Universidad del Rosario

Analisis Avanzado de Datos

W2. Regresión Lineal Múltiple

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Profesor

FERNEY ALBERTO BELTRAN MOLINA
ferney.beltran@urosario.edu.co

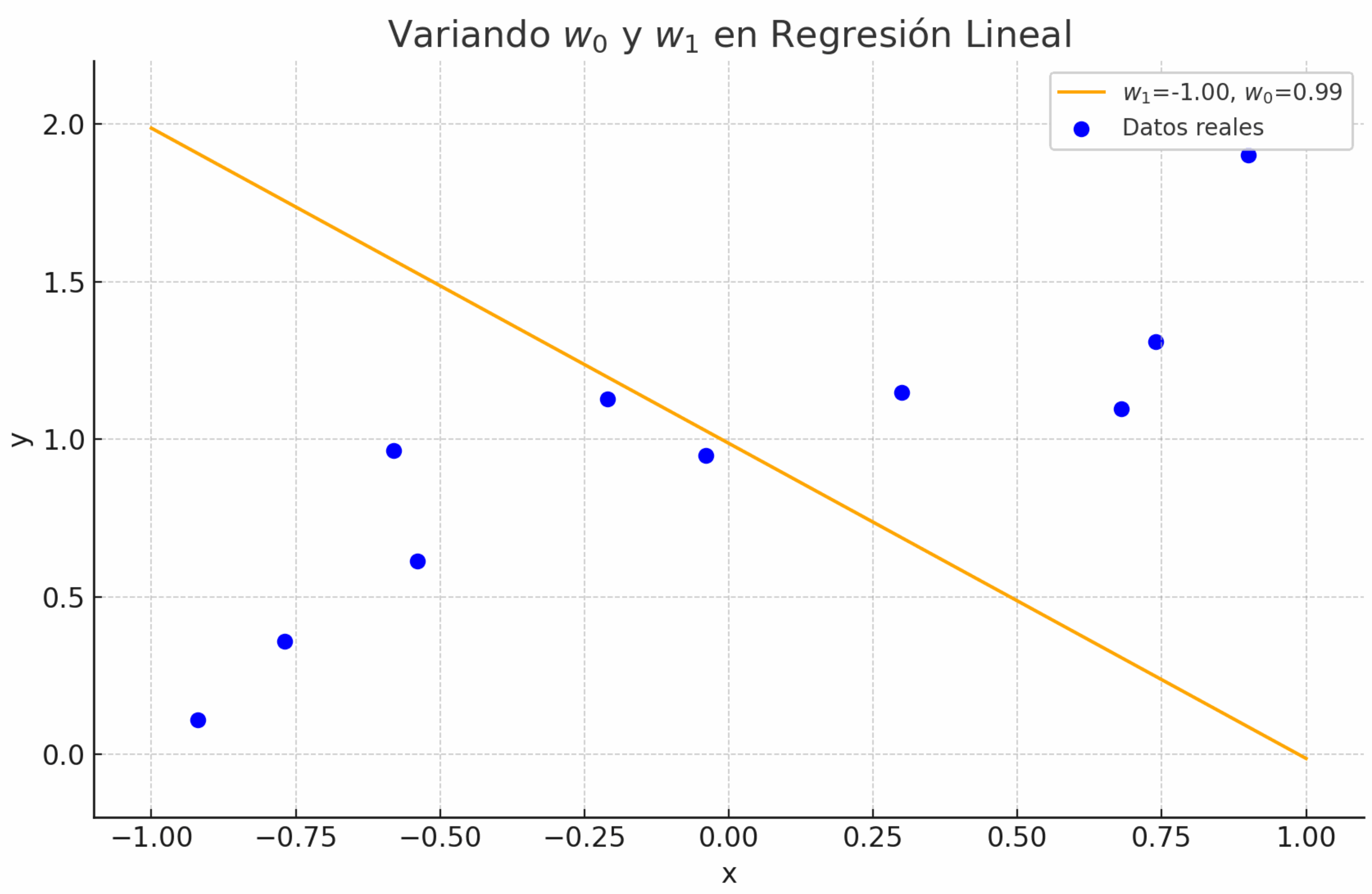
Ingeniero Electrónico.
Magister en TIC
Candidato Doctor en TIC

Director del Centro de investigación e innovación CEINTECCI.
Miembro de la junta directiva Avanciencia

Procesamiento y análisis de datos basadas en IA.
Simulación y modelado por computación,
Optimizan Sistemas de procesamiento en hardware y software
Diseño de sistemas electrónicos reconfigurables

Regresión simple

Candidatos de modelo ¿cuál fue el mejor?



