Laboratorio 1: Regresion Lineal

Tomas Lopez Pérez

April 2024

1 Introducción

La regresión lineal es un método esencial en estadística que permite analizar la variabilidad de una variable en relación con una o más variables predictoras [Ped97]. En este trabajo, se explora la implementación de la regresión lineal en Python, utilizando diversas instancias para observar el comportamiento del algoritmo. La elección de Python como entorno de desarrollo ofrece una plataforma versátil y accesible para realizar análisis estadísticos, incluida la regresión lineal. Al explorar diferentes instancias de regresión lineal en Python, podemos examinar cómo el algoritmo se adapta a distintos conjuntos de datos . En resumen, este trabajo ofrece una exploración detallada de la implementación de la regresión lineal en Python, destacando su utilidad y aplicabilidad en la investigación.

2 Fundamento teórico

2.1 Regresion lineal

La regresión se refiere a la relación entre dos variables donde una variable se regresa o se predice a partir de la otra. Esto se logra mediante una línea de regresión que se ajusta a los valores observados. La regresión se utiliza para identificar relaciones causales o predecir una variable basada en otra. Mientras que algunas relaciones son deterministas y predecibles, como la relación entre la presión y el volumen de un gas a temperatura constante, en fenómenos biológicos, la relación es aleatoria debido a la influencia de múltiples factores desconocidos. En un contexto determinista, los resultados se alinean estrechamente con la línea de regresión debido al error de medición, mientras que en un contexto aleatorio, se agrega variabilidad adicional introducida por múltiples factores. [Dag+14]

2.2 Diagrama de dispersión

Previo a cualquier análisis, es crucial realizar una primera inspección visual de los datos mediante un diagrama de dispersión o nube de puntos. Esta representación conjunta de los datos nos permite evaluar si es apropiado utilizar un modelo de regresión simple. Al observar el diagrama de dispersión, podemos determinar varios aspectos fundamentales de manera general: [CLA06]

1. La existencia de una relación entre las variables.
2. La naturaleza de esta relación, si es lineal o no.
3. El grado de concentración de los puntos, lo cual indica la fuerza de la relación.
4. La detección de valores atípicos que podrían distorsionar la relación potencial.
5. La uniformidad de la dispersión de los datos a lo largo de la nube de puntos.

Esta inspección visual inicial proporciona información valiosa que guía el proceso de análisis posterior y ayuda a tomar decisiones informadas sobre el tipo de modelo de regresión más adecuado para los datos disponibles.

2.3 Error Cuadrático Medio

Error cuadrático medio (MSE), [Mar86] la diferencia cuadrática promedio entre el valor observado en un

estudio estadístico y los valores predichos a partir de un modelo. Al comparar observaciones con valores predichos, es necesario elevar al cuadrado las diferencias, ya que algunos valores de datos serán mayores que la predicción (y por lo tanto sus diferencias serán positivas) y otros serán menores (y por lo tanto sus diferencias serán negativas). Dado que es tan probable que las observaciones sean mayores que los valores predichos como menores, las diferencias sumarían cero. Cuadrar estas diferencias elimina esta situación. La fórmula para el error cuadrático medio es figura 1 donde yi es el i-ésimo valor observado, pi es el valor predicho correspondiente para yi y n es el número de observaciones. La Σ indica que se realiza una suma sobre todos los valores de i.

$$\text{MSE} = \frac{\Sigma(y_i - p_i)^2}{n}$$

Figure 1: MSE

3 Metodología

En esta sección, describimos los pasos seguidos para explorar la implementación de la regresión lineal en Python como ejemplo , así como los métodos utilizados para evaluar su desempeño.

3.1 Selección de datos de prueba

Para este estudio, seleccionamos datos de los metros cuadrados de varias casas y sus respectivos precios, estos valores fueron elegidos al azar

En la Tabla 1, se presentan los datos de entrenamiento que serán utilizados en el experimento 1. Estos datos son relevantes para la evaluación del primer modelo. Adicionalmente, se hace uso de los datos proporcionados en la Tabla 2, los cuales contienen información sobre los pesos ww y bb.

3.2 Preprocesamiento de datos

Antes de realizar el análisis de regresión lineal, se realiza el escalamiento de los datos para tenerlos en un rango comun.

3.3 Implementación del modelo de regresión lineal

Utilizamos Python para implementar el modelo de regresión lineal. Ajustamos el modelo a cada conjunto de datos seleccionado utilizando la función de regresión lineal y ajustamos los hiperparámetros según fuera necesario.

