Práctica 1

01/03/2022

César Borja Moreno 800675

Aprendizaje Automático

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

Índice

1. Trabajo previo

```
function [J, grad] = CosteL2 (theta, X, y)
% Calcula el coste cuadratico y su gradiente
    r = X*theta - y;
    J = (1/2)*r'*r;
    grad = X'*r;
end
```

```
X = [1 ..., 1 ..., ...]    % muestras de entrenamiento
theta = [1, 1, .., 1]       % valores iniciales de θ
y = [...]                   % salidad deseadas
alpha = 0.1                 % factor de aprendizaje

[J1, grad] = CosteL2 (theta, X, y)
theta = theta - alpha*grad
[J2, grad] = CosteL2 (theta, X, y)
while (J2 < J1)
    J1 = J2
    theta = theta - alpha*grad
    [J2, grad] = CosteL2 (theta, X, y)
end
```

2. Regresión monovariable utilizando ecuación normal

Se busca encontrar los mejores valores para θ utilizando la técnica de ecuación normal con una sola variable (superficie m^2). Esto se consigue de la siguiente manera en Matlab:

$$\theta = X \setminus y$$

Siendo X la matriz de valores de entrada e y el vector de salidas de dichos datos, ambos extraídos del fichero PisosTrain.txt.

La recta de predicción obtenida es la siguiente:

$$-34313.387972 + 2608.209435 * x1$$

En la Figura 2 se muestra la recta de predicción junto con los puntos de entrenamiento:

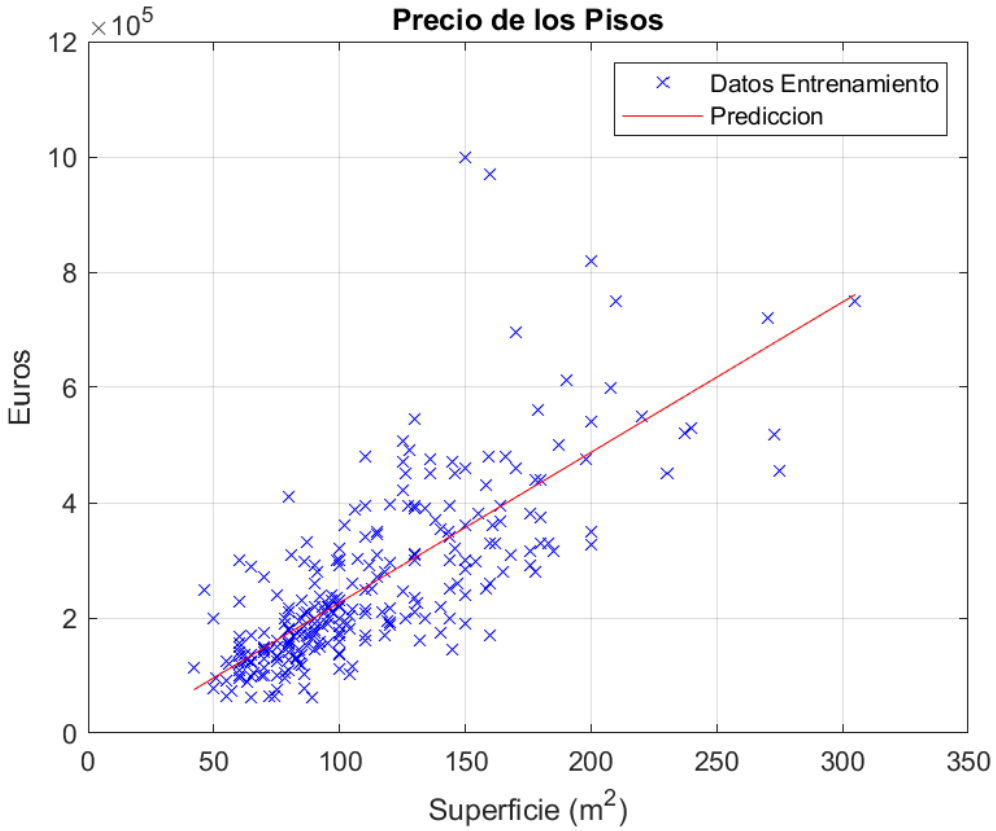

Figura 2. Recta de predicción junto con datos de entrenamiento

Los puntos de test junto con la recta de predicción se muestran en la Figura 3.