En la regresión lineal simple, los modelos candidatos se definen por

$$f(x) = w_0 + w_1 x$$

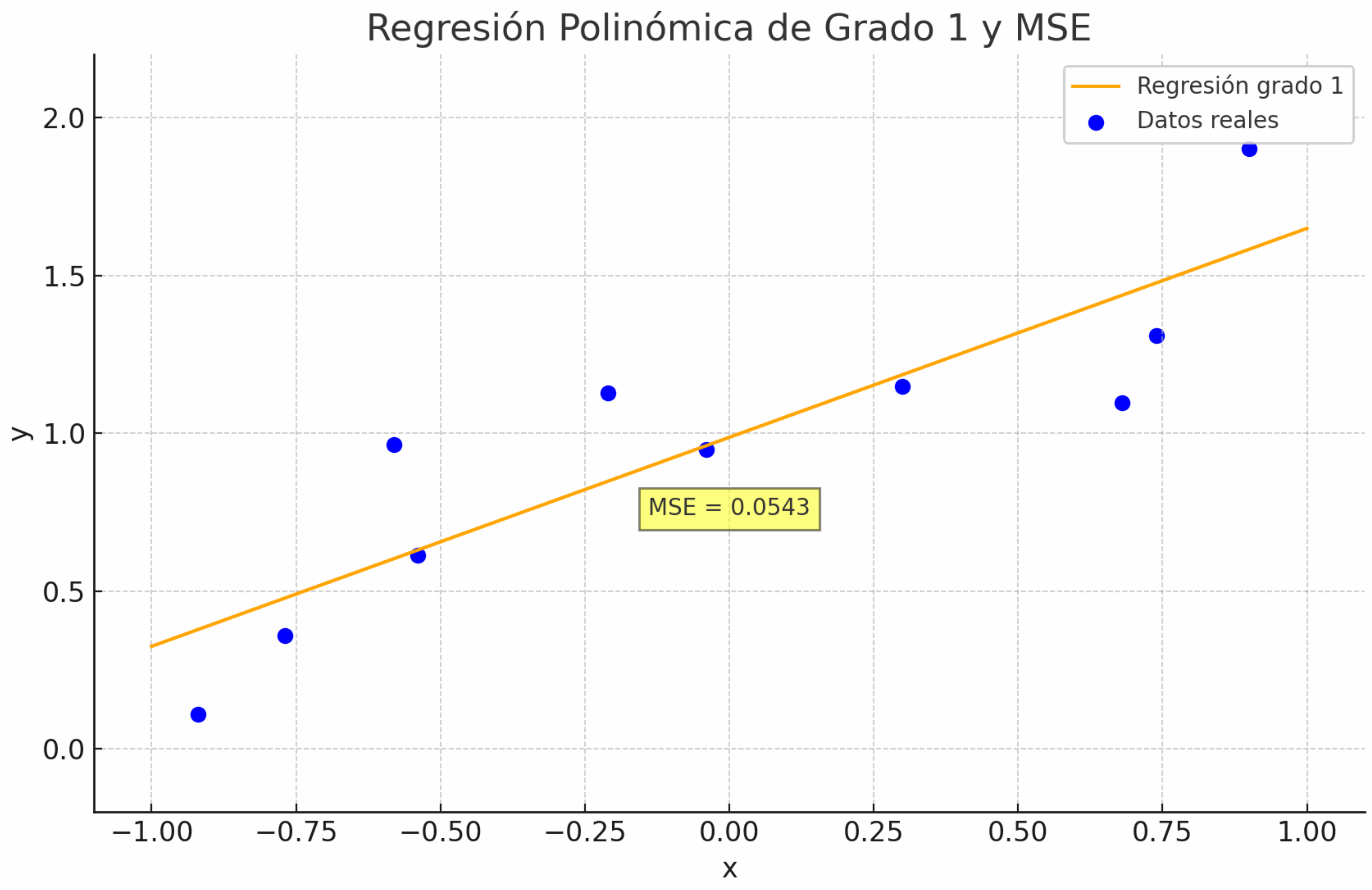
$$\hat{y}_i = f(x_i) = w_0 + w_1 x_i$$

Ecualizador

$$f_{best}(x) = \arg\min_f \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Regresión en el laboratorio

