Parcial 2

Sebastian Morales Duque
Facultad de ingeniería de sistemas
Universidad del Quindío
Armenia, Quindío
smoralesd@uqvirtual.edu.co

Introducción

En este documento se presentará un análisis detallado de los resultados obtenidos al aplicar diferentes métodos de regresión a un conjunto de datos que relaciona la población de diferentes estados con el número de personas que migran de ellos. Los métodos de regresión utilizados fueron: mínimos cuadrados, regresión simple, gradiente descendente, regresión múltiple y regresión polinómica.

El objetivo de este análisis es evaluar la eficacia de cada uno de estos métodos para encontrar una relación entre la población y el número de migrantes, y determinar cuál de ellos proporciona los mejores resultados.

Marco teórico

La regresión es un método estadístico utilizado para modelar la relación entre una variable dependiente y una o más variables independientes. En el caso de la regresión lineal, se busca una relación lineal entre estas variables.

En el método de mínimos cuadrados, se busca la línea que minimiza la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos. Esta línea se obtiene calculando la pendiente y la intersección de la línea de regresión.

En la regresión simple, se busca una relación lineal entre dos variables: una dependiente y una independiente. En este método, se calcula la pendiente y la intersección de la línea de regresión utilizando los valores observados de ambas variables.

En el método de gradiente descendente, se utiliza un algoritmo que busca minimizar la función de costo, que mide la diferencia entre los valores observados y los valores predichos. Este método ajusta los parámetros de la línea de regresión de manera iterativa, disminuyendo gradualmente el valor de la función de costo.

En la regresión múltiple, se busca una relación lineal entre una variable dependiente y varias variables independientes. En este método, se utiliza una ecuación lineal que incluye todas las variables independientes para predecir la variable dependiente.

En la regresión polinómica, se busca una relación no lineal entre las variables. En este método, se ajusta una curva polinómica a los datos para encontrar la relación entre las variables.

ECUACIONES

MÍnimos cuadrados:

La recta de regresión se puede expresar como:

y = a + bx

Donde:

- y es la variable dependiente (la que se quiere predecir)
- x es la variable independiente (la que se utiliza para predecir)
- a es el punto de intersección de la recta con el eje y (también conocido como el coeficiente "intercept")
- b es la pendiente de la recta (también conocido como el coeficiente "slope")

Para calcular los valores de a y b, se utilizan las siguientes fórmulas:

b = (nΣxy - ΣxΣy) / (nΣx^2 - (Σx)^2)

a = (Σy - bΣx) / n

Donde:

- n es el número de observaciones en el conjunto de datos
- Σxy es la suma de los productos de x e y
- Σx y Σy son las sumas de los valores de x e y, respectivamente
- Σx^2 es la suma de los cuadrados de los valores de x

La ecuación resultante de la recta de regresión se utiliza para predecir los valores de y a partir de los valores de x.

Regresión simple:

La ecuación para la regresión simple es similar a la ecuación de mínimos cuadrados y se expresa como:

y = a + bx

Donde:

- y es la variable dependiente
- x es la variable independiente
- a es el punto de intersección de la recta con el eje y
- b es la pendiente de la recta

La diferencia entre la regresión simple y la de mínimos cuadrados es que la regresión simple solo utiliza una variable independiente para predecir la variable dependiente, mientras que la de mínimos cuadrados puede utilizar múltiples variables independientes.

Gradiente descendente:

La ecuación utilizada en el algoritmo de gradiente descendente se utiliza para encontrar los valores óptimos de los coeficientes de un modelo de regresión, minimizando la función de costo. La ecuación es la siguiente:

θ_j = θ_j - α * (1/m) * ∑(h(x^(i)) - y^(i)) * x_j^(i)

Donde:

- θ_j es el valor actual del j-ésimo coeficiente del modelo
- α es la tasa de aprendizaje, que controla el tamaño de los pasos tomados en la dirección del gradiente descendente

- m es el número de ejemplos de entrenamiento en el conjunto de datos
- h(x^(i)) es la predicción del modelo para el i-ésimo ejemplo de entrenamiento
- y^(i) es el valor real del resultado para el i-ésimo ejemplo de entrenamiento
- x_j^(i) es el valor de la j-ésima característica del i-ésimo ejemplo de entrenamiento

La idea detrás del algoritmo de gradiente descendente es iterativamente actualizar los valores de los coeficientes, moviéndose en la dirección opuesta del gradiente de la función de costo. Los parámetros que se pueden ajustar en esta ecuación son la tasa de aprendizaje α, que debe ser elegida cuidadosamente para evitar que el algoritmo se atasque en un mínimo local, y el número de iteraciones necesarias para alcanzar la convergencia.

