Técnica de minería de datos

Regresión

Mariann Adaliz Avila Rios 1811303
Arturo del Ángel de la Cruz 1809895
Valeria Noemí Navarro Cabello 1820160
Magaly Rivera Valdez 1823340
Sofía Pamela Rosales Garza 1799219

Historia

1805

La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados publicado por Legendre.

Gauss desarrolló de manera más profunda el método e incluía una versión del teorema de Gauss-Markov.

Los modelos lineales son una explicación ágil y simplificada de la realidad por parte de la matemática y estadística.

1889

El término regresión fue introducido por Francis Galton en su libro “Natural inheritance” y fue confirmada por su amigo Karl Pearson.

Regresión

La regresión es un una técnica de minería de datos de la categoría predictiva. Predice el valor de un atributo en particular basándose en los datos recolectados de otros atributos.

La regresión se encarga de analizar el vínculo entre una variable dependiente y una o varias independientes, encontrando una relación matemática.

- Regresión lineal simple
- Regresión lineal múltiple

Regresión Lineal Simple

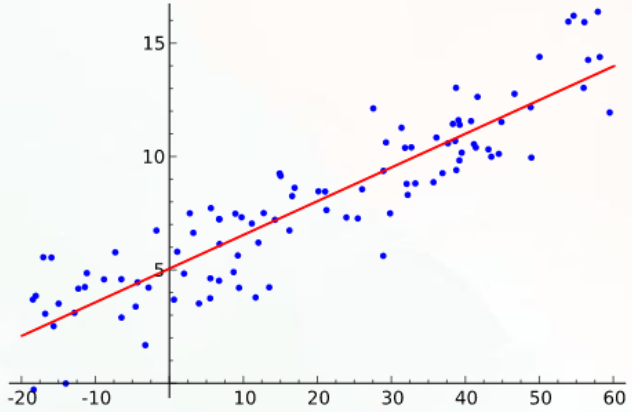
Cuando el análisis de regresión sólo se trata de una variable regresora, se llama regresión lineal simple.

La regresión lineal simple tiene como modelo:

$$y = \beta_0 + \beta_1 x + e$$

$y = \beta_o + \beta_1 x + e$

La cantidad ‘e’ en la ecuación es una variable aleatoria normalmente distribuida con $E(e)=0$ y $Var(e)= \sigma^2$



Estimación por mínimos cuadrados

La estimación de $y = \beta_o + \beta_1 x$ debe ser una recta que proporcione un buen ajuste a los datos observados. El modelo ajustado por mínimos cuadrados utiliza:

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1} x$$

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$$

Regresión Lineal Múltiple

$\beta_o, \beta_1, \dots, \beta_k$

Un modelo de regresión múltiple se dice lineal porque la ecuación del modelo es una función lineal de los parámetros desconocidos.

En general, se puede relacionar la respuesta "y" con los k regresores, o variables predictivas bajo el modelo:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + e$$

Estimación por mínimos cuadrados

Se puede escribir en la siguiente forma el modelo de regresión.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ \ldots + \beta_k x_{ik} + e_i = \beta_0 + \sum_{i=1}^{k} \beta_j x_{ij} + e_i$$

De tal manera que la función de minimos cuadrados es.

$$S(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

Se debe minimizar la función S respecto a $\beta_0, \beta_1, \ldots, \beta_k$ .Los estimadores de minimos cuadrados deben satisfacer las ecuaciones: .

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right) = 0$$

$$\frac{\partial S}{\partial \beta_j} = -2 \sum_{i=1}^{n} \left(y_i - \widehat{\beta_0} - \sum_{j=1}^{k} \widehat{\beta}_j x_{ij} \right) = 0$$

Al simplificar $\frac{\partial S}{\partial \beta_0}$ se obtienen las ecuaciones normales de mínimos cuadrados

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i$$

$$\vdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i$$

Nótese que hay p= k+1 ecuaciones normales, una para cada uno de los coeficientes desconocidos de regresión. La solución de las ecuaciones normales serán los por mínimos cuadrados $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

Es más cómodo manejar modelos de regresión múltiple cuando se expresan de forma matricial.

La notación matricial del modelo es $y = X\beta + e$ en donde

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Las ecuaciones normales de mínimos cuadrados quedan dadas por $X'X\hat{\beta} = X'y$.

Para resolverlas se multiplica ambos lados por la inversa de X'X.

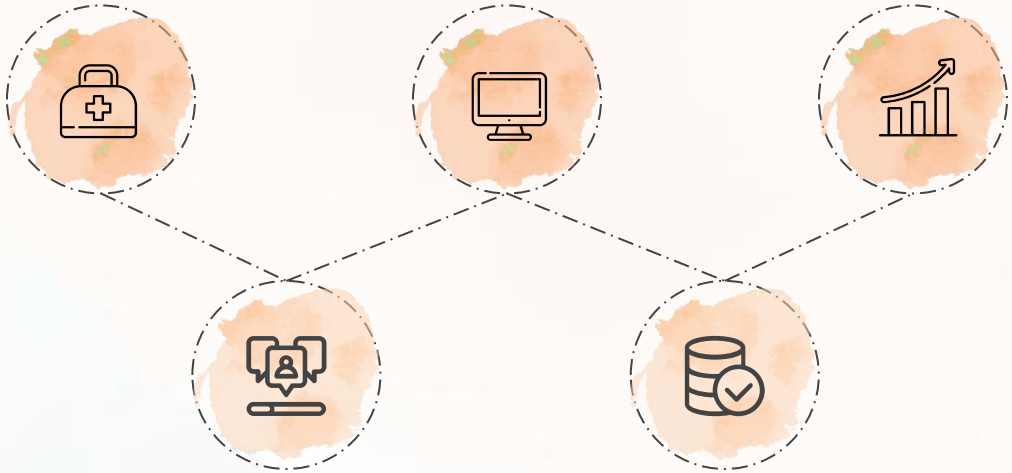
Así el estimador de mínimos cuadrados es $\boldsymbol{\hat{\beta} = (X'X)^{-1}X'y}$ siempre y cuando exista la matriz inversa $(X'X)^{-1}$; es decir, si ninguna columna de la matriz X es una combinación lineal de las demás columnas.

Aplicaciones

Medicina

Informática

Estadística



Comportamiento humano

Industria

Ejemplo

Año	Número de habitantes (millones)
1850	23.2
1860	31.4
1870	39.8
1880	50.2
1890	62.9
1900	76
1910	92
1920	105.7
1930	122.8
1940	131.7
1950	151.3
1960	179.3
1970	203.2

Ejemplo de regresión simple

Este conjunto de datos representa la población de Estados Unidos (en millones) según lo registrado por el censo decenal para el periodo 1850 - 1970

$$y = \beta_o + \beta_1 x + e$$