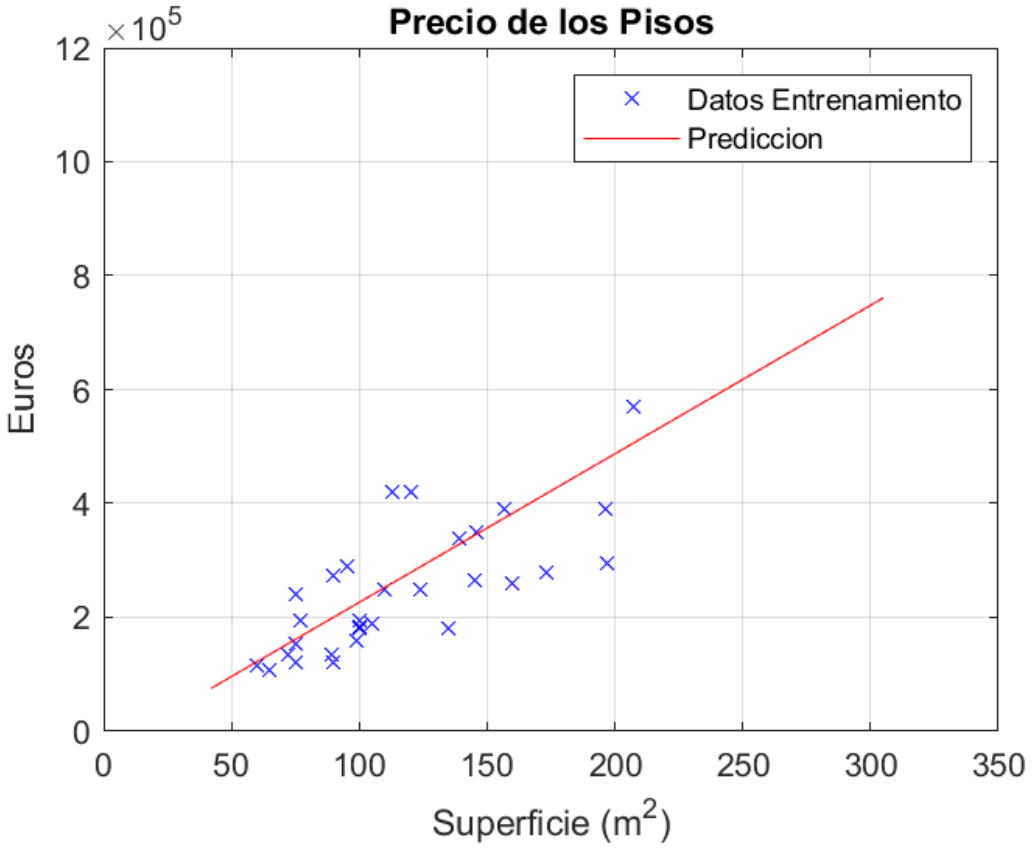

Figura 3. Recta de predicción junto con datos de test

Se pide además comparar los residuos entre los datos de entrenamiento y los de test. Para poder compararlo se han valorado las siguientes tres medidas:

- SSE (Suma de errores cuadrados): Se descarta ya que el error aumenta con el número de muestras de entrenamiento. Como el número de muestras de entrenamiento es distinto que el de las muestras de test, no nos sirve esta alternativa.
- MSE (Error cuadrático medio): Arregla el problema del SSE, pero es difícil de interpretar puesto que el resultado serían $€^2$.
- **RMSE** (Raíz cuadrada de MSE): Mucho más fácil de interpretar. Devuelve el resultado en las unidades correctas, en este caso €.