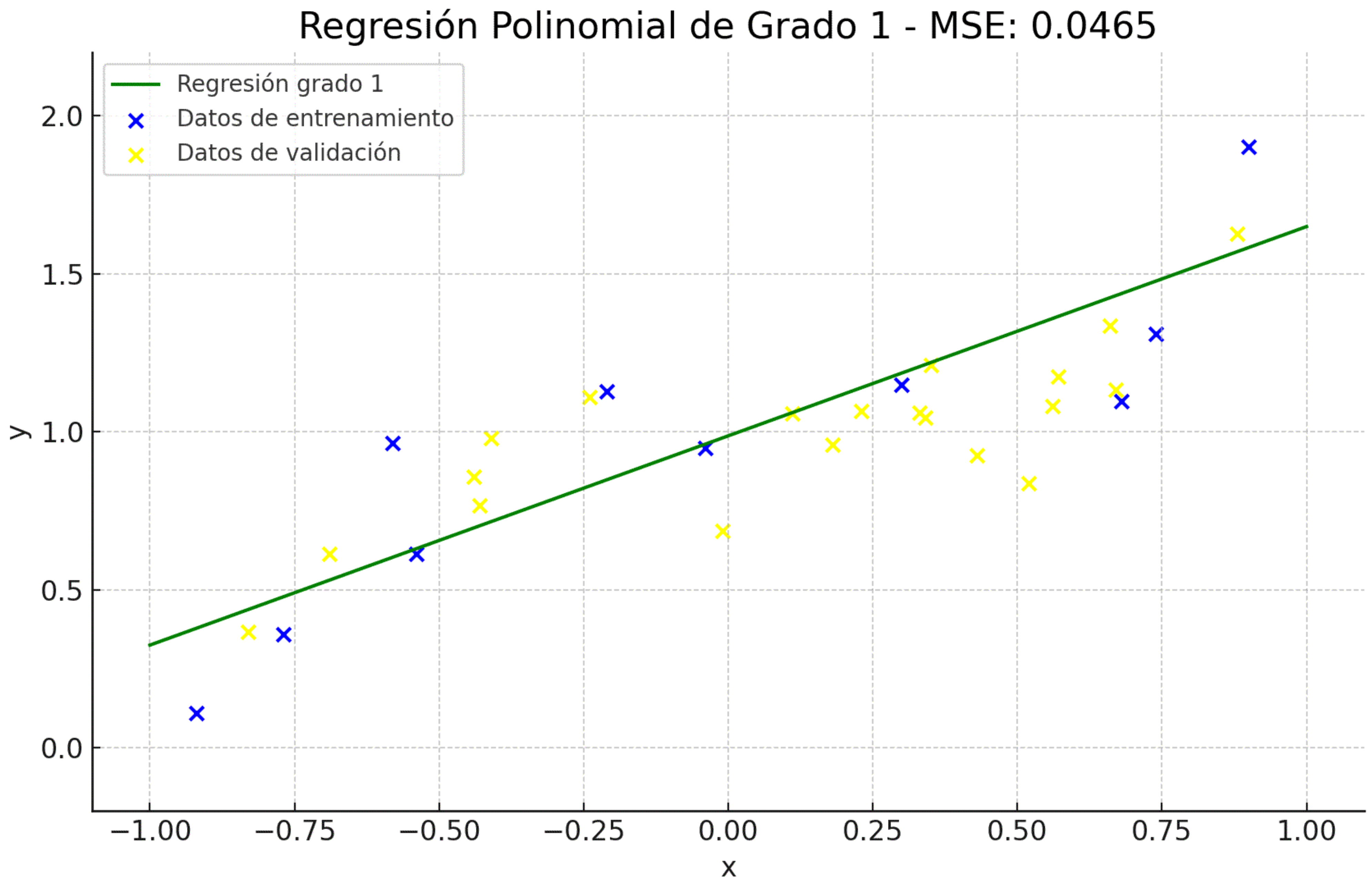
Candidatos de modelo y ranking de modelos ¿cuál fue el mejor?



