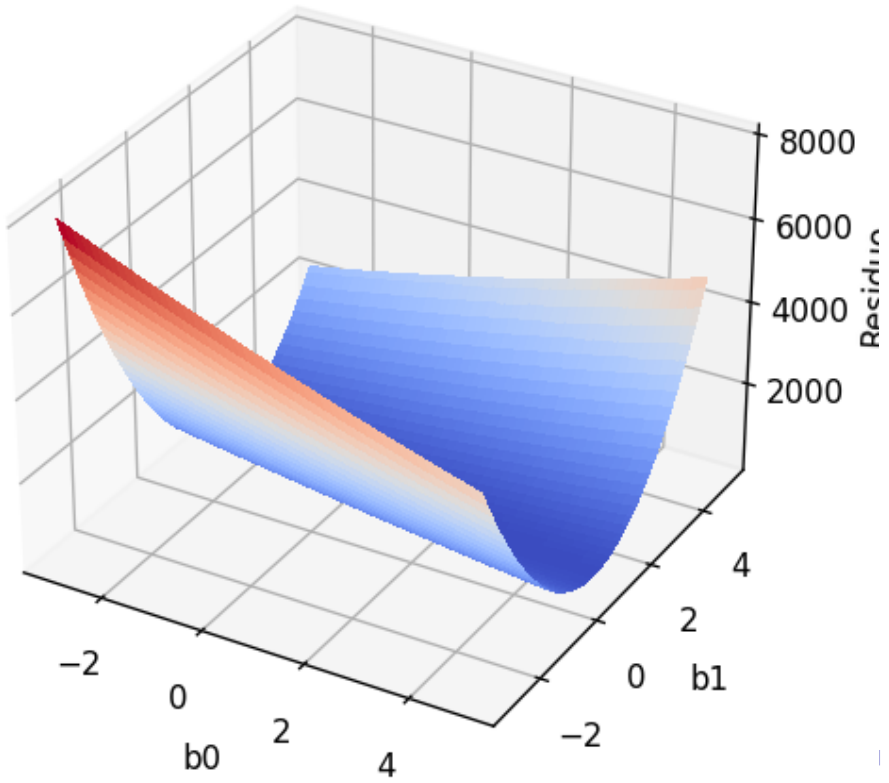
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^{N} |y_i - (\beta_0 + \beta_1 x_i)|$$



Minimizando el error

Considerando la función

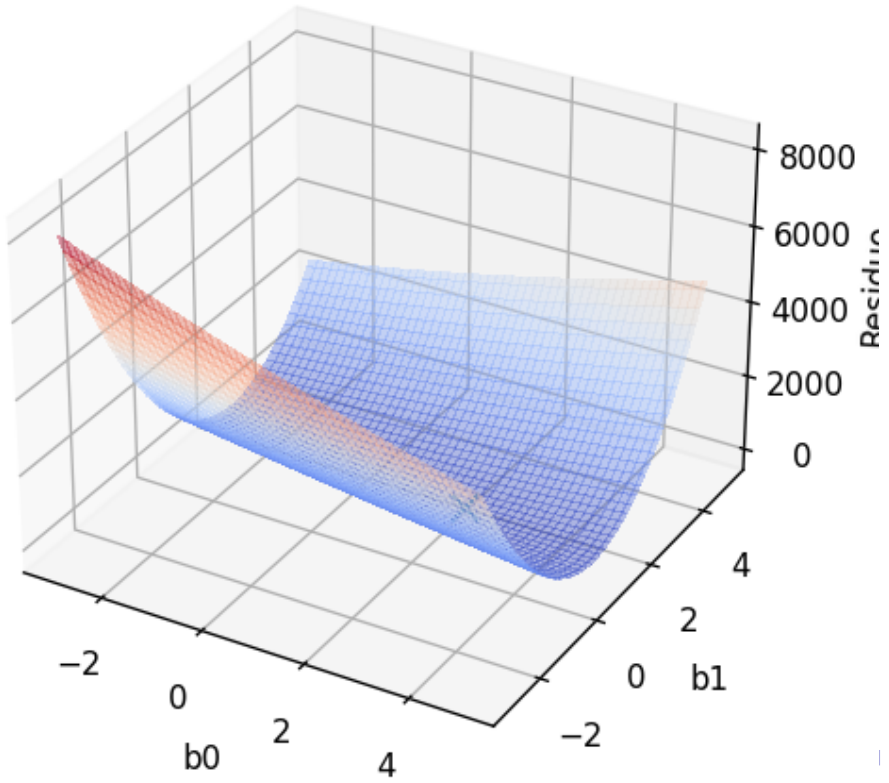
$$\mathcal{E}(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2$$



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

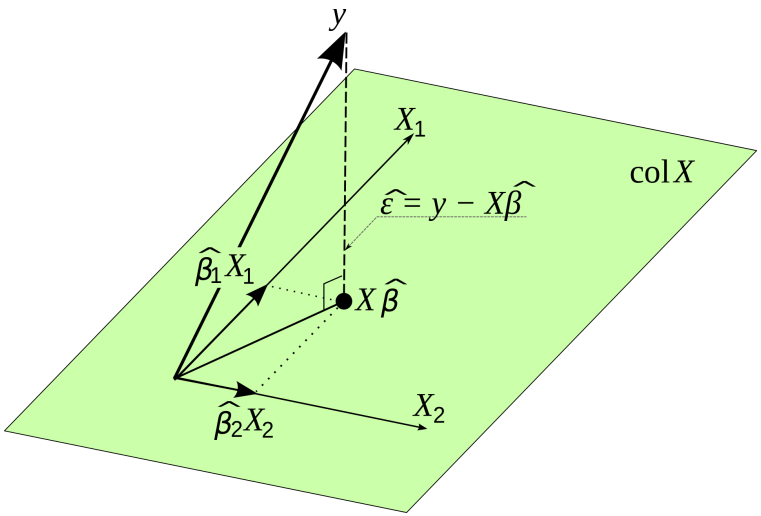
$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 + e_1 \\ y_2 &= \beta_0 + \beta_1 x_2 + e_2 \\ &\cdots \\ y_N &= \beta_0 + \beta_1 x_N + e_N \end{aligned}$$

Lo podemos reescribir como

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

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

La solución: Formulación vectorial



$$\mathbf{e}^T X = 0 \in \mathcal{M}_{1\times 2}$$
$$(\mathbf{y} - X\hat{\beta})^T X = 0$$
$$X^T(\mathbf{y} - X\hat{\beta}) = 0$$
$$X^T\mathbf{y} - X^T X\hat{\beta} = 0$$
$$X^T\mathbf{y} = X^T X\hat{\beta}$$
$$(X^T X)^{-1}X^T\mathbf{y} = \hat{\beta}$$
$$\hat{\beta} = (X^T X)^{-1}X^T\mathbf{y}$$

