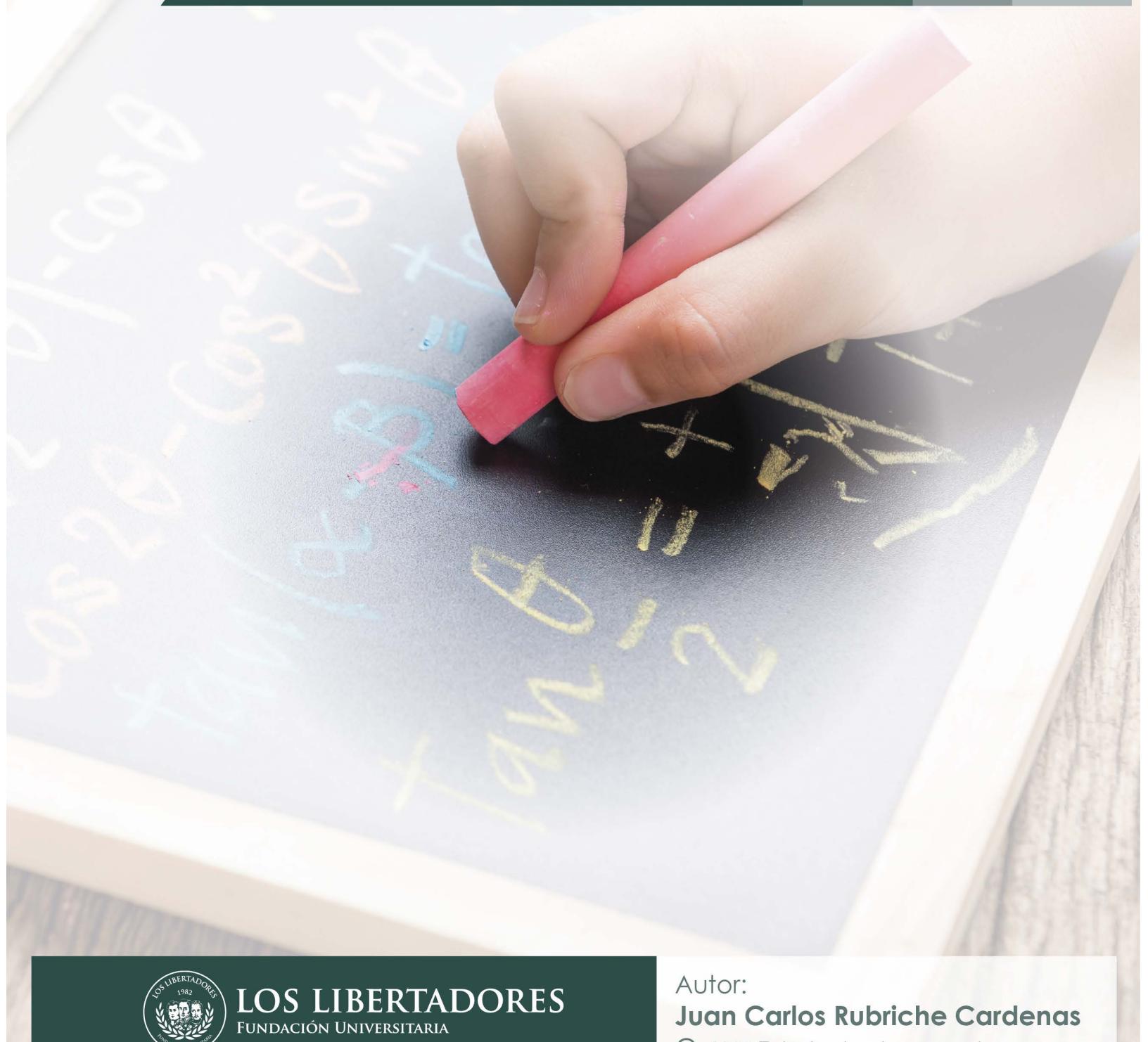




# Curso en **Análisis de regresión**

Unidad 2 • Modelo de regresión lineal múltiple



# LOS LIBERTADORES

## FUNDACIÓN UNIVERSITARIA

Autor:  
**Juan Carlos Rubriche Cárdenas**  
© Juan Carlos Rubriche Cárdenas



## Preguntas orientadoras

- ¿Qué es un modelo de regresión lineal múltiple y cuáles son sus supuestos?
- ¿Qué métodos estadísticos hay para estimar los parámetros del modelo?
- ¿Se puede hacer inferencia sobre los parámetros del modelo?
- ¿Qué tan confiables son las estimaciones de futuras observaciones usando la ecuación de regresión múltiple estimada?
- ¿Cómo podemos verificar si los supuestos del modelo se satisfacen analizando los residuales?
- ¿Qué tan bien se ajusta el modelo de regresión múltiple estimado a los datos observados?