Por esto se utiliza el RMSE para comparar los residuos entre los datos de entrenamiento y los datos de test:

$$RMSE\ Train\ data\colon\ 102739.708064\ \simeq\ 102739,71\ €$$
$$RMSE\ Test\ data\colon\ 80628.544728\ \simeq\ 80628,54\ €$$

Como se puede observar, el error de entrenamiento es considerablemente superior al error de test. Esto se debe a que la media de los errores es sensible a los datos anómalos que se presentan en los datos de entrenamiento, como se puede observar en la Figura 1.

3. Regresión multivariable utilizando ecuación normal

En este apartado se busca encontrar los mejores valores de θ, pero esta vez utilizando dos variables: superficie (m^2) y nº de habitaciones. De nuevo se utiliza la ecuación normal para calcular dichos valores.

La recta de predicción obtenida es la siguiente:

$$-\ 12132.908900\ +\ 3028.744415\ *\ x1\ -\ 18852.830946\ *\ x2$$

En la Figura 4 se muestra el plano de predicción junto con los puntos de entrenamiento.

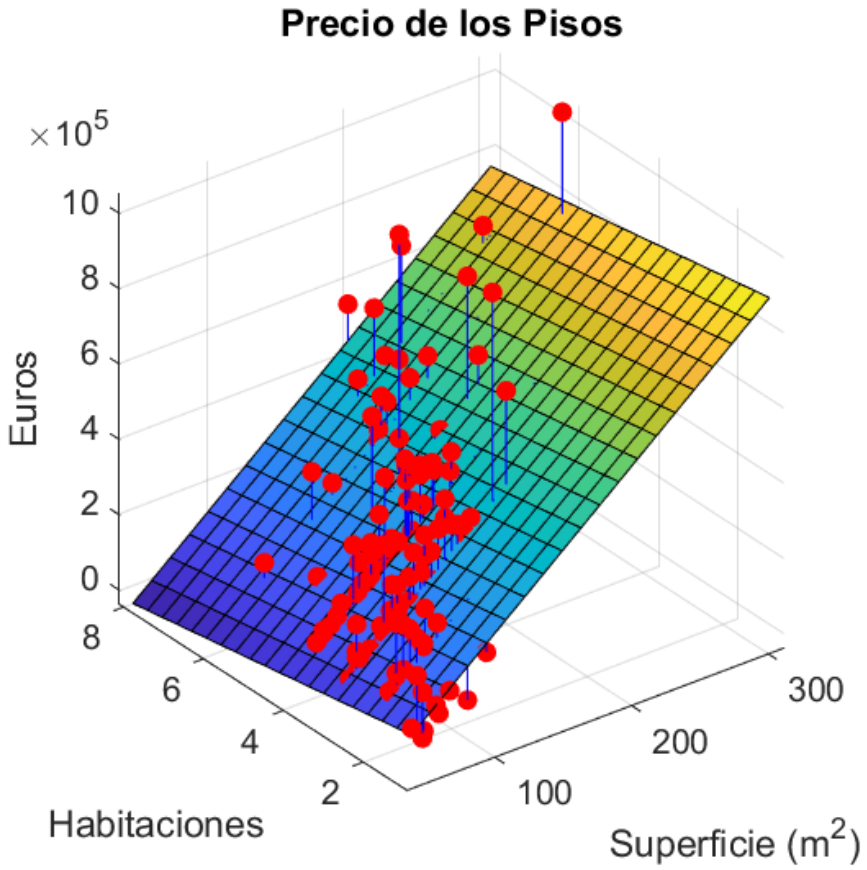

Figura 4. Plano de predicción multivariable utilizando ecuación normal

Los residuos obtenidos para los datos de entrenamiento y los datos de test son los siguientes:

$$RMSE\ Train\ data\colon\ 319546.462874\ \simeq\ 319546,46\ €$$
$$RMSE\ Test\ data\colon\ 74740.529167\ \simeq\ 74740,53\ €$$

De nuevo, el error medio en los datos de entrenamiento es mayor que el de los datos de test, debido a la presencia de datos anómalos en los datos de entrenamiento.