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

Regresión Lineal Multiple

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

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

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

Y la solución OLS es, otra vez,

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

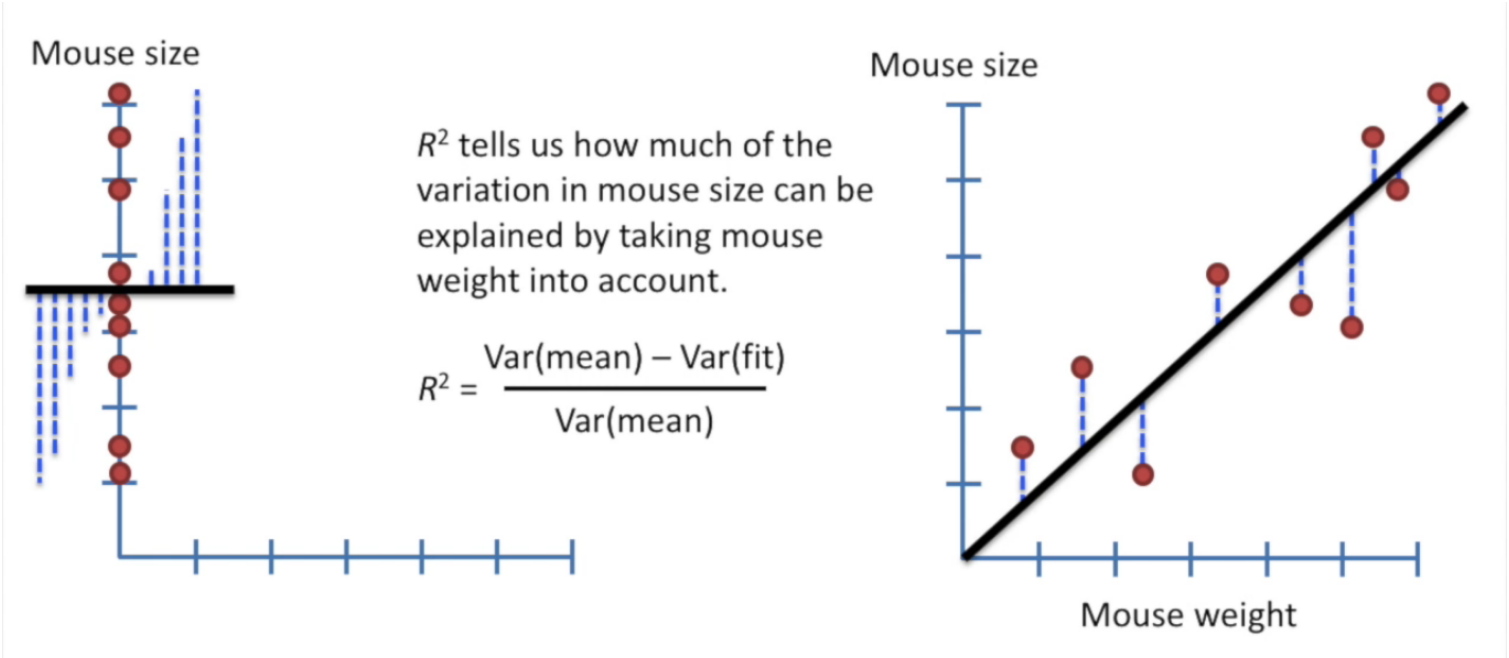
Interpretación de los coeficientes

- 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. Esta propiedad de mantener constantes las demás variables es crucial porque permite evaluar el efecto de cada variable aisladamente de las demás.

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

El coeficiente R^2

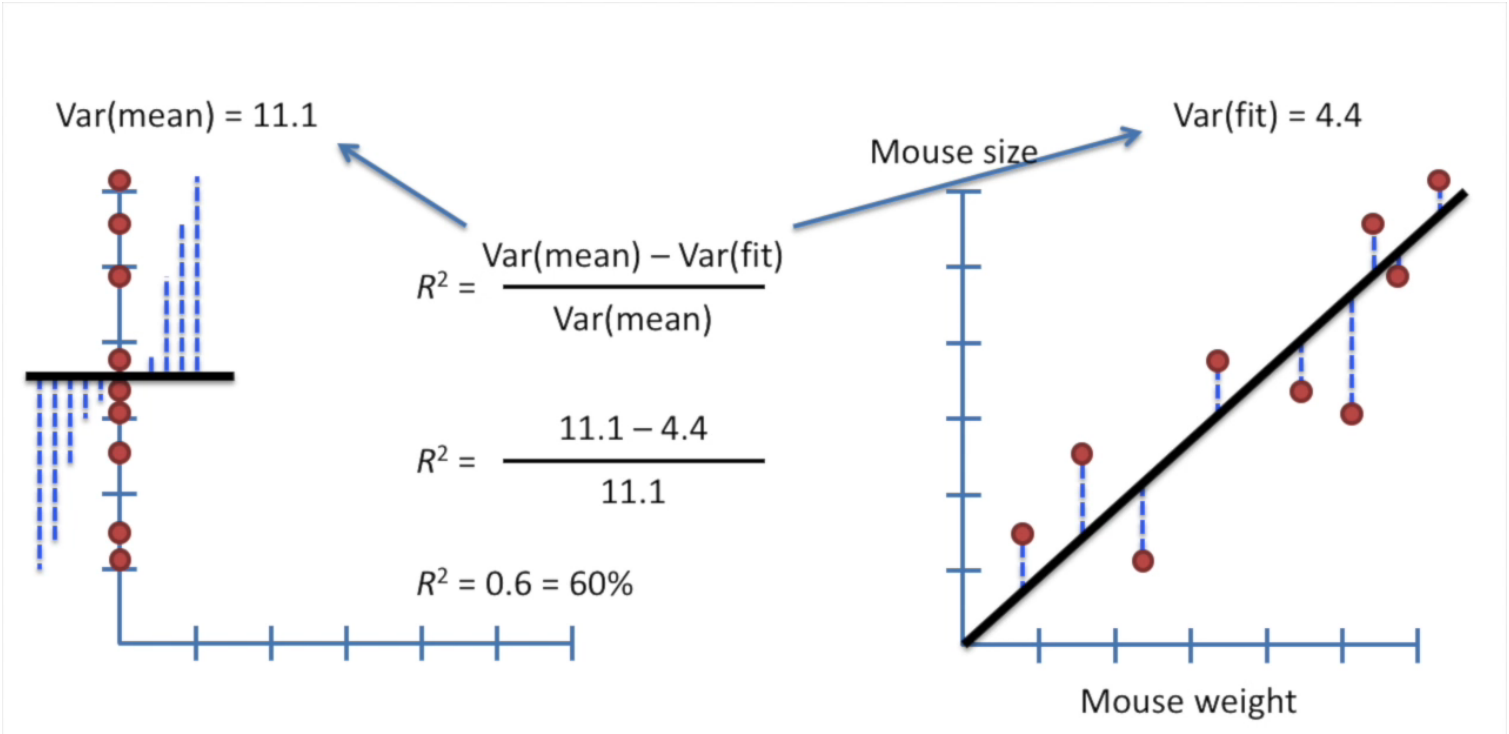
Una pregunta relevante se refiere a la incertidumbre acerca de los parámetros $\hat{\beta}$, ya que estos son variables aleatorias. El coeficiente R^2 explica la varianza de los datos que es explicada por el efecto de las variables dependientes.



