Regresión lineal simple.

Jorge Valente Hernández Castelán.

Marzo 2021

1. Introducción

Dentro del ámbito de la estadística, uno de los tópicos más importantes son las regresiones lineales (simples o múltiples), las cuales tienen como objetivo predecir valores futuros en base a aquellos valores ya existentes con anterioridad, por lo que el método más comunmente utilizado para esto es el conocido como **mínimos cuadrádos ordinarios**.

Este método de estimación se basa en la idea de que hay una variable dependiente y otra(s) variable(s) independiente(s), la cual puede explicar en cierta medida los comportamientos de la variable dependiente, teniendo como resultado la siguiente expresión matemática:

$$Y = \alpha + \beta X_1$$

Esta función nos estaría diciendo que la variable dependiente Y cambiará β veces cada que la variable independiente X se modifique en una unidad. Mátematicamente hablando esto es correcto, sin embargo, recordemos que en el ámbito de la estadística los valores (casi) nunca se cumplen a la perfección, siempre existe un margen de error que debe ser considerado, por lo que en econometría se suele usar el término de error ε o μ para denotar el componente aleatorio de la función, o mejor dicho, el margen de error de la función, teniendo finalmente la siguiente expresión:

$$Y = \alpha + \beta X_1 + \varepsilon$$

Un modelo de regresión lineal debe cumplir con ciertos supuestos, los cuales son los siguientes:

1. La relación de X y Y es lineal.
2. Los valores de X son fijos.
3. El termino de error tiene un valor esperado de 0, o sea: $E(\varepsilon) = 0$.
4. El termino de error tiene una varianza constante para todas las observaciones, o sea: $E(\varepsilon^2) = 0 = \sigma^2$.

5. Los errores son estadísticamente independientes, es decir: $E(\varepsilon_j, \varepsilon_i) = 0 \ \forall \ i \neq j$.

6. El termino de error tiene una distribución normal.

Ahora, ya que conocemos dichos supuestos, podemos pasar a estimar los respectivos valores de α y β para un modelo de regresión lineal, de lo cual sabemos que α es la ordenada al origen y β la pendiente de la función. Si no se tienen los valores entonces hay que estimarlos, teniendo que $\hat{\alpha}$ y $\hat{\beta} \rightarrow \hat{Y}$, por lo que $\hat{\varepsilon} = Y - \hat{Y}$. El método de mínimos cuadrados ordinarios se basa en-valga la redundancia-minimizar los errores al cuadrado, por lo que aplicando lo anterior a las funciones de los errores estimados y la función de Y estimada, tendríamos lo siguiente:

$$\sum(\varepsilon_i)^2 = \sum(Y_i - \hat{Y})^2$$

Donde:

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X_i$$

Ahora, derivamos $\hat{\alpha}$ y $\hat{\beta}$ e igualamos a 0. Para poder estimar $\hat{\alpha}$ debemos comenzar por $\hat{\beta}$, teniendo dos posibles métodos para ello los cuales están desarrollados a continuación:

2. Método 1.

Sabiendo que $\sum \varepsilon_i^2 = \sum(Y_i - \hat{Y})^2$ y despejando $\hat{\alpha}$ de la función original $\hat{\alpha} = \hat{Y} - \hat{\beta}X_i$, entonces tendríamos:

$$\sum(\varepsilon_i)^2 = \sum(Y_i - \bar{Y} + \hat{\beta}\bar{X} - \hat{\beta}X_i)^2$$

Reagrupando y factorizando:

$$\sum(\varepsilon_i)^2 = \sum[(Y_i - \bar{Y}) - \hat{\beta}(X_i - \bar{X})]^2$$

Después, quitamos el exponente de la ecuación anterior:

$$\sum(\varepsilon_i)^2 = \sum[(Y_i - \bar{Y})^2 - 2\hat{\beta}(Y_i - \bar{Y})(X_i - \bar{X}) + \hat{\beta}^2(X_i - \bar{X})^2]$$

Derivamos con respecto a $\hat{\beta}$, tal que:

$$\frac{\partial \sum(\varepsilon_i)^2}{\partial \hat{\beta}} = -2\sum(X_i - \bar{X})(Y_i - \bar{Y}) + 2\hat{\beta}\sum(X_i - \bar{X})^2 = 0$$

y, finalmente, despejando $\hat{\beta}$:

$$-2\hat{\beta}\sum(X_i - \bar{X})^2 = -2\sum(X_i - \bar{X})(Y_i - \bar{Y})$$

$$\hat{\beta}\sum(X_i-\bar{X})^2=\sum(X_i-\bar{X})(Y_i-\bar{Y})$$

$$\hat{\beta}=\frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sum(X_i-\bar{X})^2}$$

