**Universidad Nacional Autónoma de México**

**Escuela Nacional de Estudios Superiores**

**Unidad Juriquilla**

**Computación II**

**Practica 8**

# Método de mínimos cuadrados para regresión polinomial


**Bruno Arturo López Pacheco**

**Dr. Ulises Olivares Pinto**

**No. De Cuenta 317347140**

**Licenciatura en Tecnología**

**Objetivo**

Utilizar el método de mínimos cuadrados para realizar una regresión de un conjunto de datos de entrada.

**Marco teórico**

**Método de mínimos cuadrados**

Es una técnica en la que dados un conjunto de pares ordenados (variable dependiente e independiente) y una familia de funciones, se intenta encontrar la función continua, dentro de dicha familia, que mejor se aproxime a los datos, de acuerdo con el criterio de mínimo error cuadrático.

Sea $(x_k, y_k)_{k=1}^{n}$ un conjunto de n puntos en el plano real, y sea $\{f_j(x)\}_{j=1}^{m}$ una base de m funciones linealmente independiente en un espacio de funciones. Se requiere encontrar una función f(x) que sea combinación lineal de las funciones base, de modo que $f(x_k) \approx y_k$, esto es

$$f(\mathrm{x}) = \sum_{\mathrm{j=1}}^{\mathrm{m}} c_j f_j(x)$$

Se trata de encontrar los m coeficientes $c_j$ que hagan que la función aproximante $f(x)$ dé la mejor aproximación para los puntos dados $(x_k, y_k)$.

La aproximación por mínimos cuadrados se basa en la minimización del error cuadrático medio, definido como:

$$E_c(f) = \frac{\sum_{k=1}^{n}(e_k)^2}{n}$$

El método consiste en resolver la siguiente matriz:

$$\begin{bmatrix}
n & \sum_{i=1}^{n} x & \sum_{i=1}^{n} x^2 & \sum_{i=1}^{n} x^3 & \cdots & \sum_{i=1}^{n} x^{m} & \sum_{i=1}^{n} Yi \\
\sum_{i=1}^{n} x & \sum_{i=1}^{n} x^2 & \sum_{i=1}^{n} x^3 & \sum_{i=1}^{n} x^4 & \cdots & \sum_{i=1}^{n} x^{m+1} & \sum_{i=1}^{n} xYi \\
\sum_{i=1}^{n} x^2 & \sum_{i=1}^{n} x^3 & \sum_{i=1}^{n} x^4 & \sum_{i=1}^{n} x^5 & \cdots & \sum_{i=1}^{n} x^{m+2} & \sum_{i=1}^{n} x^2Yi \\
\vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\
\sum_{i=1}^{n} x^{m} & \sum_{i=1}^{n} x^{m+1} & \sum_{i=1}^{n} x^{m+} & \sum_{i=1}^{n} x^{m+2} & \cdots & \sum_{i=1}^{n} x^{2m} & \sum_{i=1}^{n} x^{m}Yi
\end{bmatrix}$$

Donde:

- n = numero de puntos
- X = puntos de nuestra variable independiente
- Y = puntos de nuestra variable dependiente

**Metodología**

Los pasos para este método son:

- Formar una matriz de tamaño [m+1, m+2] donde m es el grado de la regresión.
- Se introducen los resultados numéricos de las sumas de los valores de X en la matriz.
- En la columna de resultados se introducen las sumas de los valores de X y Y.
- Se resuelve este sistema de ecuaciones lineales con el método de Gauss Jordan.
- La columna solución del sistema de ecuaciones ayuda a formar el polinomio de regresión, teniendo la forma:

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_mx^m$$

- Se calcula el error cuadrático con la ecuación descrita en la descripción del método.