3.4 Experimentación y análisis de resultados

Prediga el precio de la propiedad con 1320 metros cuadrados para esto realizamos experimentos con diferentes instancias de regresión lineal en Python utilizando los conjuntos de datos seleccionados. Analizamos los resultados de cada experimento para evaluar cómo el algoritmo se adaptó a diferentes instancias. Además, utilizamos técnicas de visualización de datos para explorar la relación entre las variables predictoras y la variable objetivo.

4 Resultado

Para los experimentos se genero una arreglo de datos x_train que es el tamaño en m^2 y y_train que es el costo en millones como podemos observar en la figura **??**. Tambien se definen w y b en la figura **??** .

4.1 Experimento 1

En este experimento se realiza el ajuste de los datos w y b para adaptar nuestro modelo al conjunto de entrenamiento. En la Figura 3, con $w = 150$ y $b = 300$, se observa que el modelo aún tiene un error considerable, aunque está más próximo al conjunto de entrenamiento. Sin embargo, al utilizar $w = 400$ y $b = 200$, como se muestra en la Figura 4, el modelo se aleja significativamente del conjunto de entrenamiento. Este comportamiento se repite en las Figuras 5 con $w = 500$ y $b = 400$, así como en la Figura 6 con $w = 505$ y $b = 500$. Se puede observar cómo

las líneas azules se acercan o se alejan de los datos de entrenamiento según los valores de w y b.

Aplicando la métrica del Error Cuadrático Medio (MSE), podemos determinar los mejores valores de w y b, como se muestra en la Figura 2, donde se encuentra el menor error, 32600, con $w = 100$ y $b = 100$. Utilizando este modelo, podemos predecir el costo para una propiedad de 1.32 metros cuadrados, lo que resulta en un precio estimado de 232.

index	Size (1000 m^2)	Price (10000s de pesos)
0	1.0	300
1	2.0	500
2	3.0	600
3	4.0	660
4	5.0	800
5	6.0	900

Table 1: Datos de tamaño y precio

index	W	b
0	100	100
1	150	300
2	400	200
3	500	400
4	505	500

Table 2: Datos para w y b

4.2 Experimento 2

El experimento realiza los mismos pasos que el experimento 1, con la diferencia de que se generan nuevos datos de entrenamiento al azar, así como valores aleatorios para w y b. En las Figuras 7 y 8, con $w = 150$ y $w = 100$ respectivamente, se observa que el modelo está muy lejos del conjunto de entrenamiento. Esto se aprecia en la inclinación de la línea, que no coincide con los datos. En las Figuras 9, 10, y 11, se puede observar cómo el modelo se acerca gradualmente a los datos de entrenamiento. Aplicando la métrica MSE, se obtiene que el modelo de la Figura 9 tiene un error menor de 13303.603, en comparación con los demás modelos. Esto se evidencia en la posición de la línea azul, que se encuentra más cerca del conjunto de entrenamiento.

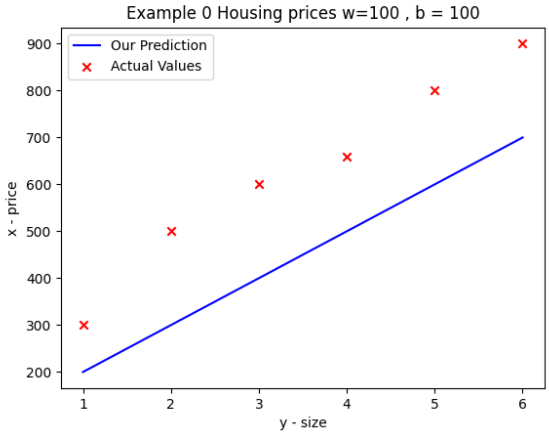

Figure 2: Example 0

Como resultado de este nuevo experimento, se genera una predicción para el costo de una propiedad de 1.32 metros cuadrados, la cual es de 728.

5 Conclusiones

Durante los experimentos realizados, se ajustaron los parámetros ww y bb en un modelo de regresión lineal, y se observó que variar estos parámetros tuvo un impacto directo en la forma de la línea de regresión. Valores más grandes de ww resultaron en una línea más inclinada, mientras que valores más pequeños la hicieron menos inclinada. Por otro lado, el parámetro bb determinó cuánto se desplazó la línea verticalmente en el eje de las yy. La elección cuidadosa de estos parámetros fue crucial para minimizar el error y lograr un buen ajuste a los datos de entrenamiento.

