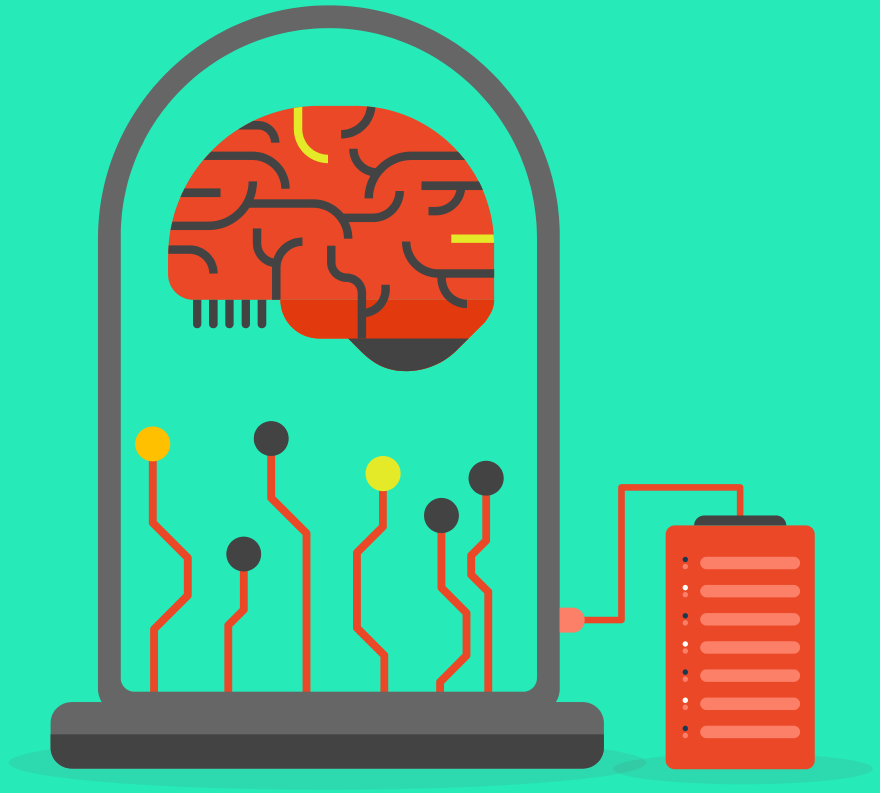
4. Regresión Lineal Múltiple



Planteamiento del Problema

Como vimos en la introducción a la Regresión Lineal, la función hipótesis que modela el caso lineal parte de una ecuación que tiene dos parámetros y una sola variable independiente. De allí a que este tipo de regresión también se llame **lineal univariada** o **regresión lineal simple**.

Sin embargo, es posible contar con problemas en donde se cuenten con dos o más variables independientes, y que de igual modo podamos realizar sobre ellas regresiones lineales. A este caso se le conoce como **Regresión Lineal Múltiple** o **Multivariada**.

Regresión Lineal Múltiple

En este caso, partiríamos de un conjunto de datos que, como se mencionó, cuenta con más de una variable independiente, y una variable dependiente que se desea modelar o predecir, por ejemplo:

x_1	x_2	x_3	x_4	y
12	-3.5	520	1	12.6
18	8.6	460	0	-26.5
32	-2.7	232	0	80.3
8	0.12	315	1	-18.3
14	3.6	178	2	40.2

Regresión Lineal Múltiple

Aunque acá solo se representan cuatro variables independientes, en realidad el problema podría tener cualquier cantidad, por lo que la hipótesis general para la Regresión Lineal Múltiple tendrá la forma:

$$h_\theta(x_1, \dots, x_n) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots \theta_n x_n$$

Y cuya manera de resolver es la misma que en el caso Lineal: aplicando la Ecuación Normal, el método de mínimos cuadrados o, por ejemplo, el método de Descenso del Gradiente. Es decir, cualquier método de optimización adecuado.

Del mismo modo, las métricas para conocer la calidad del ajuste siguen siendo las ya vistas, **RMSE** y **R^2**, aunque para el caso múltiple no resulta posible, por ejemplo, graficar las curvas de los ajustes debido a la cantidad de dimensiones.

Boston Housing Dataset

A fin de realizar la implementación práctica de la Regresión Lineal Múltiple, vamos a hacer uso de un dataset muy conocido dentro de la Ciencia de Datos, el “Boston Housing Dataset”. El mismo contiene 506 registros con 14 columnas que tienen la forma:

	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	b	lstat	medv
2	0.00632	18	2.31	0	0.538	6.575	65.2	4.09	1	296	15.3	396.9	4.98	24
3	0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	396.9	9.14	21.6
4	0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
5	0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
6	0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	396.9	5.33	36.2
7	0.02985	0	2.18	0	0.458	6.43	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
8	0.08829	12.5	7.87	0	0.524	6.012	66.6	5.5605	5	311	15.2	395.6	12.43	22.9
9	0.14455	12.5	7.87	0	0.524	6.172	96.1	5.9505	5	311	15.2	396.9	19.15	27.1
10	0.21124	12.5	7.87	0	0.524	5.631	100	6.0821	5	311	15.2	386.63	29.93	16.5
11	0.17004	12.5	7.87	0	0.524	6.004	85.9	6.5921	5	311	15.2	386.71	17.1	18.9
12	0.22489	12.5	7.87	0	0.524	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15
13	0.11747	12.5	7.87	0	0.524	6.009	82.9	6.2267	5	311	15.2	396.9	13.27	18.9
14	0.09378	12.5	7.87	0	0.524	5.889	39	5.4509	5	311	15.2	390.5	15.71	21.7
15	0.62976	0	8.14	0	0.538	5.949	61.8	4.7075	4	307	21	396.9	8.26	20.4

Boston Housing Dataset

Y en donde las columnas representan las siguientes variables:

Crim: la proporción de crimen per capita.
Zn: proporción de tierra residencial sobre 25.000 pies cuadrados.
Indus: proporción de negocios industriales en acres por población.
Chas: Variable binaria que indica si la localidad hace frontera con el Río Charles.
Nox: concentración de óxido nítrico (partes por 10 millones).
Rm: promedio de número de habitaciones por localidad.
Age: proporción de localidades ocupadas por dueños antes de 1940.
Dis: distancia ponderada a 5 centros de empleo.
Rad: índice de accesibilidad a autopistas radiales.
Tax: valor de los impuestos a la propiedad por cada $10.000.
PtRatio: proporción de estudiantes/profesores por localidad.
B: proporción de personas de color por localidad.
Lstat: porcentaje de status bajo de la población.
Medv: valor promedio de las propiedades en $1000s.

Boston Housing Dataset

El objetivo, en este caso, es determinar si es posible predecir el valor promedio de las propiedades (la variable **Medv**) en función del resto de variables que pertenecen al dataset.

Para ello, vamos a explorar el dataset e implementar la Regresión Lineal Múltiple, para luego introducir los conceptos de **P-Values** y la **Selección de Variables**.

Vayamos a nuestro **Quinto Notebook Práctico**.