En este apartado se pide también comparar el modelo monovariable del primer apartado con este modelo multivariable. En concreto se pide comparar el precio de un piso de $100m^2$ con 2, 3, 4 ó 5 habitaciones y decidir cuál es mejor.

En primer lugar, como el primer modelo es monovariable, solo será posible predecir el precio en función de la superficie, por lo que no se tendrá en cuenta el número de habitaciones.

Utilizando la ecuación de predicción del primer apartado:

$$f(x) = -34313.387972 + 2608.209435 * x$$

$$f(100) = -34313.387972 + 2608.209435 * 100 = 226507,55\,€$$

En segundo lugar, se pasa a calcular el precio con el segundo modelo (multivariable), teniendo en cuenta ahora también el número de habitaciones del piso con la ecuación de predicción del modelo multivariable:

$$f(x1, x2) = -12132.908900 + 3028.744415 * x1 - 18852.830946 * x2$$

$$\begin{aligned} f(100,2) &= -12132.908900 + 3028.744415 * 100 - 18852.830946 * 2 \\ &= 253035,87\,€ \\ f(100,3) &= -12132.908900 + 3028.744415 * 100 - 18852.830946 * 3 \\ &= 234183,04\,€ \\ f(100,4) &= -12132.908900 + 3028.744415 * 100 - 18852.830946 * 4 \\ &= 215330,21\,€ \\ f(100,5) &= -12132.908900 + 3028.744415 * 100 - 18852.830946 * 4 \\ &= 196477,38\,€ \end{aligned}$$

Se puede observar como el precio del piso disminuye con el número de habitaciones. Estos resultados indican que la predicción del modelo multivariable no es buena, pues no es lógico que un piso con 5 habitaciones sea más barato que uno de 2, teniendo la misma superficie. Esto se debe a la presencia de datos espurios.

¿Cuál de los dos modelos es mejor?

Para comparar los dos modelos se han comparado los RMSE de los datos de test de cada uno de ellos. Se concluye por tanto que el modelo multivariable es mejor ya que se obtiene un error medio menor (80628,54 < 74740,53) que en el modelo monovariable.

4. Regresión monovariable utilizando Descenso de Gradiente

En este caso se busca resolver la regresión monovariable (superficie m^2) con el método de Descenso de Gradiente.

Utilizando la implementación del algoritmo de Descenso de Gradiente planteado en trabajo previo realizado en la Figura 1, se ha obtenido la evolución de la función de coste (J(θ)) (Figura 5). Para que la evolución fuera visual, se ha mostrado la evolución de la función de coste durante las 50 primeras iteraciones del algoritmo.

Se han probado distintos valores de α hasta dar con el óptimo ($\alpha = 0.0000001$).