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

El coeficiente R^2

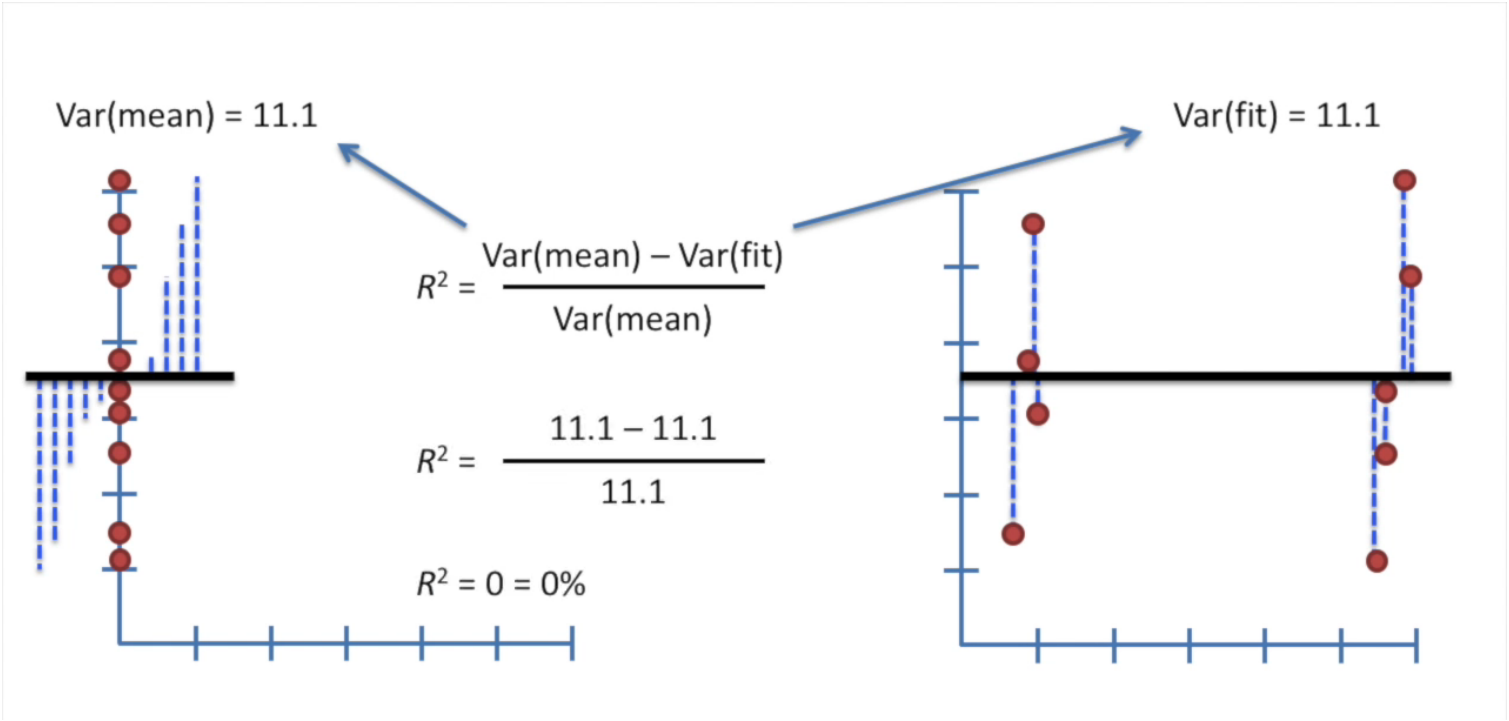
Una pregunta relevante se refiere a la incertidumbre acerca de los parámetros $\hat{\beta}$, ya que estos son variables aleatorias. El coeficiente R^2 explica la varianza de los datos que es explicada por el efecto de las variables dependientes.



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) varbiale 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) varbiale 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) varbiale dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.
- Con estos coeficientes podemos realizar predicciones de nuevos valores.
- Los coeficientes cuantifcan la relación de cada variable con la variable de salida.

Resumiendo

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

Resumiendo

- La regresión lineal descubre la relación lineal entre varias variables predictoras y (una) varbiale dependiente. Para esto usamos OLS y obtenemos coeficientes para la regresión.
- Con estos coeficientes podemos realizar predicciones de nuevos valores.
- Los coeficientes cuantifcan 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.
- Los p-values nos indican si el coeficiente R^2 es estadísticamente significativo.

Table of Contents

Regresión Polinomial

En la regresión lineal queremos ajustar un módelo

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

a los datos

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

esto se traduce en el sistema

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

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

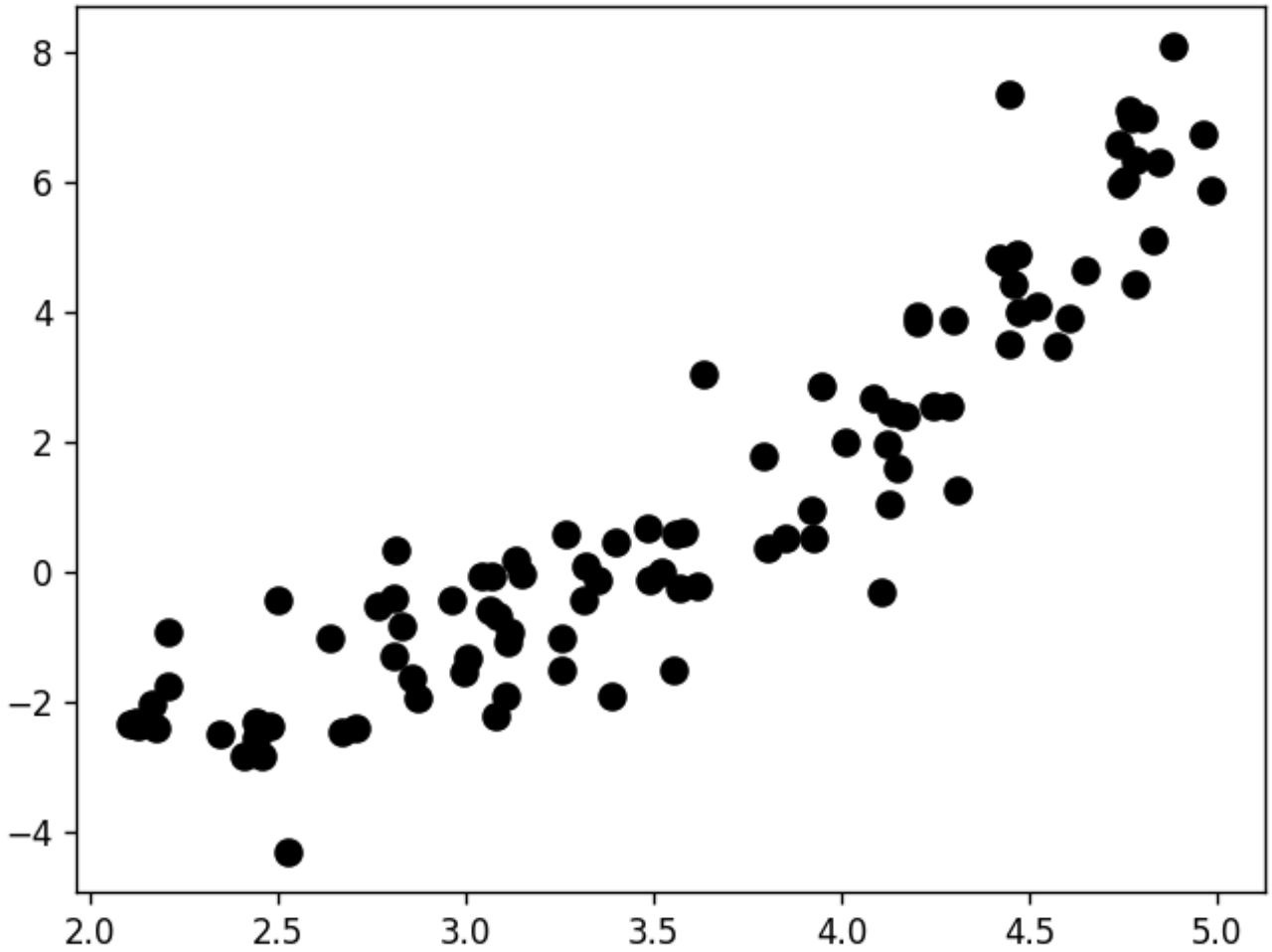
$$\begin{array}{c|c|c} x_1^{(1)} & x_1^{(2)} & y_1 \\ x_2^{(1)} & x_2^{(2)} & y_2 \\ \dots & \dots & \dots \\ x_N^{(1)} & x_N^{(2)} & y_N \end{array}$$