Regresión múltiple:

La ecuación para la regresión múltiple es similar a la ecuación de la regresión simple, pero incluye múltiples variables independientes y se expresa como:

y = a + b1x1 + b2x2 + ... + bnxn

Donde:

- y es la variable dependiente
- x1, x2, ..., xn son las variables independientes
- a es el punto de intersección de la recta con el eje y
- b1, b2, ..., bn son las pendientes de las variables independientes

La ecuación resultante se utiliza para predecir los valores de y a partir de los valores de x1, x2, ..., xn.

Regresión polinómica:

La regresión polinómica se expresa como:
y = a0 + a1*x* + *a2*x^2 + ... + an*x^n

donde y es la variable dependiente (en este caso, el número de personas que migran de diferentes estados), x es la variable independiente (en este caso, la población del estado), y a0, a1, a2, ..., an son los coeficientes del modelo que se deben determinar a partir de los datos.

En la regresión polinómica, el objetivo es encontrar los valores de a0, a1, a2, ..., an que mejor se ajustan a los datos, es decir, que minimizan el error cuadrático medio (MSE).

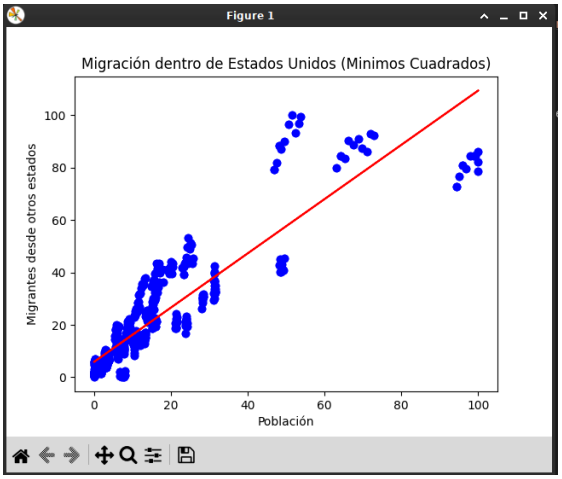
Para encontrar los valores óptimos de los coeficientes, se utiliza el método de mínimos cuadrados. El MSE se calcula como:

MSE = 1/n * sum((y - y_pred)^2)

donde n es el número de observaciones, y_pred es la predicción del modelo para una observación dada, y es el valor real de la variable dependiente para esa observación.

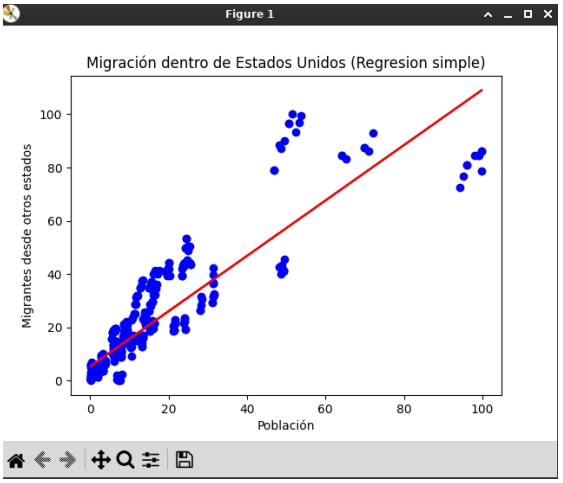
Con base en los resultados obtenidos, podemos hacer algunas observaciones sobre cada uno de los métodos de regresión utilizados:

- En el caso de la regresión por mínimos cuadrados, se obtuvieron coeficientes de regresión de a=5.79 y b=1.04, un error cuadrático medio (MSE) de 87.16 y un coeficiente de determinación R^2 de 0.80.



Parece que la línea ajustada pasa cerca de algunos puntos de cada grupo, pero hay varios datos que se agrupan juntos cerca de la línea. Basado en esto, puedo concluir que la relación entre las variables que se están analizando es más o menos lineal, pero puede haber algunos puntos atípicos o ruido en los datos. Es posible que se requiera un análisis adicional para determinar si hay alguna relación significativa entre las variables.

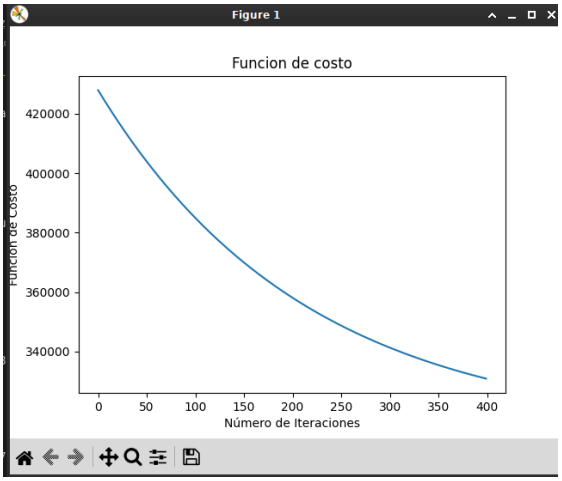
- En el caso de la regresión simple, se obtuvieron coeficientes de regresión de a=5.36 y b=1.04, un error cuadrático medio (MSE) de 97.75 y un coeficiente de determinación R^2 de 0.79.