Además, se observó que el número de datos de entrenamiento afectó la capacidad del modelo para aproximar la línea de regresión. En el experimento 1, con un conjunto de datos más pequeño, la aproximación de la línea fue más difícil, como se observó en la gráfica con solo dos datos. En contraste, en el experimento 2, con un mayor número de datos, la línea

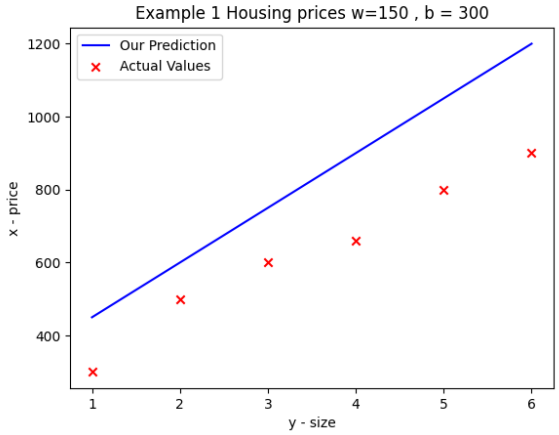

Figure 3: Example 1

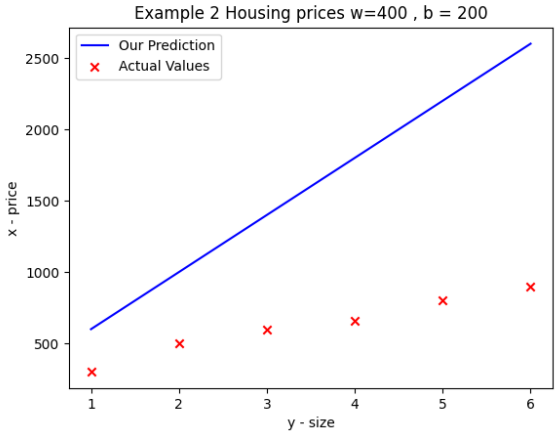

Figure 4: Example 2

de regresión pudo ajustarse mejor a los datos, como se vio en las gráficas con dos y tres puntos respectivamente.

Una vez obtenidos los valores de ww y bb con el menor error, se realizaron predicciones para nuevos valores. El precio predicho para una propiedad de 1.321 metros cuadrados fue 232 en el experimento 1 y 728 en el experimento 2. Es importante destacar que estos valores pueden variar debido a la generación aleatoria de datos.

En conclusión, los experimentos resaltan la importancia de seleccionar cuidadosamente los parámetros del modelo de regresión lineal y la influencia del tamaño del conjunto de datos en la precisión de las predicciones. Además, se demuestra que un mayor número de datos de entrenamiento puede mejorar la capacidad del modelo para ajustarse a los datos observados.

References

[Mar86] Hans Marmolin. "Subjective MSE Measures". In: *IEEE Transactions on Systems, Man, and Cybernetics* 16.3 (1986), pp. 486–489. DOI: 10.1109/TSMC.1986.4308985.

[Ped97] Elazar J Pedhazur. *Multiple regression in behavioral research: Explanation and prediction*. Fort Worth: Harcourt Brace College Publishers, 1997.

[CLA06] Carlos Camacho, A López, and M Arias. "Regresión lineal simple". In: *Documento inédito. Recuperado de http://personal.us. es/vararey/adatos2/Regsimple. pdf* (2006).

[Dag+14] Jorge Dagnino et al. "Regresión lineal". In: *Rev. Chil. Anest* 43.2 (2014).