Es necesario generar nuevas columnas con los datos de grado 2:

$$\begin{array}{c|c|c|c|c|c} x_1^{(1)} & x_1^{(2)} & {x^{(1)}}_1^2 & {x^{(2)}}_1^2 & x_1^{(1)} x_1^{(2)} & y_1 \\ x_2^{(1)} & x_2^{(2)} & {x^{(1)}}_2^2 & {x^{(2)}}_2^2 & x_2^{(1)} x_2^{(2)} & y_2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_N^{(1)} & x_N^{(2)} & {x^{(1)}}_N^2 & {x^{(2)}}_N^2 & x_N^{(1)} x_N^{(2)} & y_N \end{array}$$

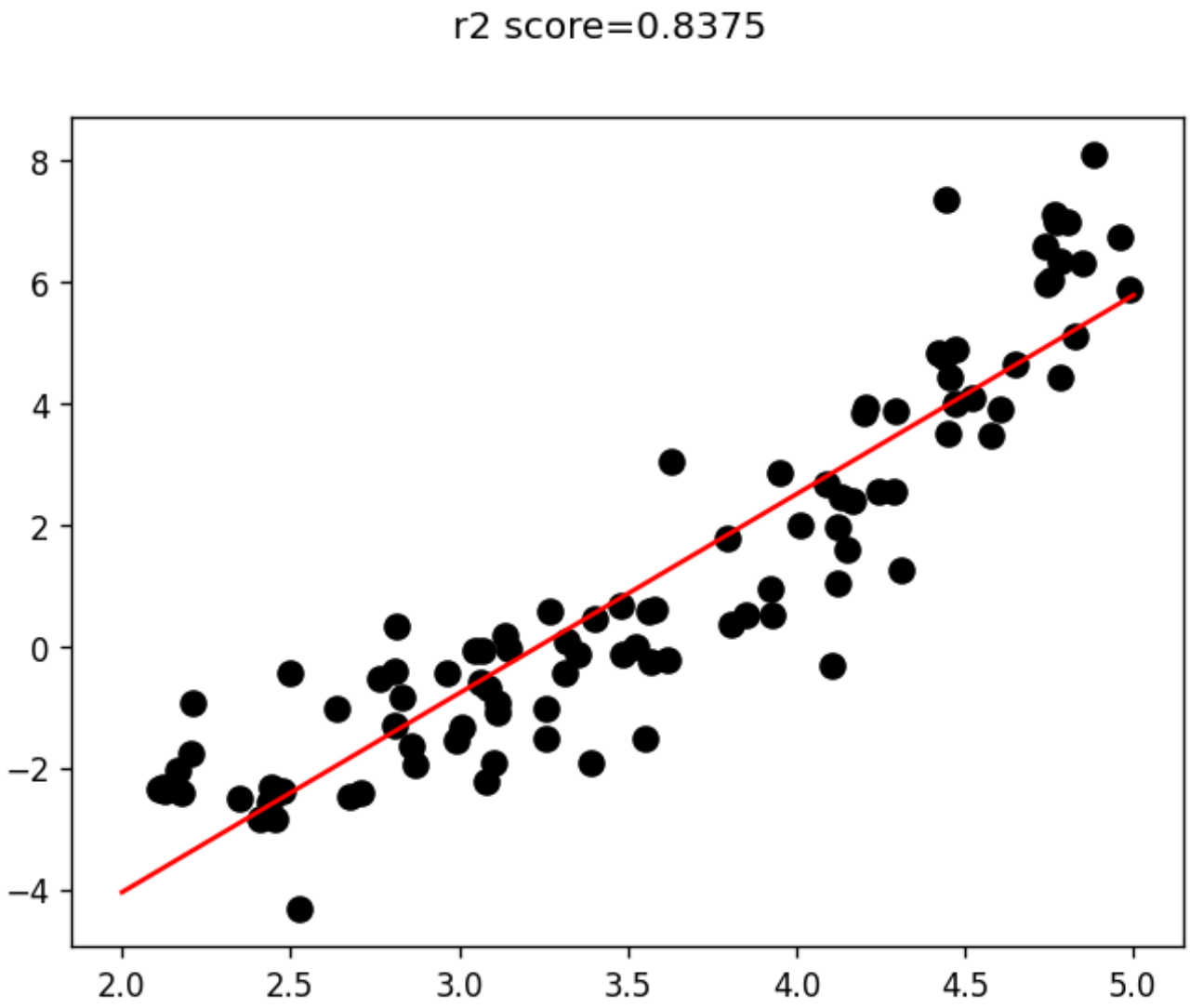
Regresión Polinomial: Ejemplo

Los datos



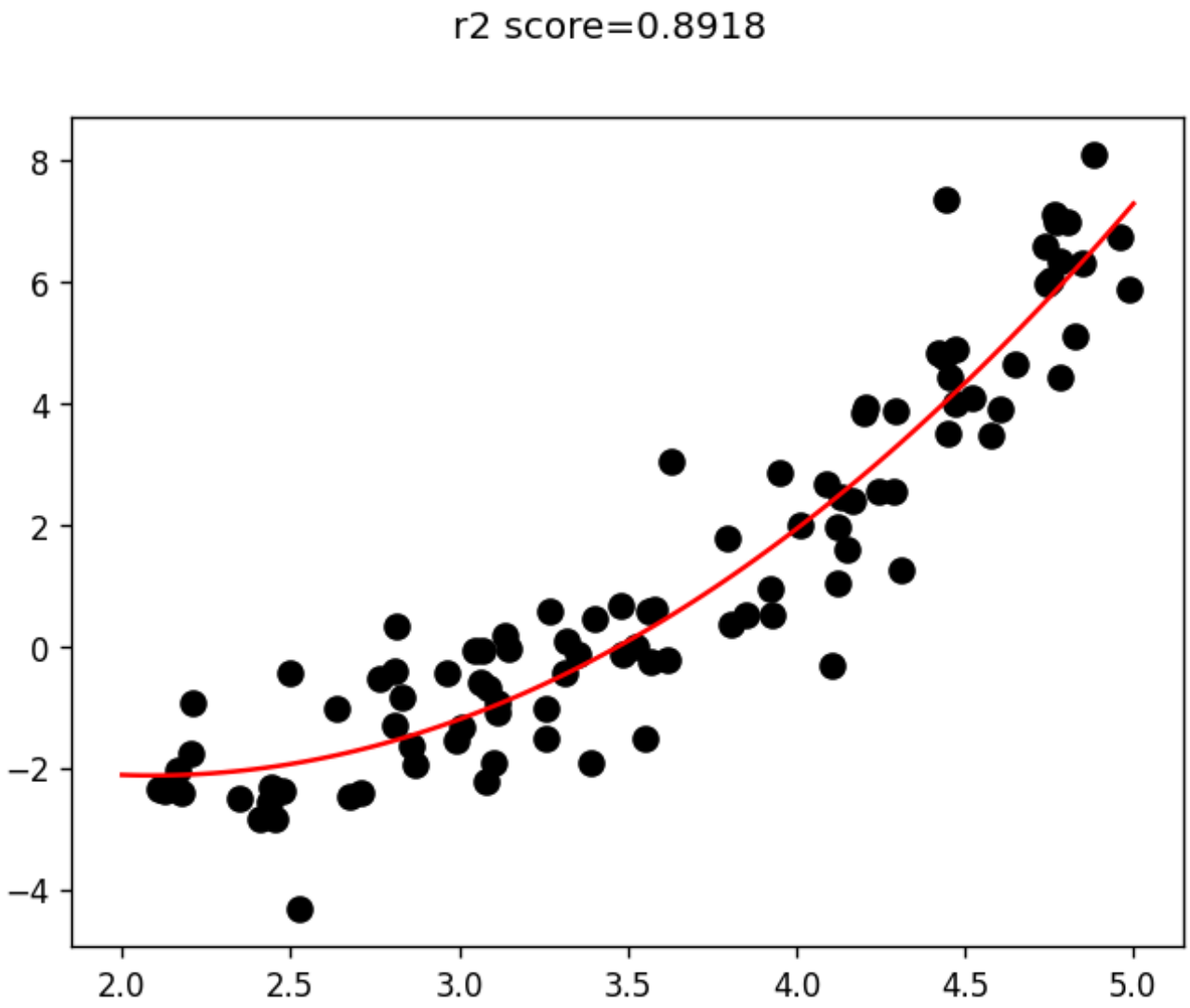
Regresión Polinomial: Ejemplo

Regresión lineal



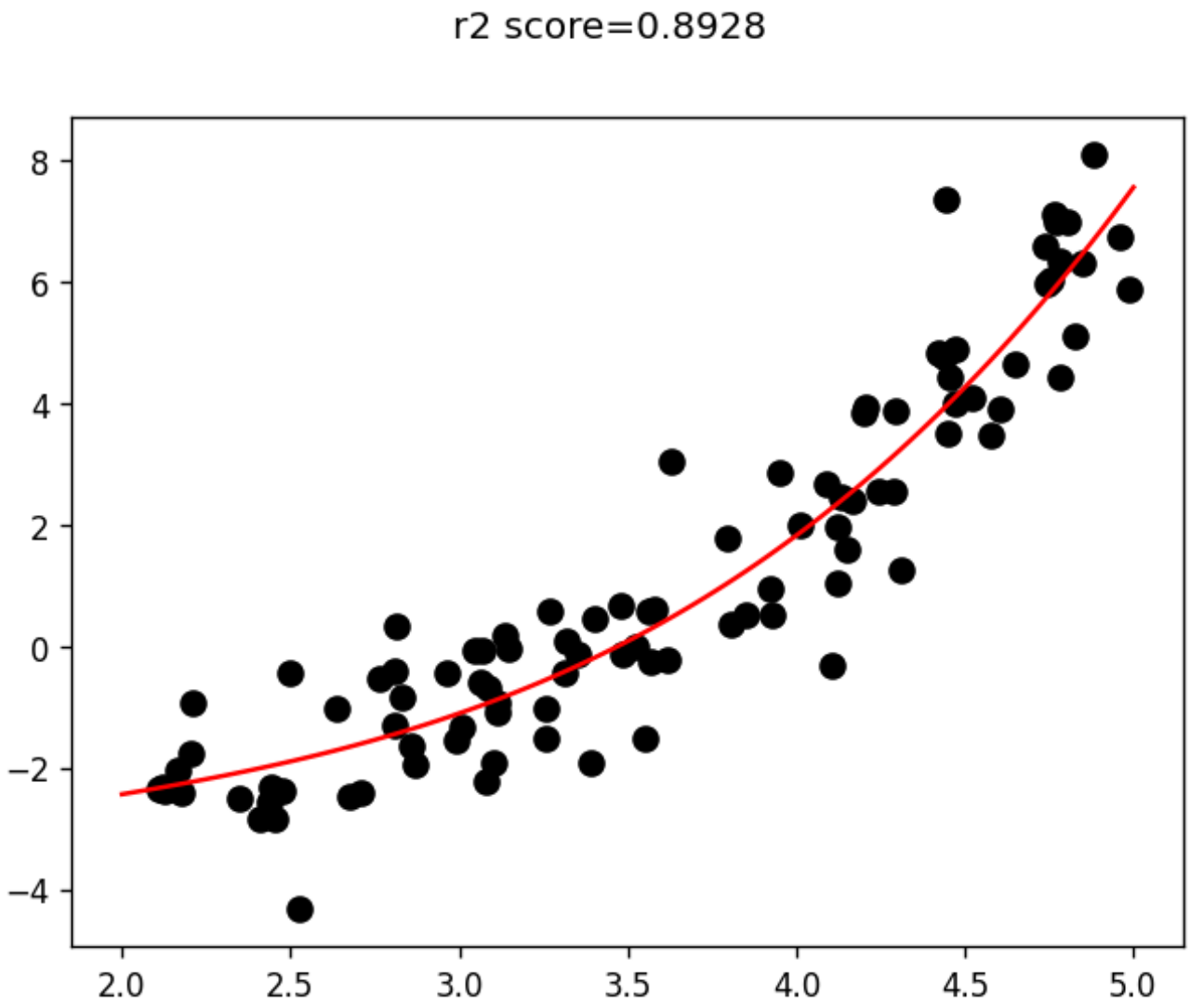
Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 2