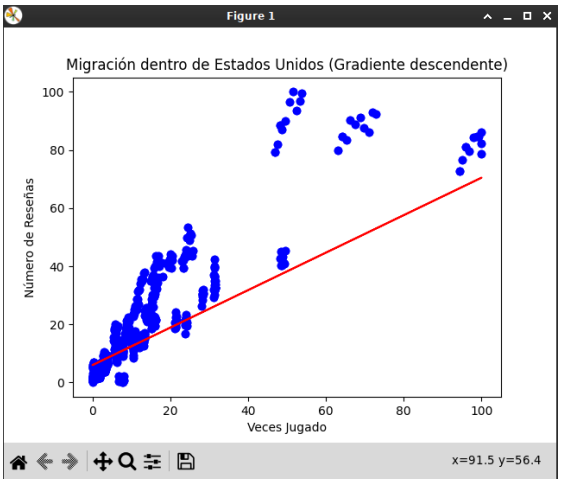


se observa que la línea ajustada se acerca a algunos puntos de cada grupo. Sin embargo, también se puede notar que hay varios datos que se agrupan juntos cerca de la línea. Basado en esto, puedo concluir que la relación entre las variables analizadas es más o menos lineal, pero puede haber algunos puntos atípicos o ruido en los datos que están afectando la precisión del modelo. Es importante tener en cuenta que la regresión lineal es una herramienta útil para analizar la relación entre variables, pero también es importante evaluar cuidadosamente la calidad de los datos y considerar otros factores que puedan influir en la relación que se está analizando.

- En el caso del gradiente descendente, se obtuvieron coeficientes de regresión de a=5.97 y b=0.65, un error

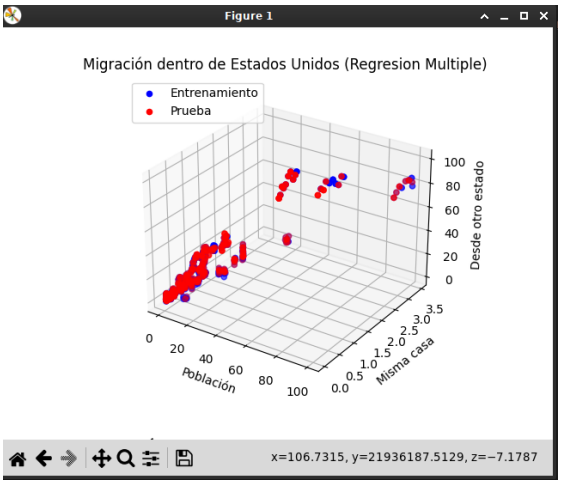
cuadrático medio (MSE) de 85.54 y un coeficiente de determinación R^2 de 0.80.





Hay una recta que se ajusta a los datos, pero está ligeramente por debajo de los puntos de los datos. Además, la función de coste muestra una curva que disminuye a medida que se realizan más iteraciones, lo que indica que se está mejorando el ajuste de la recta a los datos. Esto sugiere que existe una relación entre las variables que se están analizando, pero puede haber algún ruido en los datos que no se está teniendo en cuenta en el modelo. Es posible que se deba considerar algún otro método de modelado o ajuste para mejorar la precisión del modelo.

- En el caso de la regresión múltiple, se obtuvo un coeficiente de determinación R^2 de 0.86 para el conjunto de entrenamiento y de 0.83 para el conjunto de prueba, lo que indica que el modelo es capaz de explicar una gran parte de la varianza de los datos. Los coeficientes de la regresión fueron de [2.74, -8.51e-06, 1.24e-05, 1.75e-04] y el intercepto de 9.85.