**Pruebas**

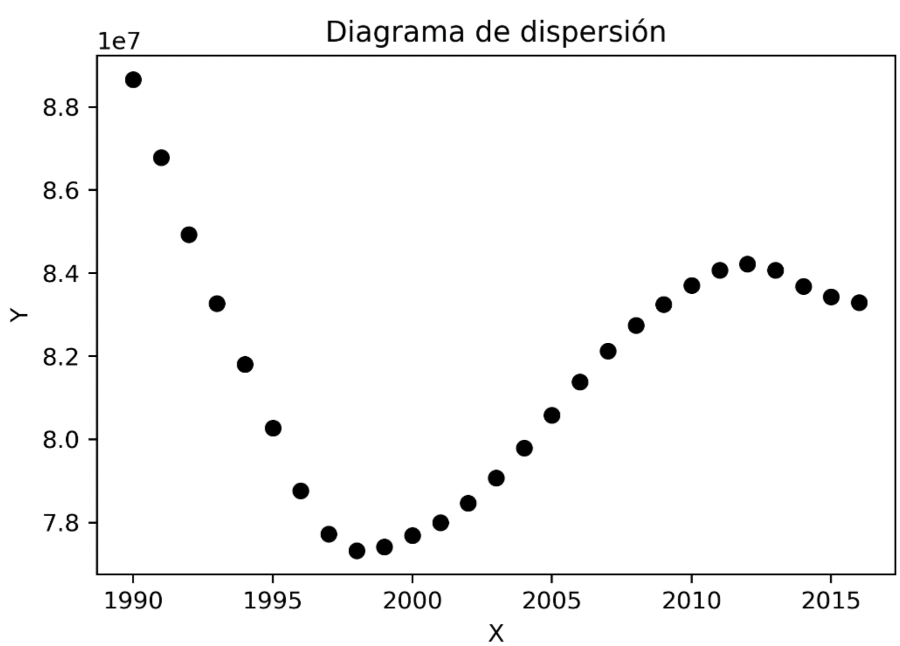
Se probó este método de regresión haciendo uso de un dataset obtenido del banco de datos "Our World in Data", del dataset "World Population Growth". Este dataset tiene información de la población por cada año. Se tomaron los datos desde el año 1990 hasta el 2016.

Se hizo la prueba con regresiones de grado 1, 2, 4, 6, 8 y 10.

También se recolectaron los errores cuadráticos medios asociados a dichas regresiones.

**Resultados**

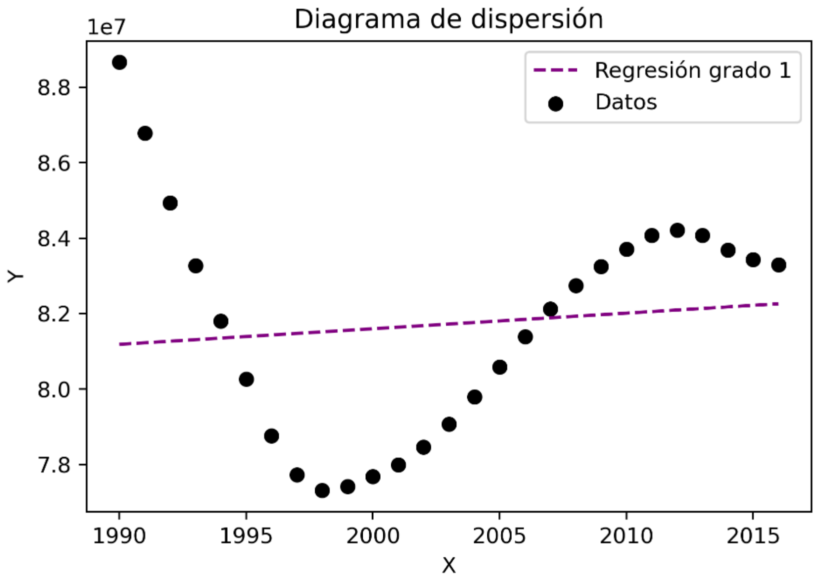
Dados los datos siguientes:



Se tienen las siguientes regresiones.

