Cuadrados mínimos lineales

Presentación TP3

Regresiones lineales

- En el TP2 resolvimos un problema de clasificación
- En el TP3 vamos a resolver un problema de regresión.

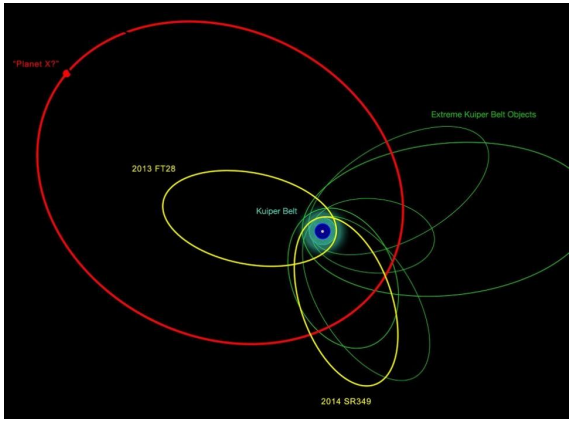
Las regresiones nos pueden ayudar en diferentes contextos:

- Sabemos exactamente a qué familia de funciones responden los datos, pero no podemos confiar completamente en los datos.
- No sabemos qué familia de funciones explica la naturaleza de los datos y queremos encontrar la mejor manera de explicarlos con alguna familia en particular.

Regresiones lineales

Un caso donde ya sabemos la familia de funciones:

- Órbitas elípticas de los planetas.

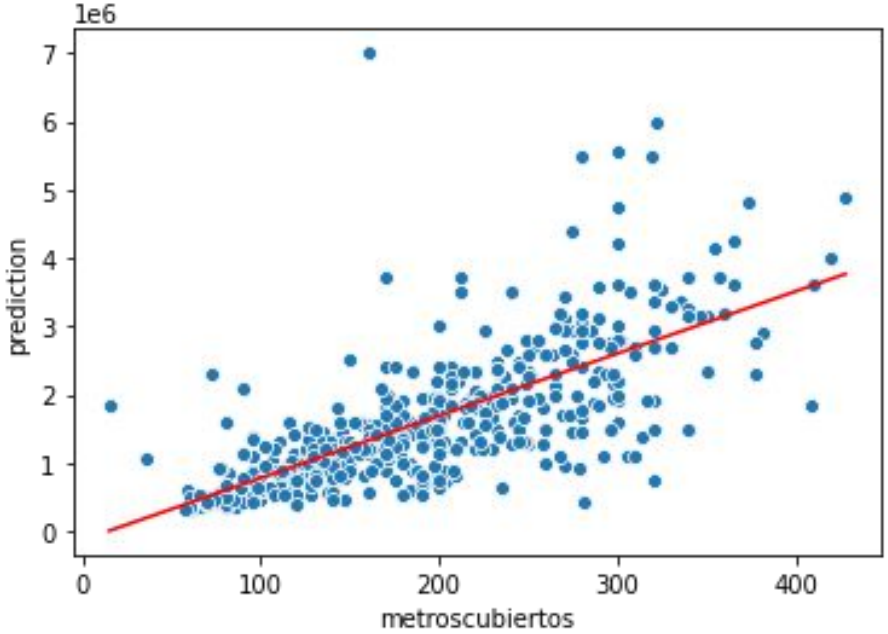


Un caso donde no tenemos ni idea la familia:

- Predicción de precios inmobiliarios

Objetivo

El objetivo del TP3 será utilizar cuadrados mínimos lineales para generar regresiones que expliquen diversas características (principalmente el precio) en un dataset de avisos inmobiliarios



