Regresión lineal

Escribe aquí tu nombre

13 de febrero de 2018

En estadística la regresión lineal o ajuste lineal es un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente Y, las variables independientes X_i y un término aleatorio ε. Este modelo puede ser expresado como:

$$Y_t = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

donde:

$Y_t Y_t$: variable dependiente, explicada o regresando.

$X_1, X_2, \cdots, X_p X_1, X_2, \cdots, X_p$: variables explicativas, independientes o regresores.

$\beta_0, \beta_1, \beta_2, \cdots, \beta_p \beta_0, \beta_1, \beta_2, \cdots, \beta_p$: parámetros, miden la influencia que las variables explicativas tienen sobre el regrediendo. donde $\beta_0 \beta_0$ es la intersección o término çonstante", las β_i $(i > 0) \beta_i$ $(i > 0)$ son los parámetros respectivos a cada variable independiente, y p p es el nÃžmero de parámetros independientes a tener en cuenta en la regresión. La regresión lineal puede ser contrastada con la regresión no lineal.

Historia

La primera forma de regresión lineal documentada fue el método de los mÃŋnimos cuadrados que fue publicada por Legendre en 1805, Gauss publicó un trabajo en donde desarrollaba de manera más profunda el método de los mÃŋnimos cuadrados,1âĂŃ y en dónde se incluÃŋa una versión del teorema de Gauss-Márkov.

El término regresión se utilizó por primera vez en el estudio de variables antropométricas: al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenÃŋan una estatura muy superior al valor medio, tendÃŋan a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendÃŋan a reducir su diferencia respecto a la estatura media; es decir, regresaban.ªl promedio.2âĂŃ La constatación empÃŋrica de esta propiedad se vio reforzada más tarde con la justificación teórica de ese fenómeno.

El término lineal se emplea para distinguirlo del resto de técnicas de regresión, que emplean modelos basados en cualquier clase de función matemática. Los modelos lineales son una explicación simplificada de la realidad, mucho más

ágiles y con un soporte teórico mucho más extenso por parte de la matemática y la estadÃŋstica.

Pero bien, como se ha dicho, se puede usar el término lineal para distinguir modelos basados en cualquier clase de aplicación.

El modelo de regresión lineal

El modelo lineal relaciona la variable dependiente Y con K variables explÃŋcitas $X_k X_k (k = 1, ...K)$, o cualquier transformación de éstas que generen un hiperplano de parámetros $\beta_k \beta_k$desconocidos:

(2) $Y = \sum \beta_k X_k + \varepsilon Y = \sum \beta_k X_k + \varepsilon$

donde $\varepsilon\varepsilon$ es la perturbación aleatoria que recoge todos aquellos factores de la realidad no controlables u observables y que por tanto se asocian con el azar, y es la que confiere al modelo su carácter estocástico. En el caso más sencillo, con una sola variable explÃŋcita, el hiperplano es una recta:

(3) $Y = \beta_1 + \beta_2 X_2 + \varepsilon Y = \beta_1 + \beta_2 X_2 + \varepsilon$

El problema de la regresión consiste en elegir unos valores determinados para los parámetros desconocidos $\beta_k \beta_k$, de modo que la ecuación quede completamente especificada. Para ello se necesita un conjunto de observaciones. En una observación i-ésima (i= 1,... I) cualquiera, se registra el comportamiento simultáneo de la variable dependiente y las variables explÃŋcitas (las perturbaciones aleatorias se suponen no observables).

(4) $Y_i = \sum \beta_k X_{ki} + \varepsilon_i Y_i = \sum \beta_k X_{ki} + \varepsilon_i$

Los valores escogidos como estimadores de los parámetros $\hat{\beta}_k \hat{\beta}_k$, son los coeficientes de regresión sin que se pueda garantizar que coincida n con parámetros reales del proceso generador. Por tanto, en

(5) $Y_i = \sum \hat{\beta}_k X_{ki} + \hat{\varepsilon_i} Y_i = \sum \hat{\beta}_k X_{ki} + \hat{\varepsilon_i}$

Los valores $\hat{\varepsilon_i} \hat{\varepsilon_i}$ son por su parte estimaciones o errores de la perturbación aleatoria.

Hipótesis del modelo de regresión lineal clásico

1. **Esperanza matemática nula:** $\mathbb{E}(\varepsilon_i) = 0$. Para cada valor de X la perturbación tomará distintos valores de forma aleatoria, pero no tomará sistemáticamente valores positivos o negativos, sino que se supone tomará algunos valores mayores que cero y otros menores que cero, de tal forma que su valor esperado sea cero.

2. **Homocedasticidad:** $\text{Var}(\varepsilon_t) = \mathbb{E}(\varepsilon_t - \mathbb{E}\varepsilon_t)^2 = \mathbb{E}\varepsilon_t^2 = \sigma^2$ para todo t. Todos los términos de la perturbación tienen la misma varianza que es desconocida. La dispersión de cada ε_t en torno a su valor esperado es siempre la misma.

3. **Incorrelación o independencia:** $\text{Cov}(\varepsilon_t, \varepsilon_s) = (\varepsilon_t - \mathbb{E}\varepsilon_t)(\varepsilon_s - \mathbb{E}\varepsilon_s) = \mathbb{E}\varepsilon_t \varepsilon_s = 0$ para todo t,s con t distinto de s. Las covarianzas entre las distintas pertubaciones son nulas, lo que quiere decir que no están correlacionadas. Esto implica que el valor de la perturbación para cualquier observación

muestral no viene influenciado por los valores de las perturbaciones correspondientes a otras observaciones muestrales. Regresores estocásticos. Independencia lineal. No existen relaciones lineales exactas entre los regresores. $T > k + 1$. Suponemos que no existen errores de especificación en el modelo, ni errores de medida en las variables explicativas. Normalidad de las perturbaciones: $\varepsilon \sim N(0, \sigma^2)$