Dado una familia de modelos de regresión, la **solución de mínimos cuadrados es el modelo que minimiza el error cuadrático medio en nuestro** conjunto de datos de entrenamiento.

Regresión en el laboratorio

Candidatos de modelo y ranking de modelos ¿cuál fue el mejor?



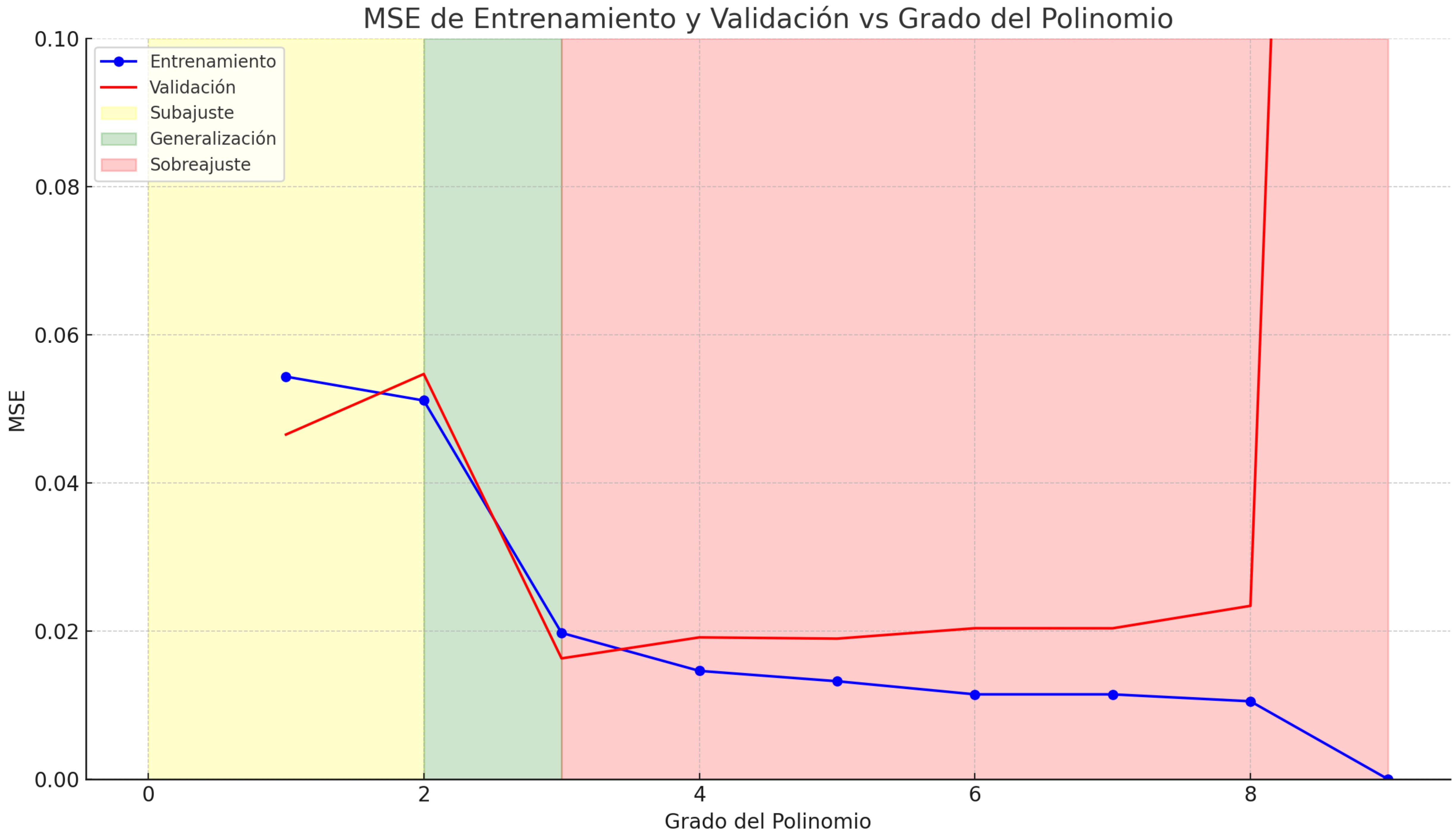
La capacidad de nuestro modelo de transferir lo que hizo durante aprendizaje a producción