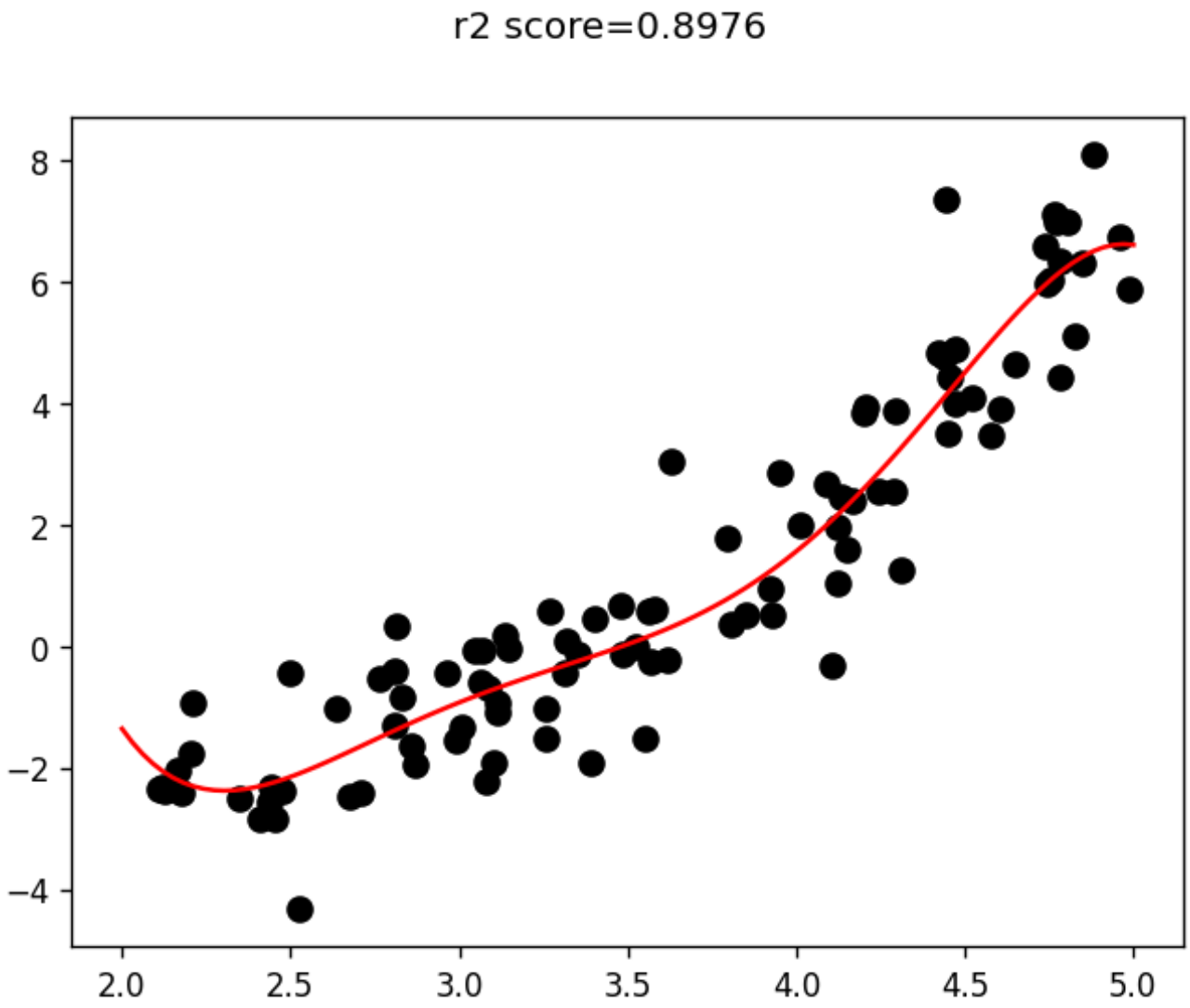
Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 3

