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

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 2. Tambien se definen w y b en la figura 3 .

4.1 Experimento 1

En las figuras 4, 5, 6, 7 y 8, podemos 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 encontrar los mejores valores de w y b, lo que nos permitirá predecir el costo para una propiedad de 1.32 metros cuadrados. Esta predicción se muestra en la figura 9.

4.2 Experimento 2

Se genera un nuevo experimento mas datos al azar figura 10, 11, 12, 13 y obteniendo los mejore w y b predecimos el costo para una propiedad de 1.32 metros cuadrados figura 14

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

Figure 2: Datos

5 Conclusiones

Durante los experimentos realizados, se modificaron los parámetros (w) y (b) en un modelo de regresión lineal. Observamos que variar estos parámetros tiene un impacto directo en el modelo de regresion lineal, valores más grandes de (w) hacen que la línea sea más inclinada, mientras que valores más pequeños la hacen menos inclinada. Por otro lado, el parámetro (b) determina cuánto se desplaza la línea verticalmente en el eje de las (y). La elección cuidadosa de estos parámetros es crucial para minimizar el error y lograr un buen ajuste a los datos de entrenamiento.

De igual manera podemos observar en las graficas que entre menos datos de entrenamiento ocupemos, es mas complicado aproximar la recta, Como lo vemos en la grafica con dos datos del primer experiemnto , pero si agremas mucho mas datos es mas facil minimizar el error como lo vemos en el experiemnto 2 en donde la recta pasa por encima de los dos y tres puntos respectivamente.

Entonces, una vez que obtuvimos w y b con el que consideramos el menor error podemos predecir un nuevo valor El precio predicho para una propiedad de [1.32] metros cuadrados es: [232.] en el experimento y en el segundo [239.04] este valor va a varias ya que estamos genera datos al azar, por el cual podemo concluir que ambos en ambos experimentos nos prorpociona una prediccion similar, aunque es posible encontrar un w y b que nos de algun menor error.

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

Figure 3: W y B

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).

