Clase:
Regresión Múltiple

Orlando Joaqui Barandica
orlando.joaqui@javerianacali.edu.co

Pontificia Universidad Javeriana de Cali

2023

Tabla de Contenido

Regresión Lineal Múltiple

En la vida real una variable se relaciona no sólo con una sino con muchas más variables, por ejemplo, un agente de bienes raíces está interesado en determinar el Valor Comercial de un inmueble con base en el área construida (en m2), el Número de habitaciones, el Valor comercial de las casas vecinas, el tipo de inmueble (casa o apartamento) y la Antigüedad de la construcción.

Los procedimientos de regresión lineal múltiple son ampliamente usados en investigación de tipo transversal (observaciones referidas en un mismo instante de tiempo, por ejemplo: análisis de una encuesta realizada a un grupo de empresarios para determinar las implicaciones del tratado de libre comercio con los Estados Unidos) y en datos de serie de tiempo.

Regresión Lineal Múltiple

El objetivo es buscar un modelo de regresión que explique a una variable dependiente (variable Y) a través de varias variables independientes (variables $X_1, X_2, ..., X_k$) relacionadas linealmente mediante la expresión:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e \tag{1}$$

- La validez de este modelo requiere del cumplimiento de varias hipótesis sobre los parámetros $\beta_1, ..., \beta_k$ y sobre el error e.
- Cada coeficiente β_i mide el efecto marginal por cada cambio unitario en X_i sobre Y dejando las demás variables explicativas constantes.

Regresión Lineal Múltiple

El modelo es estimado mediante el método de los mínimos cuadrados tal como se hizo para el modelo de regresión lineal simple. Se presenta una notación matricial para las k variables regresoras y las n observaciones representadas:

