Empecemos por suponer que tenemos una colección de **n** datos que dependen de **p** variables, de esta forma entraremos en el problema de como encontrar una regresión que nos apróxime estos puntos y esto nos ayude a predecir su comportamiento, en caso de ser un fenómeno. El "mejor" de los casos sería en el que la regresión interpole todos los puntos, sin embargo esto no es siempre posible, veámoslo a continuación.

Notemos que hallar la regresión es hallar los coeficientes b_i de la ecuación $y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$.
Llamemos $\boldsymbol{b}$ al vector que contiene los b_i, $\mathbf{Y}$ el vector que contiene los resultados de las observaciones, los cuales denotaremos por y_i, A_0 como un vector de dimensión 1xn de unos, y A_i como una vector de dimensión 1xn de los puntos considerados de la variable x_i, y por ultimo $\mathbf{X}$ como la matriz compuesta por los vectores columna $A_0, A_1, \cdots, A_i$. De esta forma, para hallar los b_i de manera que interpolara todos los puntos de nuestros datos sería resolver el siguiente sistema:

$$\begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

donde x_{ij} representa el i-ésimo valor de la variable x_j.
De esta forma, el sistema sólo tendrá solución si $\boldsymbol{Y} \in Col(X)$, lo cual no es necesariamente cierto, así que tendremos que buscar otra manera.

Vamos a buscar $\hat{\boldsymbol{b}}$ tal que la expresión $\boldsymbol{r} := \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}$ sea minimizada, donde $\boldsymbol{r}$ es nuestro vector residual y además, claramente $\hat{\boldsymbol{Y}} := \boldsymbol{X}\hat{\boldsymbol{b}} \in Col(X)$. Por lo tanto nuestro proposito es que el residuo sea el menor posible y esto será (por conveniecia) minimizar la norma al cuadrado del residual, es decir:

$$\sum_{i=1}^{n} r_i^2 = ||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}||^2$$

Donde r_i es el i-ésimo componente de $\boldsymbol{r}$. Ahora, para minimizar, usaremos cálculo de matrices.

Sabemos que $||\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}||^2 = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}})^T \cdot (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}}) = \boldsymbol{r}^T \cdot \boldsymbol{r}$, ahora derivemos esto con respecto a $\hat{\boldsymbol{b}}$ e igualemos con 0.

$$\begin{aligned} \frac{\partial \boldsymbol{r}^T \cdot \boldsymbol{r}}{\partial \hat{\boldsymbol{b}}} &= \frac{\partial}{\partial \hat{\boldsymbol{b}}}[(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}})^T \cdot (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{b}})] \\ &= \frac{\partial}{\partial \hat{\boldsymbol{b}}}[\boldsymbol{Y}^T\boldsymbol{Y} - 2\boldsymbol{Y}^T(\boldsymbol{X}\hat{\boldsymbol{b}}) + \hat{\boldsymbol{b}}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}}] \\ &= 0 - 2\boldsymbol{X}^T\boldsymbol{Y} + 2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} \end{aligned}$$

Y de esta forma, igualando con 0, tenemos que $\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}^T\boldsymbol{Y}$. Las cuales

son llamadas ecuaciones normales, donde los $\hat{\boldsymbol{b}}$ que cumplan las mismas serán nuestras soluciones.

Lo cual es congruente con el razonamiento geométrico asumiendo que el resultado es único, pues, si tomamos nuestra ecuación original y multiplicamos por $\boldsymbol{X}^T$ a ambos lados, obtenemos que $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b} = \boldsymbol{X}^T\boldsymbol{Y}$ esto es la proyección ortogonal de $\boldsymbol{Y}$ sobre $Col(X)$ y esto, si vemos los componentes de $\boldsymbol{b}$ como un punto de $\mathbb{R}^p$, es precisamente, la distancia más corta desde dicho punto hasta $Col(X)$ y dada proyección ortogonal viene dada por $\hat{\boldsymbol{Y}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, lo que tenemos en el lado derecho de la ecuación es justamente la proyección ortogonal de $\boldsymbol{Y}$ sobre $Col(X)$ y lo igualamos con $\hat{\boldsymbol{Y}}$ ya que este será el punto sobre $Col(X)$ con la distacia más corta a $\boldsymbol{Y}$, que es lo que buscamos. Así:

$$\begin{gathered}\boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\\ \hat{\boldsymbol{b}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\\ \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{b}} = \boldsymbol{X}^T\boldsymbol{Y}\end{gathered}$$

Además, notemos que cuando tenemos una única variable, es realmente fácil realizar manipulaciones que nos permitan encontrar otro tipo de funciones que nos permitan apróximar mejor que una regresión, bien sean polinomios o bien sean funciones que podamos transformar de forma que se vean cómo una lineal. Veamos como sería con polinomios

Supongamos que tenemos un conjunto de $m+1$ puntos en $\mathbb{R}^2$ $((x_0, y_0), (x_1, y_1), \cdots, (x_m, y_m))$ y queremos aproximar estos puntos usando un polinomio de grado n que llamaremos $p_n(x) = a_0 + a_1x + \cdots + a_nx^n$ por tanto quremos encontrar los a_i que nos permitan minimizar la expresión $E := \sum_{i=0}^{n}(y_i - p_n(x))^2$
Notemo que, si derivamos E con respecto a a_j, estos es:

$$\frac{\partial E}{\partial a_j} = -2\sum_{i=0}^{m} y_i x_i^j + 2\sum_{k=0}^{n} a_k \sum_{i=1}^{m} x_i^{j+k}$$

Al igualar esta expresión con 0 obtenemos las siguientes $n+1$ ecuaciones:

$$\begin{gathered}a_0\sum_{i=1}^{m} x_i^0 + a_1\sum_{i=1}^{m} x_i^1 + \cdots + a_n\sum_{i=1}^{m} x_i^n = \sum_{i=0}^{m} y_i x_i^0\\ a_0\sum_{i=1}^{m} x_i^1 + a_1\sum_{i=1}^{m} x_i^2 + \cdots + a_n\sum_{i=1}^{m} x_i^{n+1} = \sum_{i=0}^{m} y_i x_i^1\\ \vdots\\ a_0\sum_{i=1}^{m} x_i^n + a_1\sum_{i=1}^{m} x_i^{n+1} + \cdots + a_n\sum_{i=1}^{m} x_i^{2n} = \sum_{i=0}^{m} y_i x_i^n\end{gathered}$$

De esta forma, es suficiente con resolver este sistema lineal para econtrar los valores de los coeficientes a_i.

Ahora, ¿Cómo es aplicable esta teoría a nuestra realidad? Bien, pues es realmente importante para el análisis de datos, pues nos permite modelar y predecir comportamientos de fenómenos en nuestra realidad.

Cuando hablamos de la construcción de nuestro modelo, es muy bueno, porque se adapta de manera apropiada, ya que apesar de que en la vida real no tendremos comportamientos estrictos como el de una función, sí tendremos patrones y por consecuencia un método con el que podremos aproximarnos de gran manera.
Y en cuanto a los resultados también es realmente útil, pues gracias a su construcción es realmente fiable para realizar predicciones en el comportamiento de los fenómenos y en consecuencia con esto tomar acciones que nos lleven a obtener resultados distintos.

Gracias a estas cosas, el método de los mínimos cuadrados tiene aplicaciones en diversos sectores, cómo lo puede ser el crecimiento poblacional, agricultura, economía, ingeniería, róbotica, entre otras. Entre ellas el sector de ventas, o incluso en la educación, los cuales serán objeto de práctica en la siguiente sección de esta tarea.