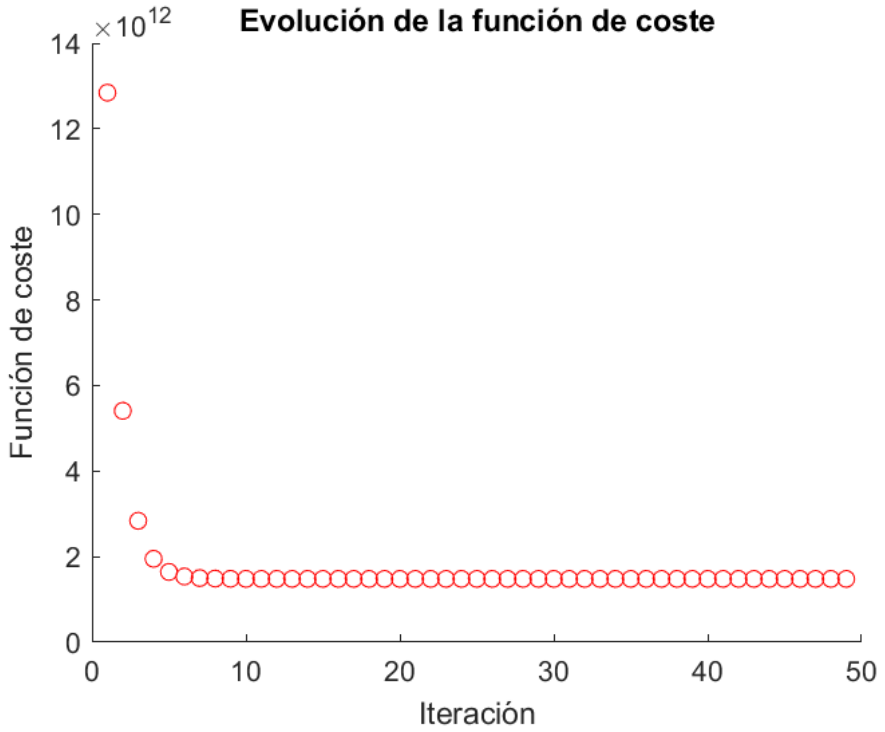

Figura 5. Evolución de la función de coste al aplicar Descenso de Gradiente

4.1. Comparación resultados con modelo monovariable usando ecuación normal

La recta de predicción obtenida utilizando Descenso de Gradiente es la siguiente:

$$f(x) = -34310.964281 + 2608.191099 * x1$$

En cuanto al RMSE:

$$RMSE\ Train\ data: 102739.708068 \simeq 102739,71\ €$$
$$RMSE\ Test\ data: 80628.396938 \simeq 80628,40\ €$$

Como se puede observar, los resultados son prácticamente idénticos. Esto se debe a que tanto en el método de descenso de gradiente como en el de ecuación normal se busca lo mismo: encontrar θ tal que minimice la función de coste cuadrático. La diferencia entre ambos métodos es la manera de hallar θ, pero buscan la misma recta.

5. Regresión multivariable utilizando Descenso de Gradiente

Ahora resolveremos la regresión multivariable con Descenso de Gradiente.

El plano de predicción obtenido es el siguiente:

$$f(x) = \ -\ 12133.172179\ +\ 3028.722615\ *\ x1\ -\ 18852.093512\ *\ x2$$

En la Figura 6 se muestra de manera gráfica dicho plano junto con los puntos de entrenamiento.