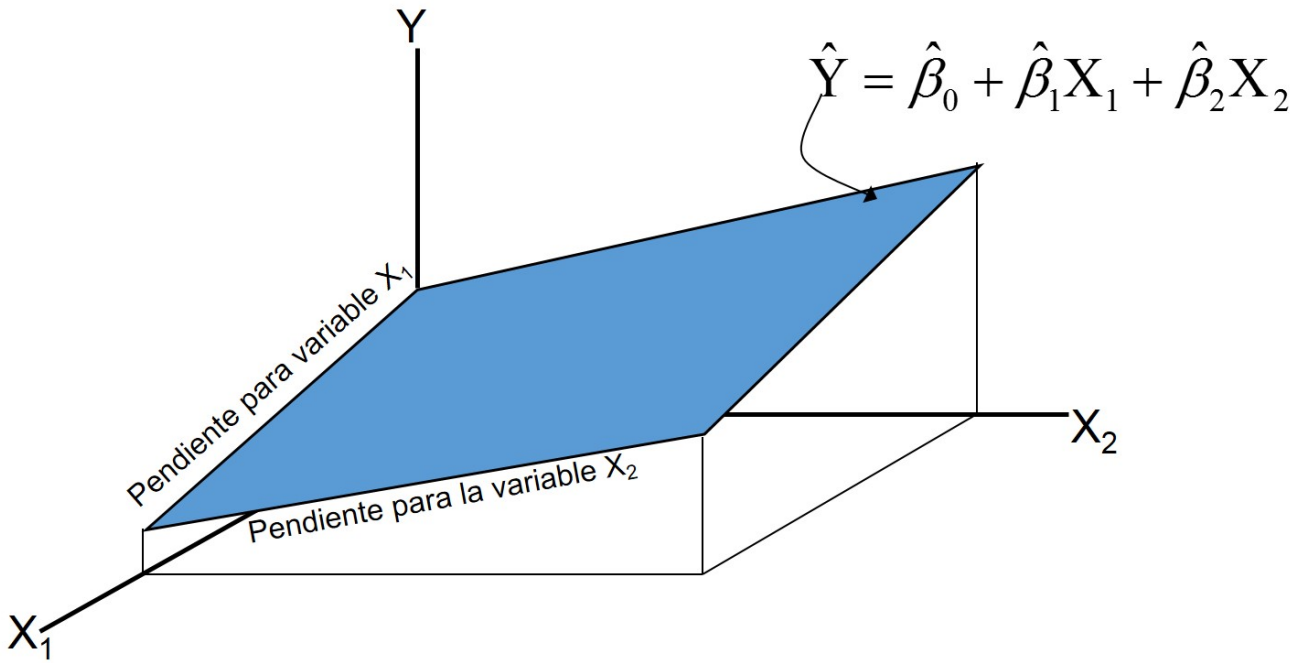
$$Y = Xb + e \tag{2}$$

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1\, x_{11}\, x_{12} \ldots x_{1k} \\ 1\, x_{21}\, x_{22} \ldots x_{2k} \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ . \quad . \quad . \quad . \\ 1\, x_{n1}\, x_{n2} \ldots x_{nk} \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \\ . \\ . \\ . \\ b_k \end{bmatrix} \quad \text{y} \quad \boldsymbol{e} = \begin{bmatrix} e_1 \\ e_2 \\ . \\ . \\ . \\ e_n \end{bmatrix}$$

Regresión Lineal Múltiple

Aplicando el método de mínimos cuadrados a la forma matricial se encuentran las estimaciones de los parámetros β

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{3}$$



Regresión Lineal Múltiple

Ajuste un modelo de regresión lineal múltiple a los datos del Caso de estudio de Calimuebles con variable dependiente: Ausentismo (ausen) y variables explicativas: Edad (X_1), antigüedad (X_2), Salario (X_3) y Número de hijos (X_4) y responda:

- Escriba el modelo de regresión lineal múltiple
- Interprete cada uno de los coeficientes del modelo

Regresión Lineal Múltiple

Respuestas:

- $\hat{(Ausentismo)} = 10{,}321 - 0{,}051 * Edad - 0{,}149 * Antiguedad - 0{,}271 * Salario + 1{,}054 * No_{hijos}$

Regresión Lineal Múltiple

Respuestas:

- $\hat{(Ausentismo)} = 10{,}321 - 0{,}051 * Edad - 0{,}149 * Antiguedad - 0{,}271 * Salario + 1{,}054 * No_{hijos}$
- $\hat{\beta}_0 = 10{,}321$: Como no tiene sentido una Edad de cero años, se concluye que 10.321 es una valor de ajuste al modelo.

Regresión Lineal Múltiple

Respuestas:

- $\widehat{(Ausentismo)} = 10{,}321 - 0{,}051 * Edad - 0{,}149 * Antiguedad - 0{,}271 * Salario + 1{,}054 * No_{hijos}$
- $\hat{\beta}_0 = 10{,}321$: Como no tiene sentido una Edad de cero años, se concluye que 10.321 es una valor de ajuste al modelo.
- $\hat{\beta}_1 = -0{,}051$: Dejando fijas las variables Antigüedad, Salario, No hijos, por cada año adicional en la Edad, el Ausentismo laboral disminuye en 0.051 días en promedio.

Regresión Lineal Múltiple

Respuestas:

- $\hat{(Ausentismo)} = 10{,}321 - 0{,}051 * Edad - 0{,}149 * Antiguedad - 0{,}271 * Salario + 1{,}054 * No_{hijos}$
- $\hat{\beta}_0 = 10{,}321$: Como no tiene sentido una Edad de cero años, se concluye que 10.321 es una valor de ajuste al modelo.
- $\hat{\beta}_1 = -0{,}051$: Dejando fijas las variables Antigüedad, Salario, No hijos, por cada año adicional en la Edad, el Ausentismo laboral disminuye en 0.051 días en promedio.
- ...