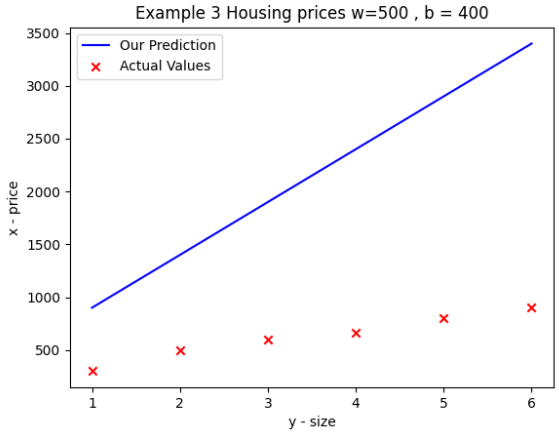

Figure 5: Example 3

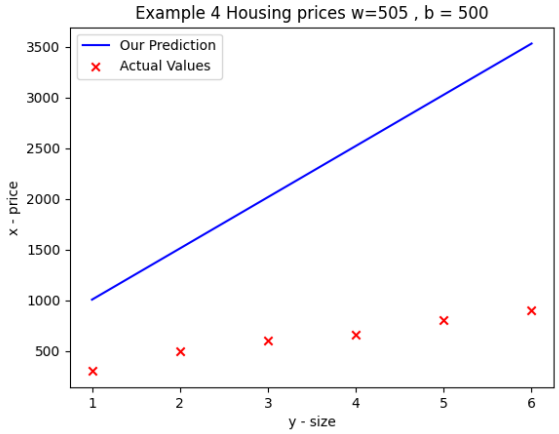

Figure 6: Example 4

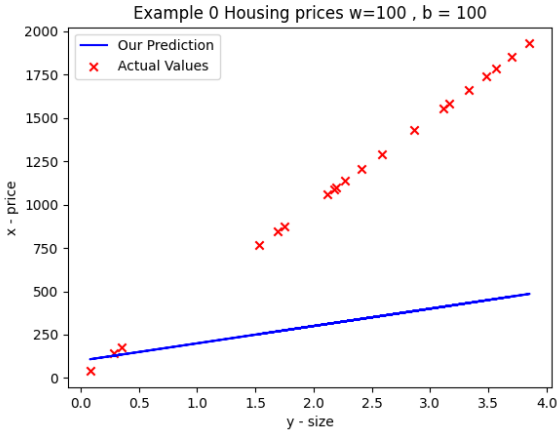

Figure 7: Example 5

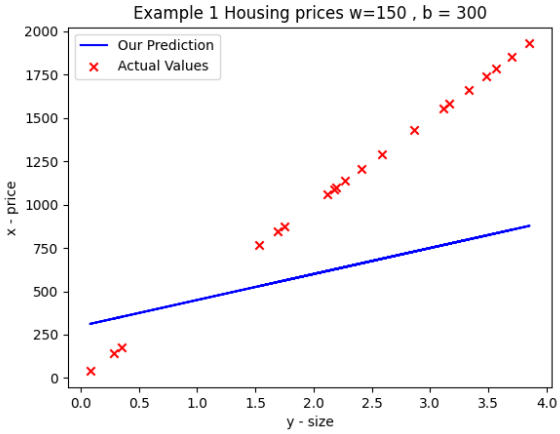

Figure 8: Example 6

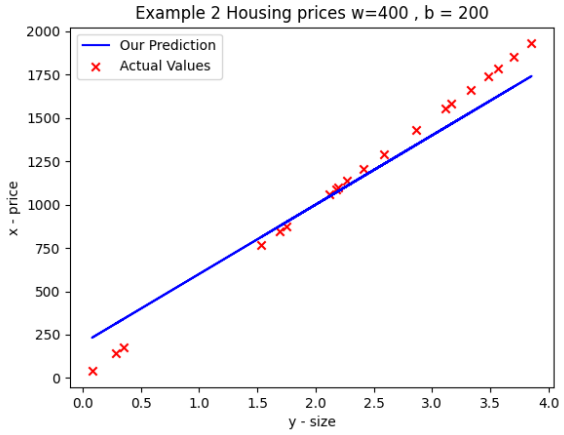

Figure 9: Example 7

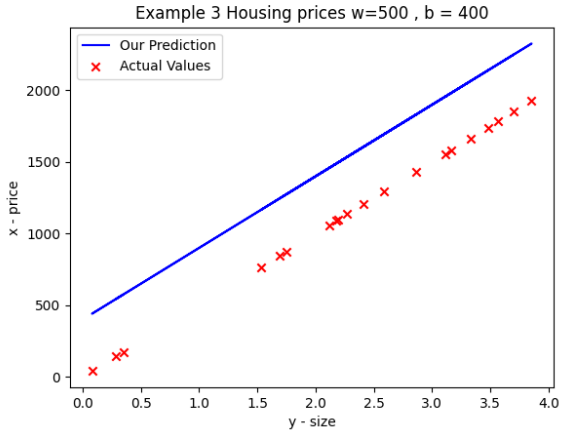

Figure 10: Example 8

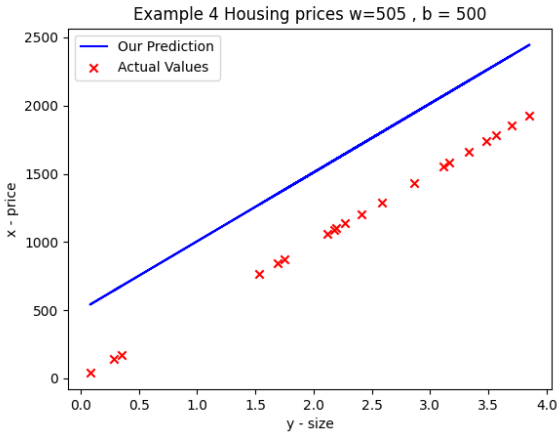

Figure 11: Ejemplo 9