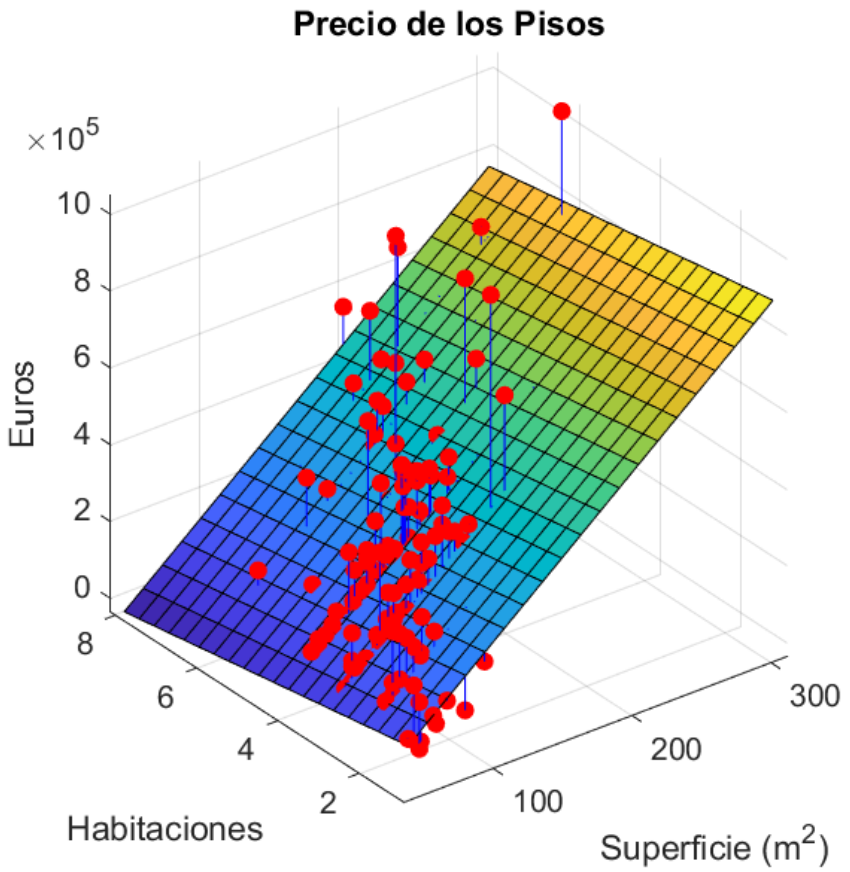

Figura 6. Plano de predicción de regresión multivariable utilizando descenso de gradiente

La función del plano es prácticamente igual a la obtenida utilizando ecuación normal, por lo que el plano de predicción también lo es.

Como consecuencia de esto, el RMSE de los datos también será igual que en el apartado 3.

$$RMSE\ Data\ train\colon\ 319546.472633\ \simeq\ 319546,47\ €$$
$$RMSE\ Data\ test\colon\ 74740.572288\ \simeq\ 74740,57\ €$$

Esto se traduce en que no hay diferencia entre los modelos con el mismo número de variables pero que utilizan ecuación normal y descenso de gradiente.

6. Regresión robusta con coste de Huber

En este último apartado, se resolverá la regresión multivariable utilizando regresión robusta con el coste de Huber, con el objetivo de que los datos espurios sean menos significativos en la predicción.

Para utilizar la función de coste de Huber, hay que determinar el parámetro δ que especifica el error "razonable". Este valor se puede estimar mirando el error aproximado de los datos "buenos" en el plano de predicción junto con los datos de entrenamiento del modelo multivariable. En la Figura 7 se puede observar como la mayor parte de los datos (datos "buenos" tienen un residuo en torno a 10^5 €.

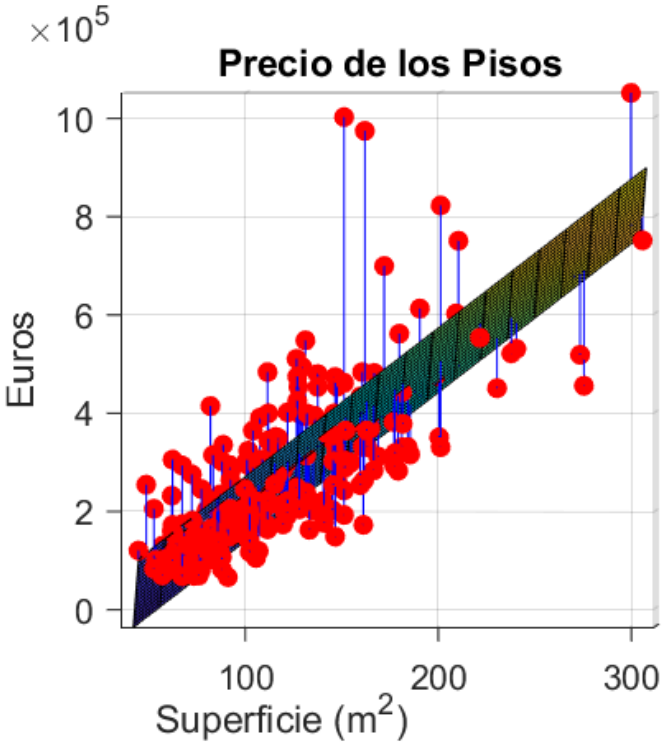

Figura 7. Vista de lado del plano de predicción del modelo multivariable en el que se puede apreciar el residuo aproximado de la mayoría de los datos

Con este valor de δ ($\delta = 10^5$), se aplica el descenso de gradiente utilizando el coste de Huber en lugar del cuadrático para calcular θ.

