**Informe Trabajo Practico 1**

**Métodos Computacionales**

**Problema**

Buscamos realizar una regresión lineal en datos reales y así poder analizarlos. Para esto partimos desde la deducción de la fórmula para la solución óptima del problema, planteamos la regresión lineal y luego la aplicamos a los datos.

**Parte 1**

Para resolver cada paso de la deducción de la fórmula para β\* nos apoyamos fuertemente en el teorema presentado en la consigna, además de en las propiedades de las operaciones matriciales y vectoriales. Haciendo los despejes correctos se podía alcanzar el objetivo en cada inciso sin dificultad.

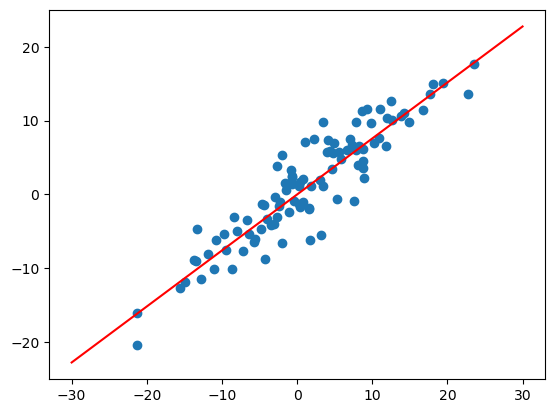
**Parte 2**

Comenzamos por graficar los puntos en el plano para luego superponer la recta de la regresión.

Primero, utilizando la fórmula obtenida en la parte 1, conseguimos β\* tomando los datos de ejercicio\_1.csv para los vectores X e Y.

Luego, sabiendo que la recta toma la forma por lo definido en las consignas del problema, tomamos el β\* calculado y a lo establecemos como un dominio con np.linspace(-30,30,1000) para poder visualizar mejor la regresión.

Así logramos una primera resolución.



Pero cuando se nos plantea aumentar en 12 unidades los valores de la coordenada para cada punto, nos damos cuenta que nuestro modelo está incompleto.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Observamos que, a la recta para seguir la estructura típica de una ecuación lineal, le está faltando una ordenada al origen. En nuestro modelo se traduce a una ausencia de β0. Fue casualidad que para el primer conjunto de datos estos estuvieran “centrados en 0” y por lo tanto tomando β0 = 0 la recta de la regresión se ubicara correctamente.

Para conseguir el β0 ampliamos la matriz de X agregando una columna llena de 1s, para que luego de hacer las operaciones correspondientes y llegue el momento de multiplicar por Y, se genere el vector .

Por último, construimos la recta de la forma y así obtenemos el grafico adecuado.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Para los datos provistos por ejercicio\_2.csv ocurre algo muy similar y recurrimos a la solución propuesta (incluir en los cálculos).

Gráfico, Gráfico de dispersión

Descripción generada automáticamenteGráfico, Gráfico de dispersión

Descripción generada automáticamente

**Parte 3**

Para esta ultima parte incorporamos datos reales para poder realizar las estimaciones correspondientes.

Primero se nos pide estimar los parámetros de que minimizaran el error cuadrático medio usando los datos de entrenamiento. Como ahora tenemos 6 categorías de datos que van a conformar las columnas de nuestra matriz X, la ecuación de la recta para la regresión lineal deberá tener la forma:

La matriz X además de tomar por columnas a las 6 categorías del dataset, debe tener como primera columna una llena de 1s con tal de poder generar como ya explicamos en la parte 2. Luego de aplicar las transformaciones correspondientes β\* es un vector de 7 filas donde cada una contiene el valor de con de 0 a 6.

Ahora para estimar la variable respuesta buscamos y esto es repetir la misma cuenta de siempre: .

Por último, se calcula el error cuadrático medio utilizando la formula provista y obtenemos un ECM = 83.165, lo cual es bastante elevado para dar una buena estimación.

Este proceso se repite para los datos de test y concluye en un ECM = 58.664, el cual es mucho menor que el de los datos de entrenamiento. Observando el data set notamos la existencia de valores atípicos para el primer conjunto de datos que hacían que el ECM sea mucho mayor.

Al calcular el ECM con el data set completo vemos que disminuye aún más dado que el efecto de estos valores atípicos se pierde entre muchos típicos. Graficamos entonces el ECM para cada casa y así podemos ver que casas son las que contienen diferencias más atípicas.

Gráfico, Histograma

Descripción generada automáticamente

Finaliza el experimento con la pregunta de si agregar una variable más disminuiría el ECM. La información que se quiere agregar específicamente es el año de construcción de la casa. Observamos que esta variable es linealmente dependiente de la edad de la casa por lo que la nueva columna no nos brinda información adicional. Incluso, en una regresión lineal, cuando agregamos información irrelevante genera “ruido” en la estimación y entonces el ECM no solo no disminuye, sino que es altamente probable que aumente.