Lo que en esencia es la covarianza de X y Y entre la varianza de X, es decir:

$$\hat{\beta}=\frac{Cov(X,Y)}{Var(X)}$$

3. Método 2.

Si tenemos la derivada siguiente:

$$\frac{\partial\sum(Y_i-\hat{\alpha}-\hat{\beta}X_i)}{\hat{\beta}}=-2\sum(Y_i-\hat{\alpha}-\hat{\beta}X_i)X_i=0$$

Entonces podemos despejar el -2 para cancelarlo, ya que $\frac{0}{X}=0$, tal que:

$$\hat{\beta}'=\sum(Y_i-\hat{\alpha}-\hat{\beta}X_i)X_i=0$$

$$\hat{\beta}'=\sum(Y_iX_i-\hat{\alpha}X_i-\hat{\beta}X_i^2)=0$$

Considerando la propiedad distributiva de $\sum$ en estadística:

$$\hat{\beta}'=\sum Y_iX_i-\hat{\alpha}\sum X_i-\hat{\beta}\sum X_i^2=0$$

Despejamos $\sum Y_iX_i$:

$$\hat{\beta}'=\sum Y_iX_i=\hat{\alpha}\sum X_i+\hat{\beta}\sum X_i^2$$

Ahora, sabemos que $\hat{\alpha}=\bar{Y}-\bar{\beta}\bar{X}$, por lo que podemos reemplazar en la ecuación, teniendo:

$$(\bar{Y}-\hat{\beta}\bar{X})\sum X_i+\hat{\beta}\sum X_i^2=\sum Y_iX_i$$

Desarrollando y despejando $\hat{\beta}$:

$$\sum X_i\bar{Y}-\sum X_i\hat{\beta}\bar{X}+\hat{\beta}\sum X_i^2=\sum Y_iX_i$$

$$-\sum X_i\hat{\beta}\bar{X}+\hat{\beta}\sum X_i^2=\sum Y_iX_i-\sum X_i\bar{Y}$$

Factorizamos $\hat{\beta}$ y la despejamos:

$$\hat{\beta}(-\bar{X}\sum X_i+\sum X_i^2)=\sum Y_iX_i-\sum X_i\bar{Y}$$

$$\hat{\beta}=\frac{\sum Y_iX_i-\sum X_i\bar{Y}}{\bar{X}\sum X_i-\sum X_i^2}$$

Finalmente, aplicando la propiedad $\sum X_i = n\bar{X}$ y redistribuyendo la expresión, tenemos:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

Llegando al mismo resultado que en el método anterior.

Ahora bien, para calcular el valor de $\hat{\alpha}$ simplemente hay que derivar lo siguiente e igualar a 0:

$$\frac{\partial \sum(Y_i - \hat{\alpha} - \hat{\beta}X_i)^2}{\partial \hat{\alpha}} = -2\sum(Y_i - \hat{\alpha} - \hat{\beta}X_i) = 0$$

Despejamos -2 y desarrollamos la expresión, teniendo:

$$\sum Y_i - n\hat{\alpha} - \hat{\beta}\sum X_i = 0$$

Finalmente, despejamos a $\hat{\alpha}$ para calcularla, tal que:

$$n\hat{\alpha} = \sum Y_i - \hat{\beta}\sum X_i$$

$$\hat{\alpha} = \frac{\sum Y_i}{n} - \frac{\hat{\beta}\sum X_i}{n}$$

Por lo que el valor final de $\hat{\alpha}$ estaría dado por la expresión:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

4. Interpretación.

La lógica de los valores de $\hat{\alpha}$ y $\hat{\beta}$ es que encontremos una función que pueda explicar de la mejor manera posible la tendencia que hay entre la relación de dos variables, suponiendo que están relacionadas. La forma de interpretar dichos valores es la siguiente: Sabiendo que Y es la variable dependiente o endógena y X la variable independiente o exógena, si X incrementa en una unidad (sea el caso de lo que estamos midiendo) entonces Y va a tender a variar en $\hat{\beta}$ unidades, pero si X se mantiene constante (no varía en el tiempo) entonces Y mantendrá un valor de $\hat{\alpha}$.

Referencias

[1] Hanck, C., Arnold, M., Gerber, A., & Schmelzer, M. (2019). Introduction to Econometrics with R. Essen: University of Duisburg-Essen.[Google Scholar].

[2] Kleiber, C., & Zeileis, A. (2008). Applied econometrics with R. Springer Science Business Media.

[3] Wooldridge, J. M. (2006). Introducción a la econometría. Un enfoque moderno: un enfoque moderno. Editorial Paraninfo.