Una vez encontrados los valores de θ óptimos nos queda la siguiente ecuación de predicción:

$$f(x) = \ -7727.527378 \ + \ 2774.122235 * x1 \ + \ 15069.306429 * x2$$

que define el siguiente plano:

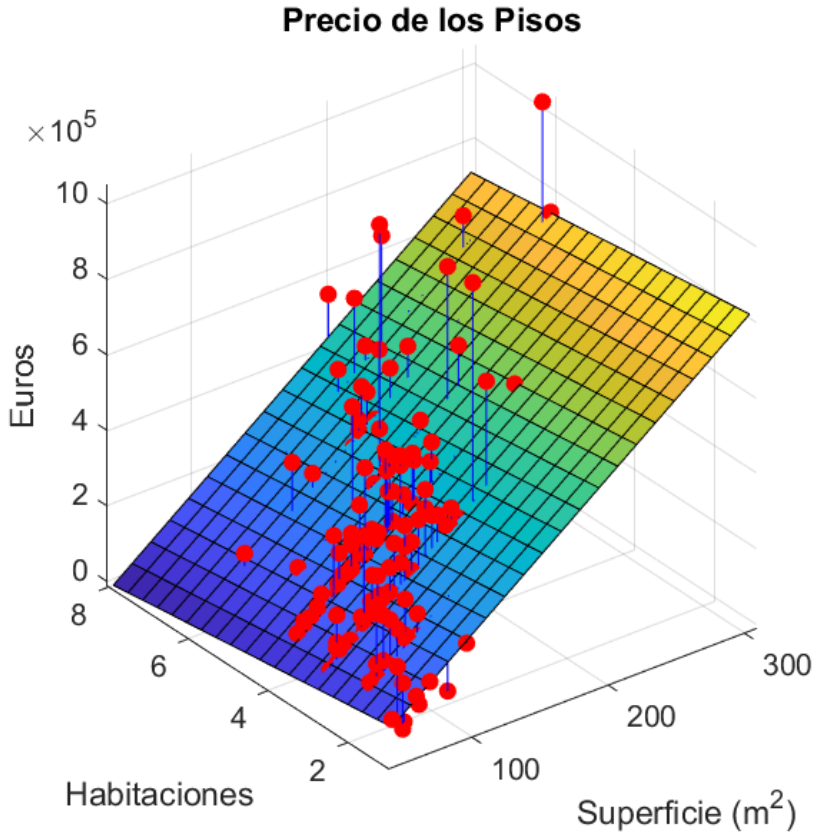

Figura 8. Plano de predicción de regresión multivariable utilizando coste de Huber y descenso de gradiente

El plano obtenido es ligeramente distinto al obtenido en el apartado anterior, presentando una pendiente un poco menor con respecto a la superficie.

Al calcular el RMSE de los datos de entrenamiento y de test se obtiene:

$$RMSE\ Data\ train\colon\ 314763.993555 \simeq 314763,99\ € < 319546,47\ €$$
$$RMSE\ Data\ test\colon\ 70831.375330 \simeq 70831,37\ € < 74740,57\ €$$

El error medio en ambos casos es menor que en el modelo del apartado anterior, lo que indica que el modelo que utiliza regresión robusta es algo mejor que el modelo que no la utiliza.

Conclusiones

Como resultado de esta práctica se puede concluir que el modelo multivariable funciona mejor que el modelo monovariable. Además se remarca la importancia que tienen los datos anómalos y cómo pueden afectar al modelo de predicción si se utiliza una función de coste cuadrático.