Regresión Lineal Múltiple

Respuestas:

- $\hat{(Ausentismo)} = 10{,}321 - 0{,}051 * Edad - 0{,}149 * Antiguedad - 0{,}271 * Salario + 1{,}054 * No_{hijos}$
- $\hat{\beta}_0 = 10{,}321$: Como no tiene sentido una Edad de cero años, se concluye que 10.321 es una valor de ajuste al modelo.
- $\hat{\beta}_1 = -0{,}051$: Dejando fijas las variables Antigüedad, Salario, No hijos, por cada año adicional en la Edad, el Ausentismo laboral disminuye en 0.051 días en promedio.
- ...
- $\hat{\beta}_4 = 1{,}054$: Dejando fijas las variables Edad, Antigüedad y Salario, por cada hijo adicional, el Ausentismo laboral aumenta en 1.054 días en promedio.

Anova

Evaluación de la significancia del modelo: Evalúa que tan bien se ajustan los datos a una ecuación de regresión múltiple.

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Promedio de los Cuadrados	f	Valor P
Regresión	k	SSR	$MSR = \frac{SSR}{k}$	$f_c = \frac{MSR}{MSE}$	$Valor\ p = P(f > f_c)$
Residuos	$n - (k + 1)$	SSE	$MSE = \frac{SSE}{n-(k+1)}$		
Total	$n - 1$	Total			

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Donde:

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$: Suma de cuadrados Total

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$: Suma de Cuadrados de la Regresión

$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$: Suma de Cuadrados del Error

Los grados de libertad se obtienen de la siguiente manera:

Para la regresión: k, numero de variables independientes.

Para los residuos: $n - (k + 1)$, numero de datos menos el número de parámetros estimados.

Anova

Prueba global: Prueba del modelo de regresión múltiple.

- Prueba F para la significación global del modelo. Muestra si existe una relación lineal entre todas las variables X e Y conjuntamente consideradas.
- El valor del estadístico F se obtiene mediante la razón:

$$F = \frac{SSR/k}{SSE/(n-k-1)} \tag{4}$$

$H_0 : \beta_1 = ... = \beta_k = 0$ (Ninguna relación lineal)

H_1 : al menos un $\beta_i \neq 0$ (al menos una variable independiente afecta Y)

Coeficiente de determinación múltiple R^2

Informa de la proporción de la variación total en Y explicada por todas las variables X en su conjunto.

$$R^2 = \frac{SSR}{SST} \tag{5}$$

El coeficiente de determinación tiene problemas ya que su valor aumenta introduciendo nuevas variables en el modelo aunque su efecto no sea significativo, por lo que siempre se puede aumentar artificialmente R^2 , lo que llevaría a malas interpretaciones.

Coeficiente de determinación múltiple R^2

Informa de la proporción de la variación total en Y explicada por todas las variables X en su conjunto.

$$R^2 = \frac{SSR}{SST} \tag{5}$$

El coeficiente de determinación tiene problemas ya que su valor aumenta introduciendo nuevas variables en el modelo aunque su efecto no sea significativo, por lo que siempre se puede aumentar artificialmente R^2 , lo que llevaría a malas interpretaciones.

Por esta razón se define el coeficiente de determinación ajustado que tiene en cuenta los grados de libertad implicados en el modelo.

R^2 Ajustado

Muestra la proporción de variación en Y explicada por todas las variables X ajustados por el número de variables X utilizados

$$R^2_{ajus} = R^2 - (1 - R^2)\frac{k}{n - k - 1} \tag{6}$$

Útil para comparar los modelos

Significancia individual de los β_i

El aporte de cada variable al modelo se puede evaluar planteando las hipótesis:

- $H_0 : \beta_i = 0$ (La contribución de la variable i no es significativa)
- $H_1 : \beta_i \neq 0$ (La contribución de la variable i es significativa)

Para la prueba de estas hipótesis se recurre al análisis de varianza dada las estadísticas de prueba o los valores de probabilidad $(Valores\_p)$ respectivos a cada hipótesis.

Recuerde

$$t = \frac{\beta_i - 0}{Se(\beta_i)} \tag{7}$$