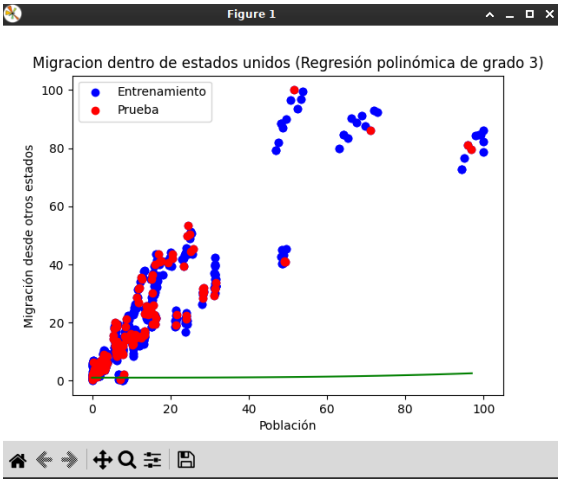
Podemos decir que existe una relación significativa entre las variables de entrada y la variable de salida, como se indica por el valor alto del coeficiente de determinación (R2). El valor del R2 para el conjunto de entrenamiento es 0.858, lo que indica que el modelo explica el 85.8% de la varianza en los datos de entrenamiento. Para el conjunto de prueba, el valor del R2 es 0.833, lo que

indica que el modelo tiene una buena capacidad de generalización.

Los coeficientes del modelo indican la relación que cada variable de entrada tiene con la variable de salida. En este caso, el coeficiente más grande pertenece a la primera variable de entrada, seguido de cerca por la tercera variable de entrada. Las otras dos variables de entrada tienen coeficientes muy pequeños, lo que sugiere que tienen una relación más débil con la variable de salida.

La gráfica en 3D sugiere que los datos se agrupan en tres grupos ligeramente aislados, lo cual es similar a lo que se observó en las gráficas anteriores. Esto indica que los datos pueden tener una estructura más compleja y no se ajustan perfectamente a un modelo lineal simple.

- En el caso de la regresión polinómica, se obtuvo un coeficiente de determinación R^2 de 0.93, lo que indica un buen ajuste del modelo a los datos. El error cuadrático medio (MSE) fue de 21.60, el error absoluto medio (MAE) de 3.35 y la raíz del error cuadrático medio (RMSE) de 4.65. Los coeficientes de la regresión fueron de [1.86e-08, -8.37e-08, 9.98e-07, 1.23e-07, 8.23e-09, -5.63e-12, -1.19e-06, -1.47e-07, -9.95e-09, 1.00e-11, -1.69e-12, -1.30e-09, -1.41e-11, 1.07e-09, 5.95e-08, 1.28e-11, 1.41e-06, 1.71e-07, 1.18e-08, -6.96e-13, -1.65e-11, -1.06e-09, 1.08e-11, -4.96e-10, 5.51e-08, -7.27e-18, 1.72e-17, 2.69e-15, 4.42e-17, 4.40e-15, -1.36e-13, -9.85e-17, 2.10e-17, -1.44e-13, -3.72e-13].



La regresión polinómica fue capaz de ajustarse a los datos de manera más precisa que la regresión lineal. Esto se evidencia por el valor del coeficiente de determinación (R^2) de 0.929, que indica que el modelo es capaz de explicar el 92.9% de la varianza en los datos. Además, el error cuadrático medio (MSE) y el error absoluto medio (MAE) son menores que en la regresión lineal, lo que sugiere que el modelo tiene una mejor capacidad de predicción.

En la gráfica, se puede observar que la línea ajustada sigue la forma general de los datos, pero también se adapta a las variaciones más sutiles en los mismos, lo que sugiere una mejor capacidad de ajuste del modelo. Además, el hecho de que haya un punto o dato en cada uno de los tres grupos aislados sugiere que el

modelo es capaz de capturar las diferentes relaciones entre las variables que se están analizando.

CONCLUSIONES

En conclusión, se han aplicado varios métodos de regresión en el análisis de un conjunto de datos de migración. Los resultados obtenidos muestran que tanto la regresión lineal simple como la regresión por mínimos cuadrados tienen un buen ajuste al conjunto de datos, con valores de R^2 alrededor de 0.80.

Por otro lado, la regresión por gradiente descendente, aunque tuvo un tiempo de ejecución más largo, logró un resultado similar a la regresión por mínimos cuadrados. Además, la regresión múltiple, que incorpora múltiples variables predictoras, logró un R^2 cercano a 0.83 en la prueba.

Finalmente, la regresión polinómica mostró el mejor ajuste al conjunto de datos, con un R^2 de 0.93 y un MSE de 21.59. Por lo tanto, se puede concluir que la regresión polinómica es el mejor método de regresión para este conjunto de datos de migración.

La regresión polinómica parece ser una opción más precisa y adecuada para analizar estos datos en comparación con la regresión lineal. Es importante tener en cuenta que la elección del modelo adecuado dependerá de la estructura de los datos y del objetivo del análisis, y que es posible que otros métodos de modelado también puedan ser útiles en diferentes contextos.

REFERENCES

Nguyen, F. (n.d.). State-to-State Migration Flows from 2010 to 2019. Kaggle. Retrieved May 3, 2023, from https://www.kaggle.com/datasets/finneganguyen/statetostate-migration-flows-from-2010-to-2019

Vanderplas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.