Modelo "Polinomio grado 3"

Generalización pasar de aprendizaje a producción

Generalización (principio fundamental

Que vimos

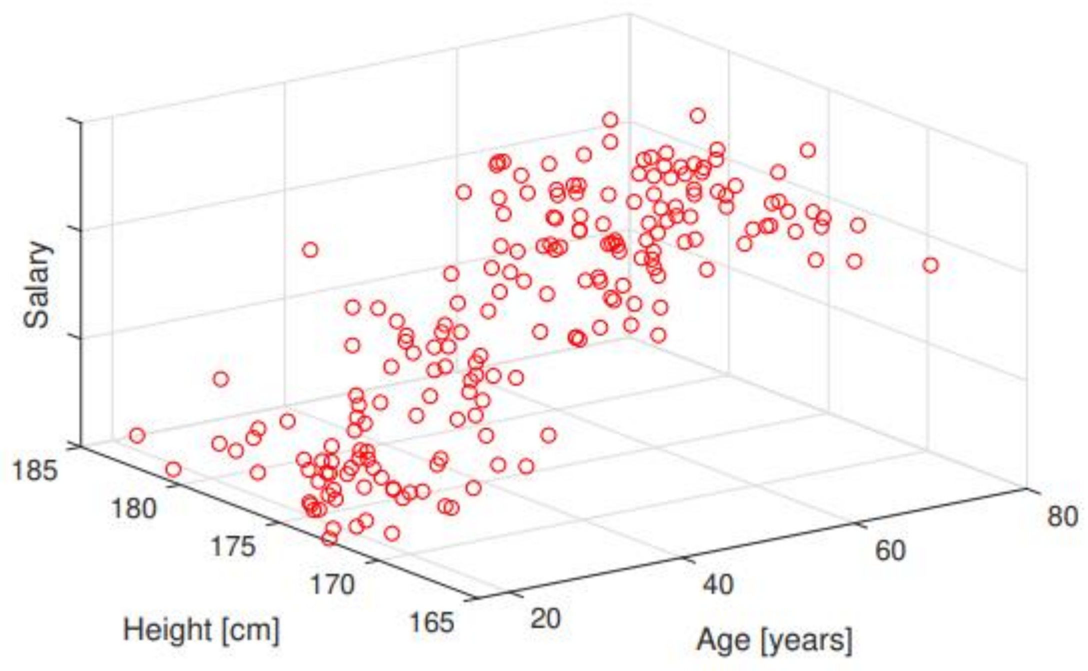


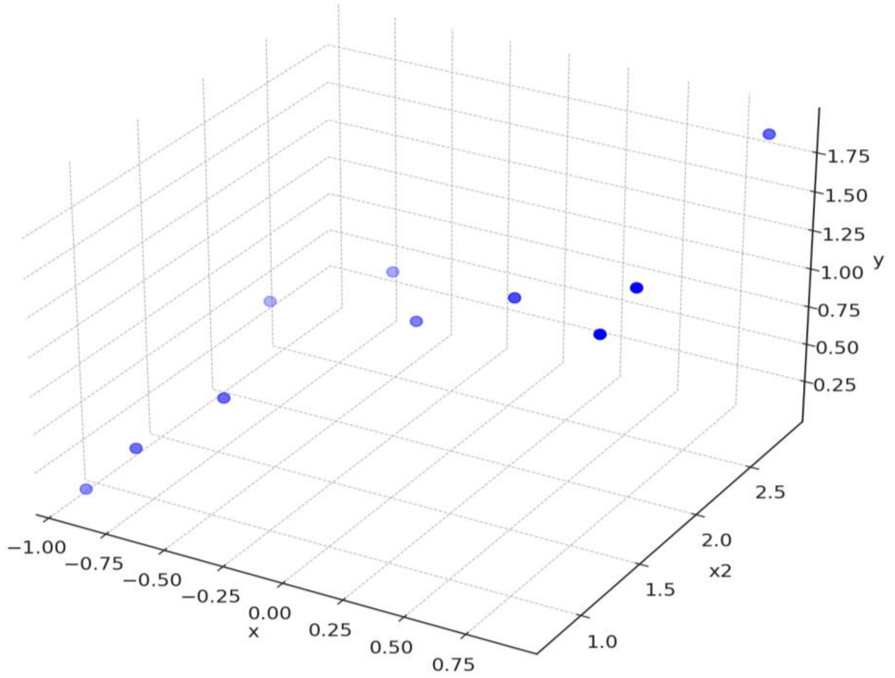
- Subajuste (Underfitting)
- Generalización (Good Fit)
- Sobreajuste (Overfitting)

Grado 3
¡es decir, el mejor modelo durante la implementación!