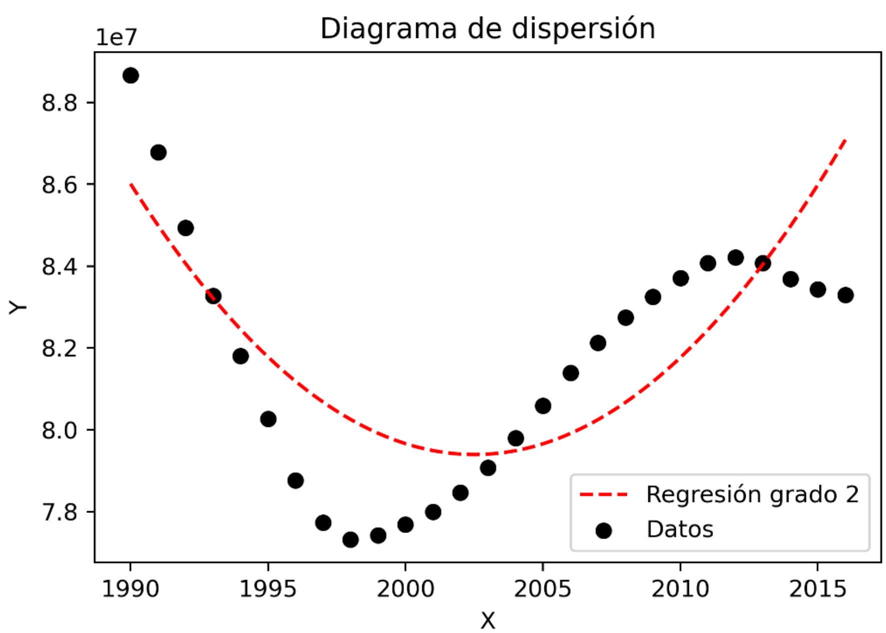
De grado 1, se tiene la regresión

$$f(x) = 41352.4884004884x - 1105416.45136344$$



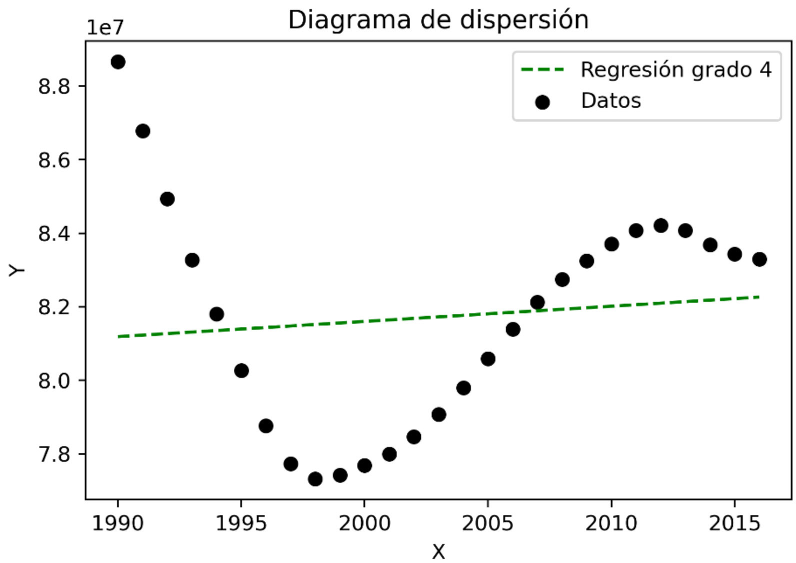

De grado 2, se tiene la regresión

$$f(x) = 42270.9263411674x^2 - 169295978.434378x + 169587912054.584$$



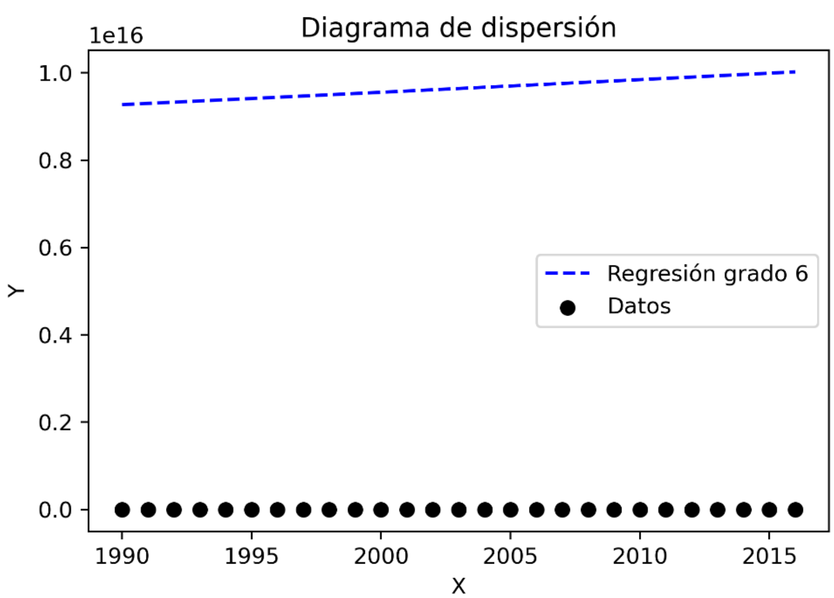
De grado 4, se tiene la regresión