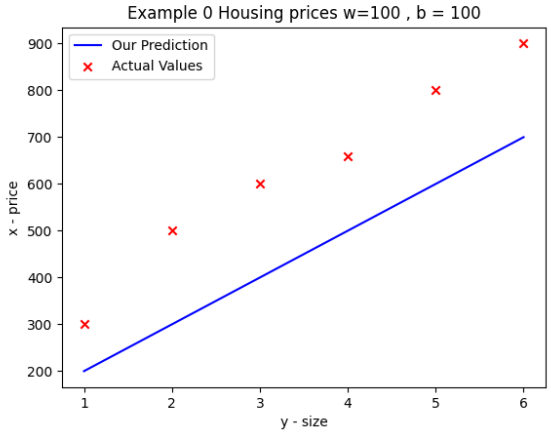

Figure 4: Example 0

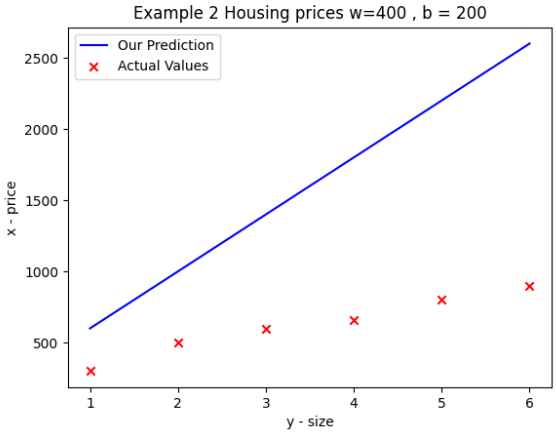

Figure 6: Example 2

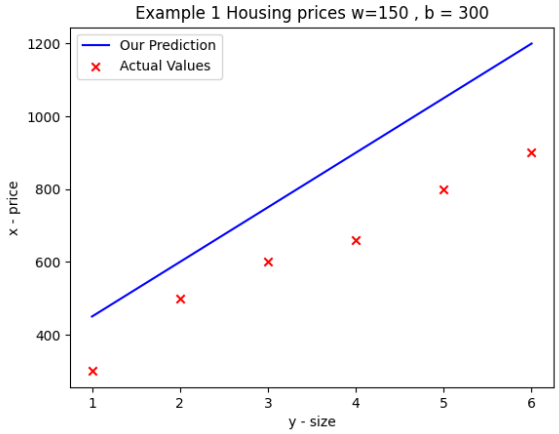

Figure 5: Example 1

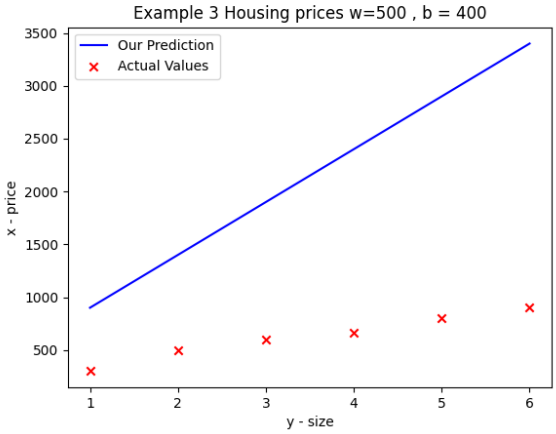

Figure 7: Example 3

Figure 8: Example 4

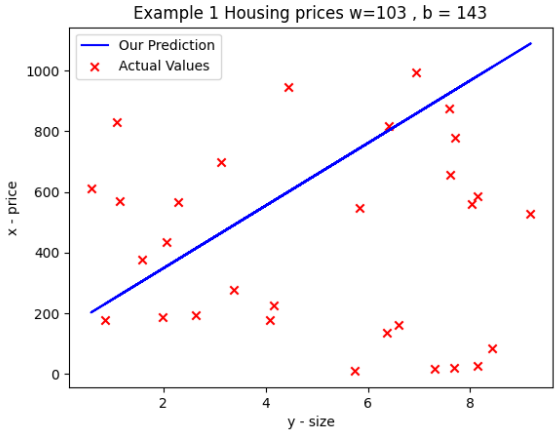

Figure 11: Example 6

```
El mejor valor de w: 100
El mejor valor de b: 100
MSE correspondiente: 32600.0
El precio predicho para una propiedad de [1.32] metros cuadrados es: [232.]
```

Figure 9: El mejor w y b

Figure 10: Example 5

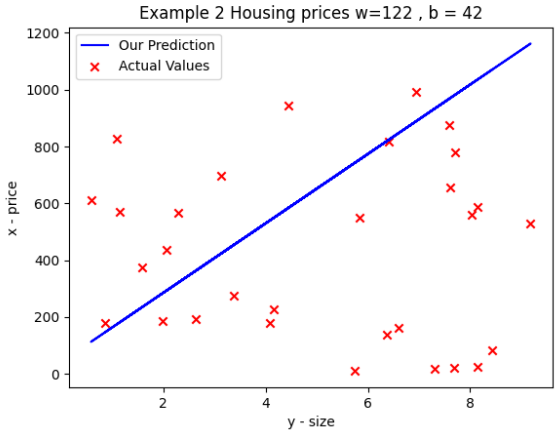

Figure 12: Example 7

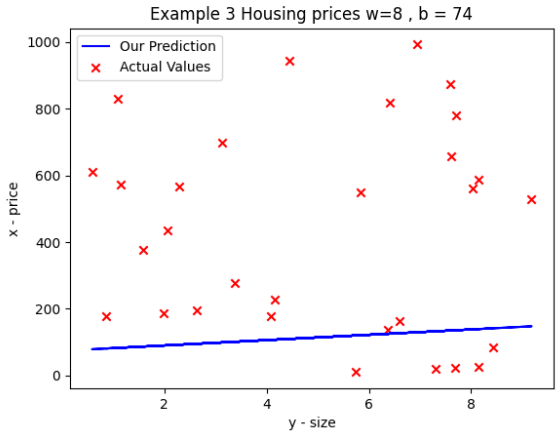

Figure 13: Example 8

```
El mejor valor de w: 97
El mejor valor de b: 111
MSE correspondiente: 192364.50178451175
El precio predicho para una propiedad de [1.32] metros cuadrados es: [239.04]
```

Figure 14: El mejor w y b