Regresión Múltiple

Cuando tenemos dos o mas predictores





Regresión múltiple puede ser expresada como: $\hat{y}_i = f(x_i)$

Notemos entonces que una regresión simple puede ser traslada a un escenarios de multivariable

Considerando el siguiente modelo

$$f(x) = w_0 + w_1 x_1 + \ldots + w_K x_K$$

$$\mathbf{x} = [1, x_{i,1}, x_{i,1}, \ldots x_{i,k}]^T$$

Si contamos con un conjunto de datos de N muestras, podemos determinar los parámetros de la solución de mínimos cuadrados mediante las ecuaciones normales

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

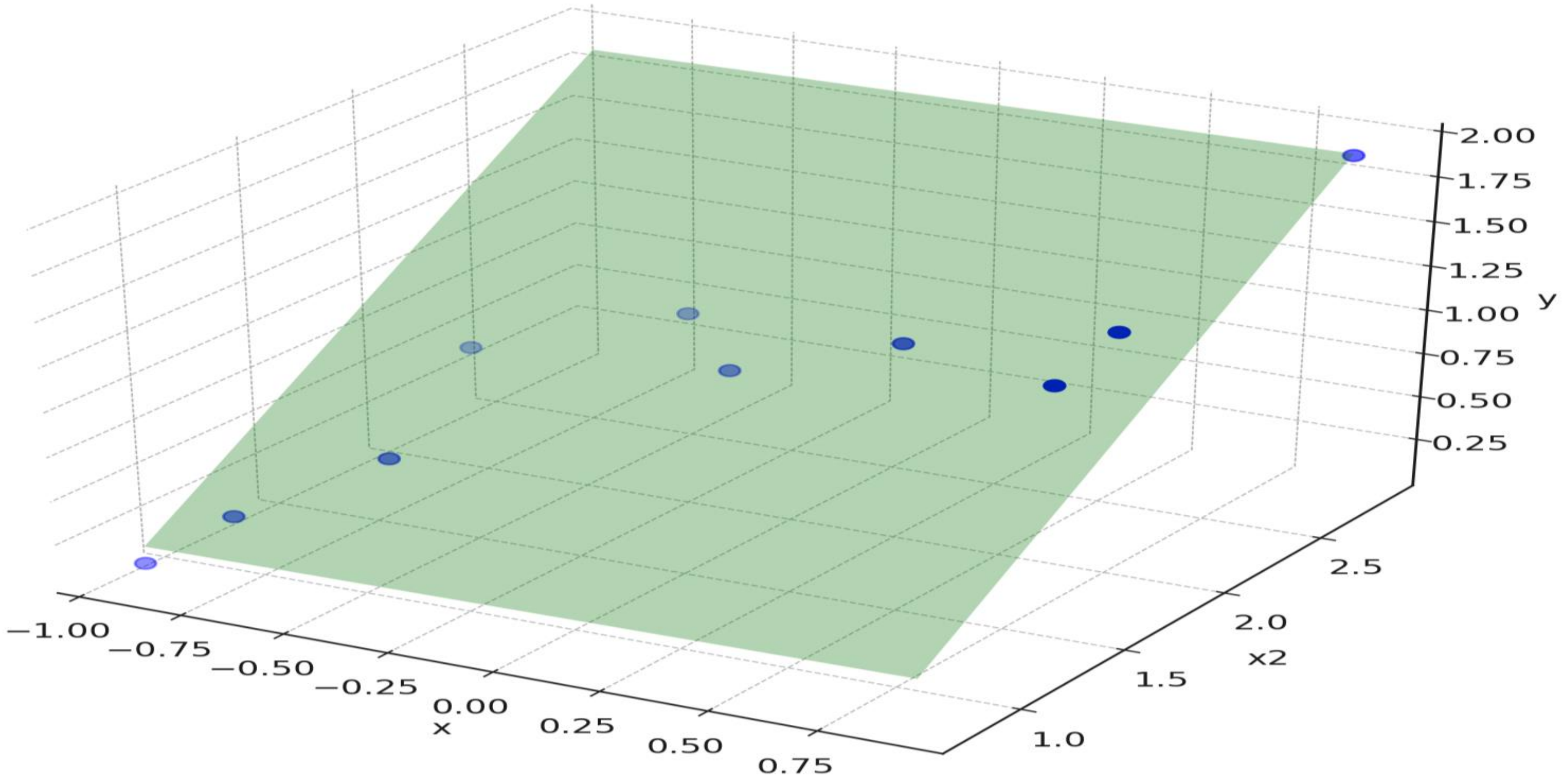
Regresión Múltiple: formulación

Los modelo **de regresión lineal múltiple** pueden ser expresado como:

$$f(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i = w_0 + w_1 x_{i,1} + \cdots + w_K x_{i,K}$$

Donde: $\boldsymbol{w} = [w_0, w_1, \ldots, w_K]^T$ son los son los parámetros del modelo

1		......		k
l				
	Xi,2			Xi,k



MSE = 0,0032

W0=−0.4215 (intercepto)
W1=0.0160 (coeficiente para x)
W2=0.8146 (coeficiente para x2)

Regresión Múltiple: formulación

Para **regresión lineal múltiple** el data set puede ser representado por la **matriz de diseño**:

$$\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_N^T \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \ldots & x_{1,K} \\ 1 & x_{2,1} & x_{2,2} & \ldots & x_{2,K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N,1} & x_{N,2} & \ldots & x_{N,K} \end{bmatrix}$$

Y el vector de etiquetas por:

$$\boldsymbol{y} = [y_1, \ldots, y_N]^T = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

Y por lo tanto los coeficientes del mejor modelo esta dado :

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

Regresión polinómica

Recuerda que el modelo polinomio es $f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_D x_i^D$

Donde *D* es el grado del polinomio

Al tratar las potencias del predictor $\mathscr{X}$ como predictores en sí mismos, los modelos polinómicos simples se pueden expresar como modelos lineales múltiples.

$$f(x_i) = w_0 + w_1 x_i + w_2 x_i^2 + w_3 x_i^3 = \boldsymbol{w}^T \boldsymbol{\phi}_i$$

Donde,

$$\boldsymbol{\phi}_i = [1, x_i, x_i^2, x_i^3]^T$$

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix}$$

Por lo tanto, Existe una solución exacta de mínimos cuadrados para la regresión polinómica simple.