$$f(x) = -1.90967100070718e^{-12}x^4 - 7.65787248372509e^{-9}x^3 - 2.29946511430185e^{-5}x^2 + 41352.7340748678x - 1105723.99787777$$



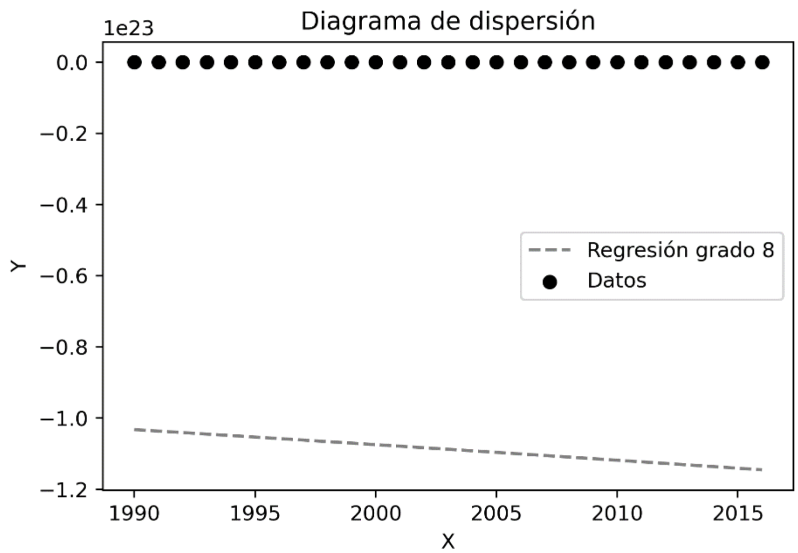

De grado 6, se tiene la regresión

$$f(x) = 0.000148551836650756x^6 + 0.00143274000229368x^5 + 0.0128511485761778x^4 + 0.113879471670699x^3 - 7.00696891636043x^2 + 2127.29358584939x + 109646421.958607$$



