Clase N°8. Modelo de regresión lineal múltiple

Es aquel que aproxima linealmente a un conjunto de datos con n variables observables $(x_1, x_2, \ldots x_n)$, donde la ecuación de aproximación es la de un *hiperplano*, representada por:

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon,$$

siendo $\beta_o, \beta_1, \beta_2, \ldots, \beta_n$ los coeficientes de la ecuación del *hiperplano*, Y la variable predictiva y ε el error o residuo cometido en la aproximación.

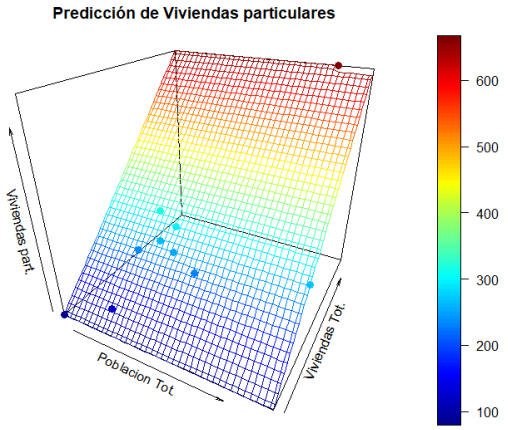
Ejemplo 1.

Queremos analizar cómo varía el volumen de viviendas particulares en relación a la población y a la cantidad de viviendas totales en un barrio de CABA.

Utilizando un modelo de regresión lineal múltiple se puede representar el fenómeno a través de la sig. ecuación:

$$Viviendas\ particulares = \beta_o + \beta_1 . población\ Tot. + \beta_2 . viviendas\ Tot.$$

Esta relación se puede mostrar también mediante el sig. gráfico:



Hiperparámetros del modelo de regresión lineal múltiple en Python

Se obtienen los mismos hiperparámetros que le modelo de regresión lineal simple, utilizando la opción *modelo.**get.param*** **()**

```
'copy_X': True.
'fit_intercept': True.
```

'n_jobs': None.

'normalize': 'deprecated'.

'positive': False.

Métricas del modelo de regresión lineal múltiple en Python

Se usan las mismas métricas que en el modelo de regresión lineal simple:

MAE: The Mean Absolute Error.

MSE: The Mean Squared Error.

RMSE: The Root Mean Squared Error.

R^2**:** El coeficiente de determinación.

Análisis multivariado

Analiza el grado de asociación entre un conjunto de *n* variables independientes, el cual se suele representar mediante una matriz de correlación para mostrar el grado de asociación entre las variables de un conjunto de datos.

Se parte de una **matriz de varianza-covarianza** para determinar el grado de asociación entre las variables numéricas. Por ejemplo para la base de datos iris, resulta:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	0.685694	-0.039268	1.273682	0.516904
sepal width (cm)	-0.039268	0.188004	-0.321713	-0.117981
petal length (cm)	1.273682	-0.321713	3.113179	1.296387
petal width (cm)	0.516904	-0.117981	1.296387	0.582414

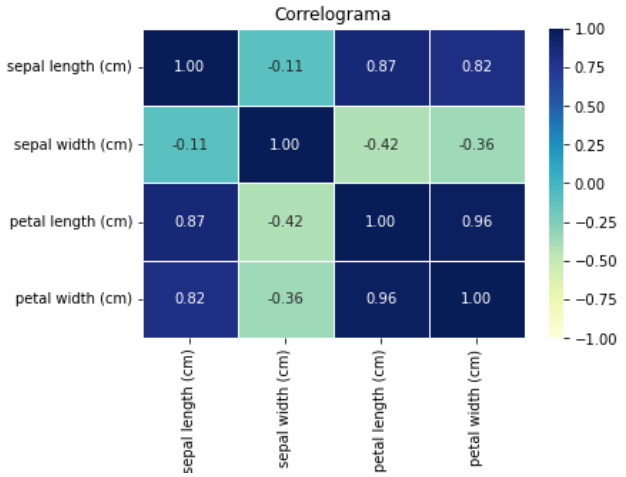
Donde los valores de la diagonal principal muestran las varianza y los resultados restantes los valores de las covarianzas. La matriz de varianza-covarianza cumple la propiedad de ser una matriz simétrica y cuadrada.

Al normalizar la matriz de varianza-covarianza se obtiene una **matriz de correlación:**

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
sepal length (cm)	1.000000	-0.109369	0.871754	0.817954
sepal width (cm)	-0.109369	1.000000	-0.420516	-0.356544
petal length (cm)	0.871754	-0.420516	1.000000	0.962757
petal width (cm)	0.817954	-0.356544	0.962757	1.000000

Los valores de la diagonal principal muestran correlación perfecta entre las variables y los valores por encima y por debajo reflejan diferentes niveles de asociación. Los negativos correlación inversa, los positivos correlación directa, los cercanos a uno correlación fuerte y próximos a cero correlación débil.

La matriz de correlación se puede graficar en un **correlograma** o mapa de calor, llamado: "heatmap", como se observa a continuación:



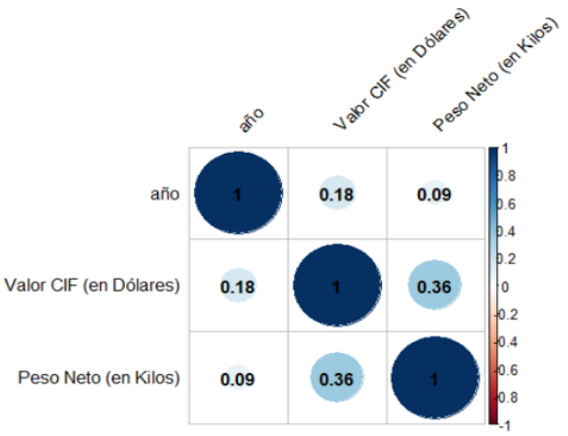
Ejemplo 2.

Se quiere determinar la relación entre las variables: *precio de importación*, *Peso Neto* importado y *Año* de la importación de mercaderías en la provincia entre los años 2016 y 2018. Para ello se muestra la sig. matriz de correlación:

	Año	Valor CIF (u$s.)	Peso Neto (kg.)
Año	**1,00**	0,18	0,09
Valor CIF (u$s.)	0,18	**1,00**	0,36
Peso Neto (kg.)	0,09	0,36	**1,00**

En la tabla prevalece un grado de asociación positivo entre las variables. El mayor grado de asociación está presente entre las variables: *"Valor CIF"* y *"Peso Neto"*, no obstante, su grado de asociación es medio. Por otra parte, las variables *"Peso Neto"* y *"Año"* presentan el menor grado de correlación.

La relación entre las variables se puede mostrar en un gráfico llamado *correlograma*:



La *correlación es positiva* si el color de la esfera es azul y si el color es rojo, la *correlación es negativa*. Por otra parte, el tamaño de las esferas y la intensidad del color es proporcional al grado de asociación, es decir, esferas grandes de color intenso indican *correlación fuerte* y esferas pequeñas de color claro expresan *correlación débil*.

Actividad: Preguntas de opción múltiple.

Responder con verdadero o falso a los siguientes enunciados:

1. Un modelo de regresión lineal múltiple busca la relación entre n variables respuestas.
2. Una herramienta usada en análisis multivariado es la matriz de correlación.
3. Un correlograma muestra el grado de dispersión entre dos variables.