Regresión polinómica multivariable

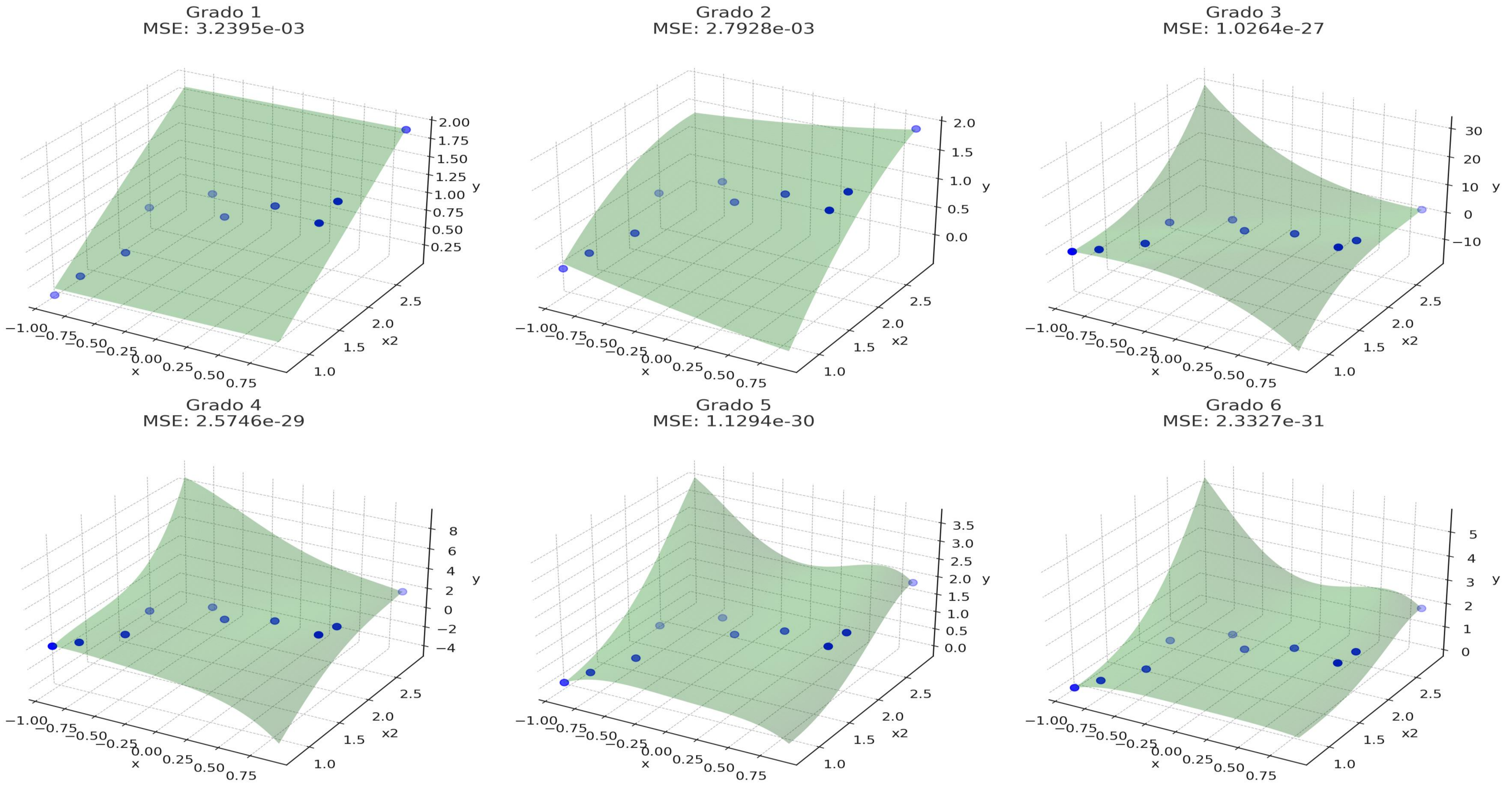
La matriz de diseño para una regresión polinómica bivariada de grado 3 se vería así:

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,1}^2 & x_{1,1}^3 & x_{2,1} & x_{2,1}^2 & x_{2,1}^3 & x_{1,1}x_{2,1} & x_{1,1}^2x_{2,1} & x_{1,1}x_{2,1}^2 \\ 1 & x_{1,2} & x_{1,2}^2 & x_{1,2}^3 & x_{2,2} & x_{2,2}^2 & x_{2,2}^3 & x_{1,2}x_{2,2} & x_{1,2}^2x_{2,2} & x_{1,2}x_{2,2}^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{1,n}^2 & x_{1,n}^3 & x_{2,n} & x_{2,n}^2 & x_{2,n}^3 & x_{1,n}x_{2,n} & x_{1,n}^2x_{2,n} & x_{1,n}x_{2,n}^2 \end{bmatrix}$$

Donde:

La primera columna (de unos) corresponde al término constante (intercepto).
Las siguientes tres columnas corresponden a las potencias de ***x1*** hasta el grado 3.
Las próximas tres columnas corresponden a las potencias de ***x2*** hasta el grado 3.
Las últimas tres columnas corresponden a las interacciones entre ***x1*** ***y*** ***x2*** hasta el grado 3.
n es el número de observaciones

Regresión polinomial: formulación



Ejercicio

1. Genera los mejores modelo a partir de mínimo cuadrático para regresión de grado 1 hasta 6, de lso datos de prueba del laboratorio incluyendo un segundo predictor

   X_2 =[1.9214, 0.9160, 2.8463, 1.7907, 2.0481, 1.6341 , 0.7874, 1.9281, 1.1887, 1.8683)

2. Calcula el MSE de cada modelo

x2val_array = np.array([1.8158, 1.911, 1.8713, 2.0062, 2.141, 1.3796, 1.1813, 1.5382, 2.0696, 1.4034, 1.9903, 2.4266, 1.3875, 2.1728, 2.2521, 1.7206, 1.0702, 2.1807, 1.9983, 2.26])

3. Y cuál es el modelo generalizado