Dataset

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 240000 entries, 0 to 239999
Data columns (total 23 columns):
 #   Column                        Non-Null Count   Dtype  
---  ------                        --------------   -----  
 0   id                            240000 non-null  int64  
 1   titulo                        234613 non-null  object 
 2   descripcion                   238381 non-null  object 
 3   tipodepropiedad               239954 non-null  object 
 4   direccion                     186928 non-null  object 
 5   ciudad                        239628 non-null  object 
 6   provincia                     239845 non-null  object 
 7   antiguedad                    196445 non-null  float64
 8   habitaciones                  217529 non-null  float64
 9   garages                       202235 non-null  float64
 10  banos                         213779 non-null  float64
 11  metroscubiertos               222600 non-null  float64
 12  metrostotales                 188533 non-null  float64
 13  idzona                        211379 non-null  float64
 14  lat                           116512 non-null  float64
 15  lng                           116512 non-null  float64
 16  fecha                         240000 non-null  object 
 17  gimnasio                      240000 non-null  float64
 18  usosmultiples                 240000 non-null  float64
 19  piscina                       240000 non-null  float64
 20  escuelascercanas              240000 non-null  float64
 21  centroscomercialescercanos    240000 non-null  float64
 22  precio                        240000 non-null  float64
dtypes: float64(15), int64(1), object(7)
memory usage: 42.1+ MB
```

Preguntas a responder

Si bien queremos explicar el precio como objetivo principal, hay muchas otras relaciones entre variables que podríamos analizar.

- ¿Podemos aproximar la cantidad de baños según la antigüedad y el tamaño en metros cuadrados del inmueble?

- ¿Podemos explicar la cantidad de piscinas según la latitud del inmueble?

- ¿Podemos aproximar la antigüedad de un inmueble según su ciudad y zona?

Variables categóricas

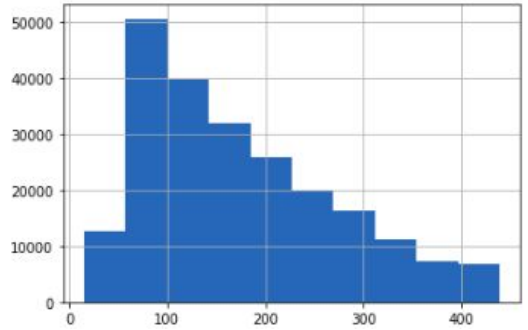
- Solo toman ciertos valores predeterminados.
- No tienen una noción de distancia entre valores.
- No las usamos en cuadrados mínimos, pero sí en el algoritmo.

```
df['tipodepropiedad'].value_counts()
```

```
Casa                                141717
Apartamento                          57341
Casa en condominio                   19297
Terreno                               9945
Local Comercial                       3055
Oficina comercial                     1741
Bodega comercial                      1406
Edificio                              1396
Terreno comercial                     1326
Casa uso de suelo                      708
Quinta Vacacional                      395
Duplex                                 343
Villa                                  340
Inmuebles productivos urbanos          200
Rancho                                 170
Local en centro comercial              165
Departamento Compartido                141
Otros                                  134
Nave industrial                         76
Terreno industrial                      31
Huerta                                  20
Lote                                     5
Hospedaje                                1
Garage                                   1
Name: tipodepropiedad, dtype: int64
```

```
df['metroscubiertos'].hist()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f27f2192e80>
```




```
df['banos'].value_counts()
```

```
2.0    87683
1.0    58173
3.0    49365
4.0    18558
Name: banos, dtype: int64
```

RMSE vs RMSLE

Una vez hecha la regresión, necesitamos poder medir la calidad del algoritmo.

Para el precio utilizaremos RMSE y RMSLE.

Diferencias:

- Ponderación de muestras (proporcionalidad).
- Ponderación de aproximación.

$$RMSE(\hat{f}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} e_{(i)}^2}$$

$$RMSLE(\hat{f}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\log(y_{(i)}+1) - log(\hat{y}_{(i)}+1))^2}$$

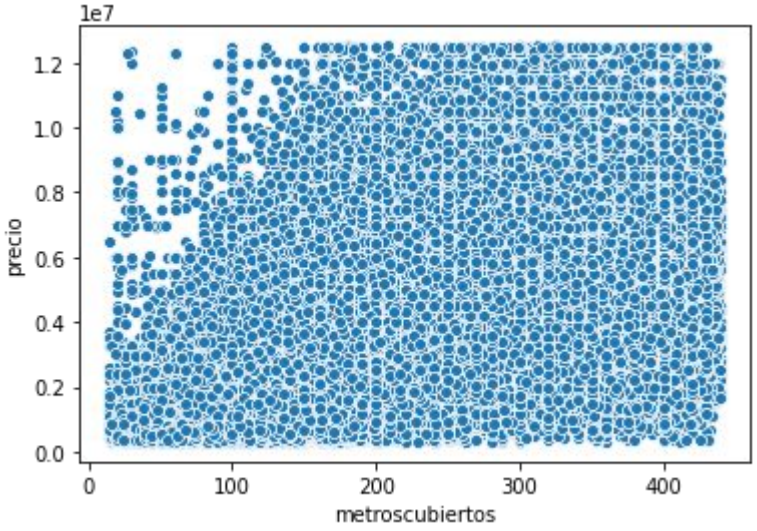
Segmentación del dataset

Es muy complejo intentar explicar todo el dataset con una única función.

Podemos segmentar el dataset para aplicar CML.

Algunas segmentaciones posibles:

- Por variables categóricas.
- Por variables numéricas cuantizadas.
- Incluso por la variable que se quiere predecir aunque no la sepamos.



Feature Engineering

Feature engineering se le llama a producir nuevas características para utilizar en las regresiones lineales.

Posibles formas hacer feature engineering:

- Combinación de características: copado para el verano sí y solo sí el inmueble tiene pileta o se encuentra al norte del paralelo 26.
- Extracción de información de los campos de texto: Tomar la descripción del aviso y ver qué palabras aparecen.
- Características de fuentes externas: Utilizar fuentes externas como indicadores socioeconómicos.

Enunciado

Vamos al enunciado!