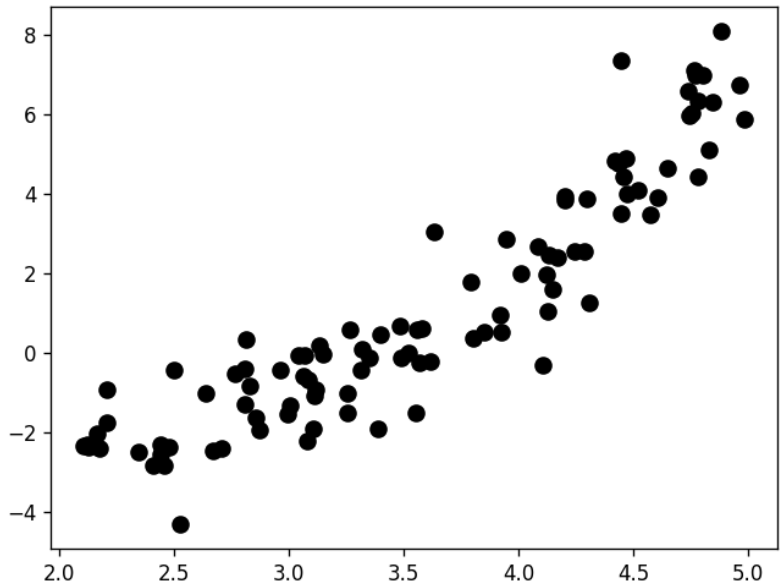


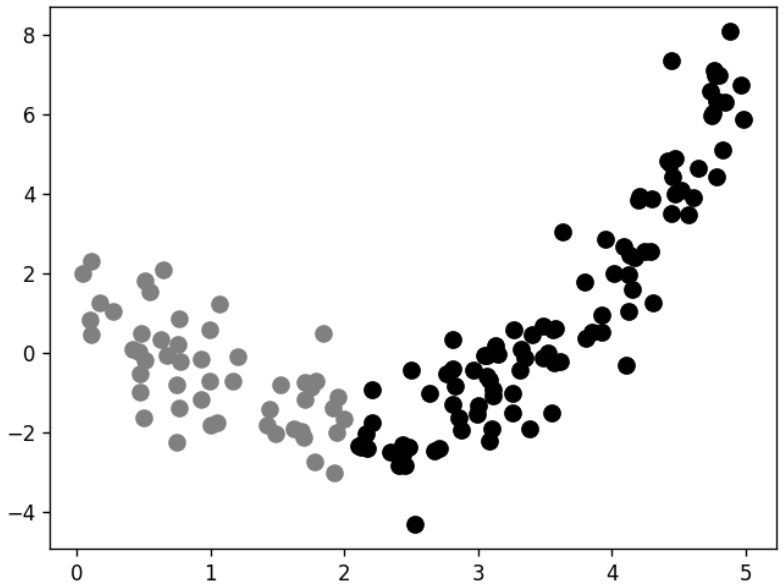
Regresión Polinomial: Ejemplo

Regresión lineal con un polinomio de grado 5

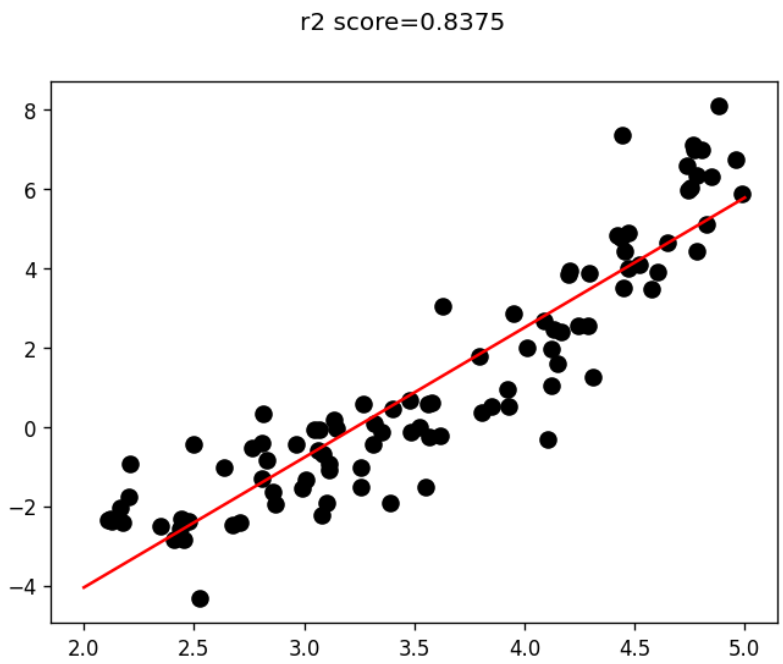


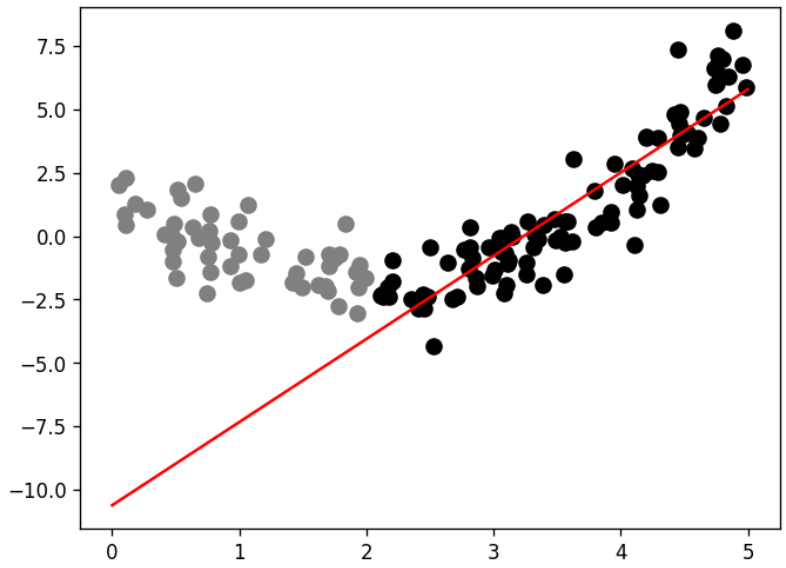
Regresión Polinomial: Ejemplo con nuevos datos



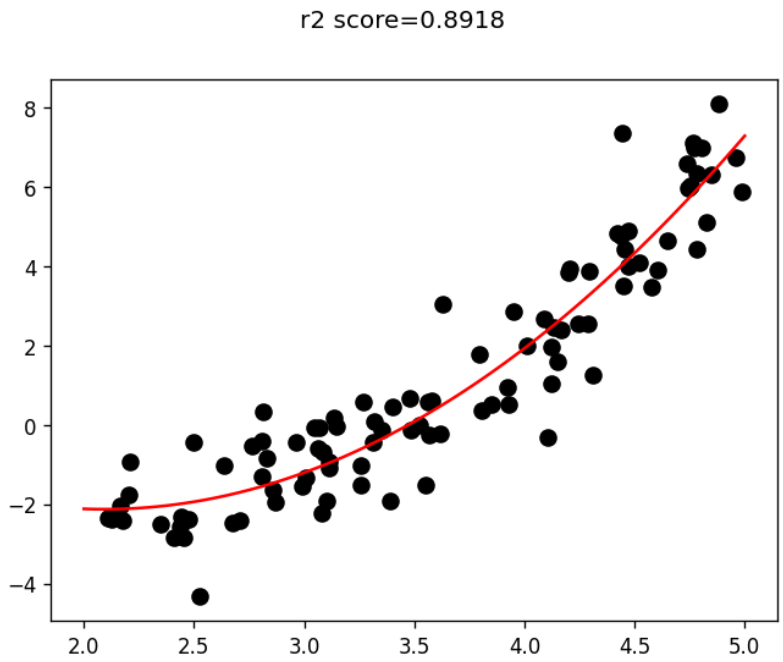


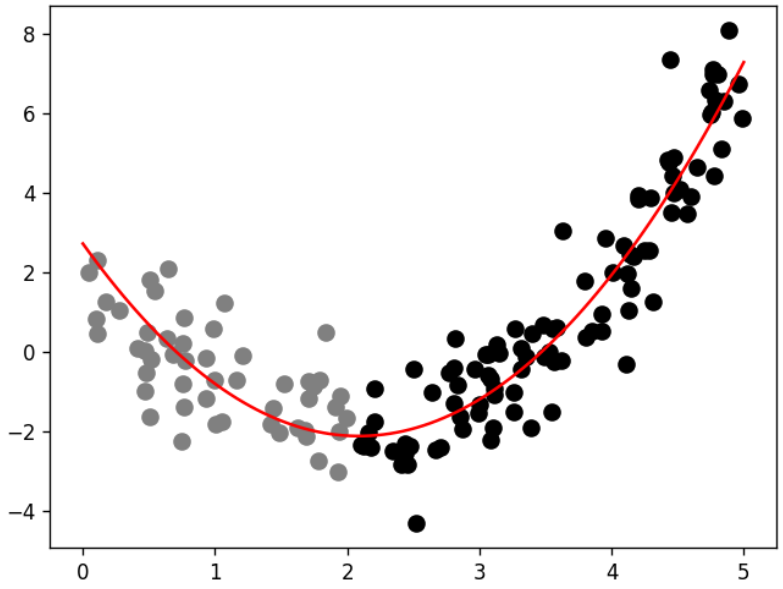
Regresión Polinomial: Ejemplo con nuevos datos