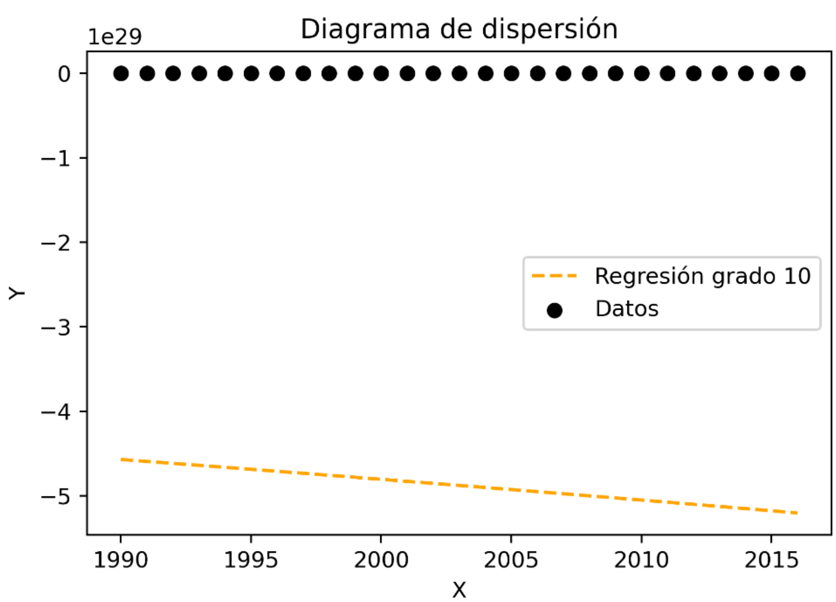
De grado 8, se tiene la regresión

$$f(x) = -0.000420143547406988x^8 + 0.000612573274132258x^7 + 0.000783360634030299x^6 + 0.0056522696626126x^5 + 0.0432570107562799x^4 + 0.385401588318926x^3 - 96.4338456388362x^2 + 186527.057960764x + 108760110.459845$$

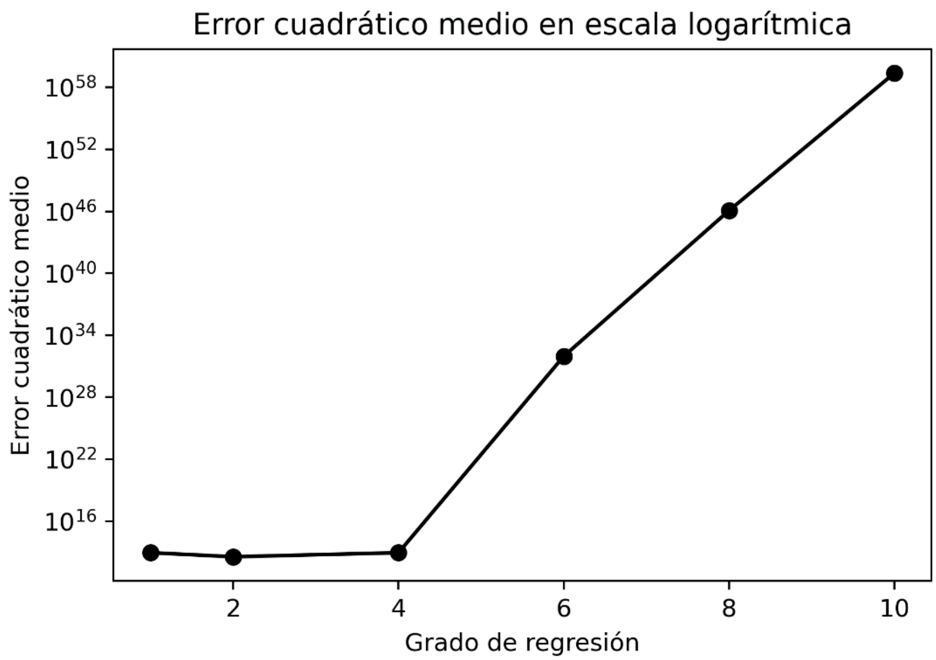



De grado 10, se tiene la regresión

$$f(x) = -0.000466685867808578x^10 - 0.00506525583413082x^9 - 0.00931294799763957x^8 + 0.00266785810325715x^7 + 0.00148131353003657x^6 + 0.0103991508876913x^5 + 0.0602982832957368x^4 + 0.615212993438354x^3 - 635.874902164347x^2 + 1988004.34578887x - 1327102244.03733$$



Por último, los errores asociados se graficaron en una escala logarítmica, pues los errores crecieron al aumentar el grado del polinomio para la regresión. Se obtuvo la siguiente gráfica.




**Conclusión**

Este método de los mínimos cuadrados es de rápida implementación pues no utiliza funciones base como en otros métodos de regresión. Aparte, utiliza métodos conocidos como la eliminación por Gauss Jordan.

La eficiencia de este método también depende de la naturaleza de los datos, pues antes de utilizar este dataset, y era aun mayor el error asociado a esa interpolación.

**Referencias**


Chapra, S., & Canale, R. (2011). METODOS NUMERICOS PARA INGENIEROS (Spanish Edition) (6th ed.). McGraw-Hill Interamericana de España S.L.

Datos recabados de: Our World in Data https://ourworldindata.org/world-population-growth