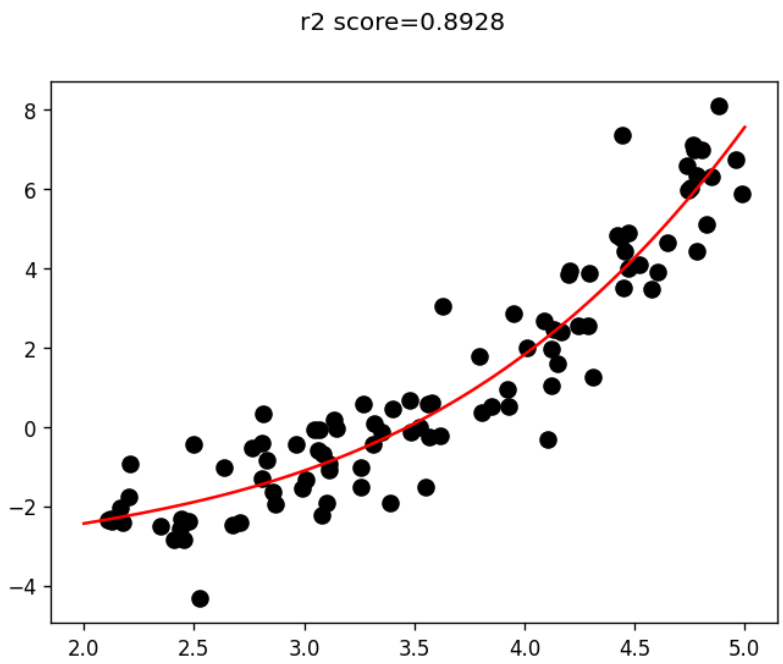


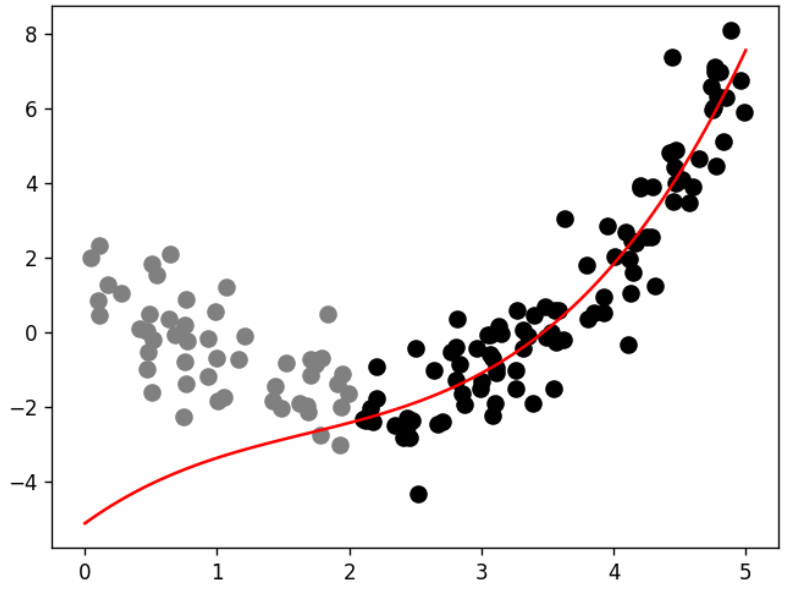
Regresión Polinomial: Ejemplo con nuevos datos



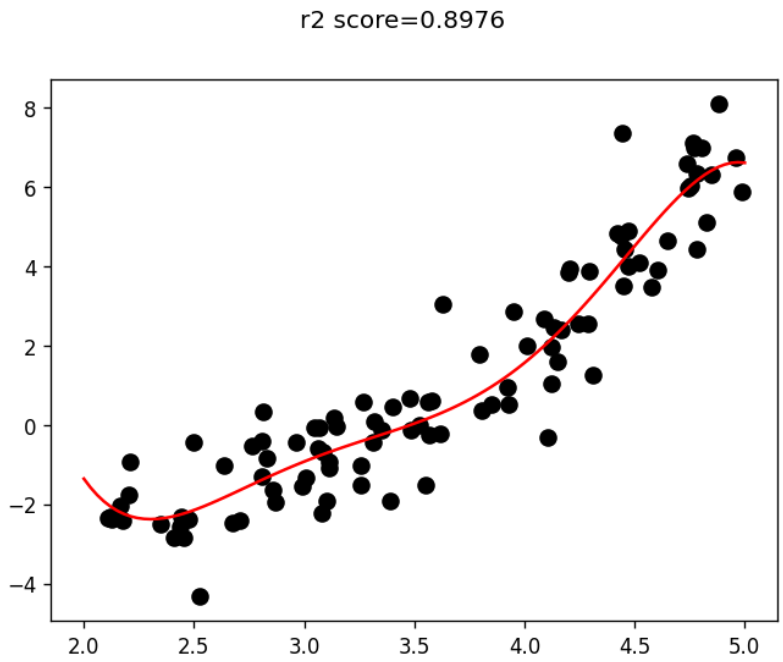


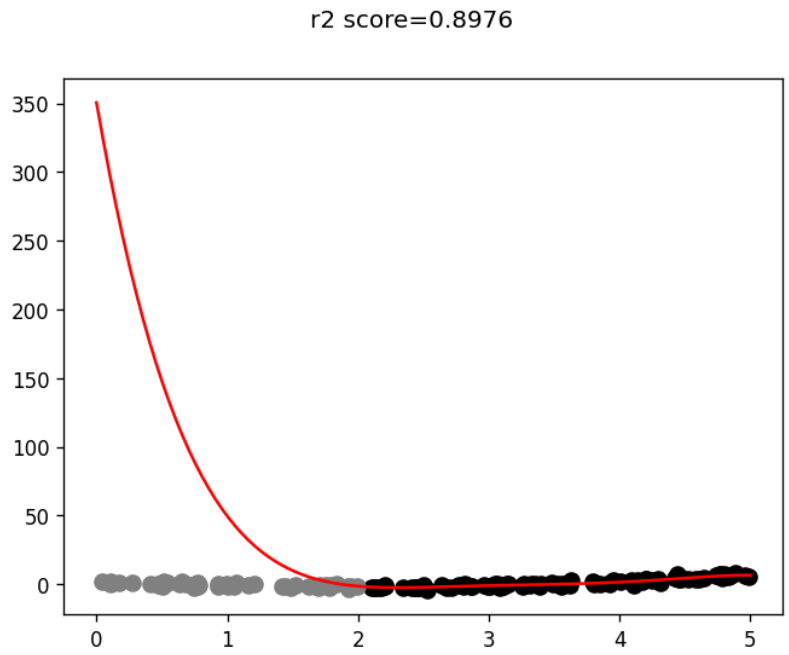
Regresión Polinomial: Ejemplo con nuevos datos





Regresión Polinomial: Ejemplo con nuevos datos





Conclusión

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

Conclusión

No es bueno usar un módelo 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.