## Síntesis

En esta unidad se estudia el modelo de regresión lineal múltiple utilizado para describir la relación de dependencia entre una variable respuesta y múltiples variables predictoras. Se muestra cómo usar el método de los mínimos cuadrados ordinarios para estimar los parámetros del modelo, usando álgebra de matrices y cómo construir intervalos de confianza para estimar los parámetros y la respuesta media; además de pruebas de hipótesis para mostrar la significancia de las variables en el modelo. Finalmente, se hace una breve introducción al análisis residual para verificar los supuestos del modelo.

## Palabras clave

- Modelo de regresión múltiple
- Ecuación de estimación
- Método de mínimos cuadrados
- Prueba de hipótesis
- Intervalo de confianza
- Residuales





## Tema 1

### Modelo de regresión múltiple

En esta sección mostramos como describir la relación de dependencia entre una variable respuesta y dos o más variables predictoras usando un modelo de regresión lineal múltiple. Por ejemplo, las directivas de un hospital están interesadas en investigar la relación de dependencia entre la satisfacción de un paciente y las variables: Edad, índice de gravedad de su enfermedad, tipo de paciente ( 0: médico o 1: quirúrgico ) y un índice de ansiedad. En la tabla 1 se muestra un conjunto de 25 observaciones de pacientes atendidos por este hospital.

Obs.	Edad	Gravedad	Tipo-paciente	Ansiedad	Satisfacción
1	55	50	0	2.1	68
2	46	24	1	2.8	77
3	30	46	1	3.3	96
4	35	48	1	4.5	80
5	59	58	0	2.0	43
6	61	60	0	5.1	44
7	74	65	1	5.5	26
8	38	42	1	3.2	88
9	27	42	0	3.1	75
10	51	50	1	2.4	57
11	53	38	1	2.2	56
12	41	30	0	2.1	88
13	37	31	0	1.9	88
14	24	34	0	3.1	102
15	42	30	0	3.0	88
16	50	48	1	4.2	70
17	58	61	1	4.6	52
18	60	71	1	5.3	43
19	62	62	0	7.2	46
20	68	38	0	7.8	56
21	70	41	1	7.0	59
22	79	66	1	6.2	26
23	63	31	1	4.1	52
24	39	42	0	3.5	83
25	49	40	1	2.1	75

Tabla 1. Datos pacientes hospital

En este ejemplo tenemos una variable respuesta (nivel de satisfacción del paciente) y varias variables predictores (edad, gravedad, tipo-paciente, ansiedad). Esta relación de dependencia entre estas variables se podría describir por medio de un modelo de





regresión lineal múltiple, que cuando relaciona la variable respuesta  $Y$  y  $k$  variables predictoras  $X_1, X_2, \dots, X_k$ , tiene la siguiente forma general:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon, \quad (1)$$

donde las constantes  $\beta_0, \beta_1, \dots, \beta_k$  son los parámetros desconocidos del modelo a estimar y  $\varepsilon$  es una variable aleatoria que se interpreta como la variabilidad en  $Y$  que no puede ser explicada por el efecto lineal de  $k$  variables predictoras.

## Tema 2

### Interpretación de $\beta_i$

El parámetro  $\beta_i$  del modelo (1) se interpreta como el cambio esperado en la variable respuesta  $Y$  por un cambio de una unidad en la variable  $X_i$  cuando las demás variables predictoras  $X_j$  ( $j \neq i$ ) se mantienen constantes.

Si suponemos que el término de error aleatorio satisface que  $E(\varepsilon)=0$  cuando las variables predictoras toman valores fijos, concluimos que:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k, \quad (2)$$

La expresión (2) se denomina ecuación de regresión múltiple cuya representación gráfica es un hiperplano en el espacio  $k + 1$  dimensional.





## Tema 3

### Estimación de parámetros por el método de los mínimos cuadrados

Vamos a estimar los parámetros  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  del modelo de regresión lineal múltiple (1) a partir de un conjunto de  $n$  observaciones, suponiendo que el número de observaciones es mayor que el número de variables predictoras ( $n > k$ ). En este problema, las observaciones son una muestra de datos multivariados que podemos representar como sigue:

Obs.	Respuesta Y	Predictoras			
		X <sub>1</sub>	X <sub>2</sub>	...	X <sub>k</sub>
1	Y <sub>1</sub>	X <sub>11</sub>	X <sub>12</sub>	...	X <sub>1k</sub>
2	Y <sub>2</sub>	X <sub>21</sub>	X <sub>22</sub>	...	X <sub>2k</sub>
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
n	Y <sub>n</sub>	X <sub>n1</sub>	X <sub>n2</sub>	...	X <sub>nk</sub>

Usando estos datos podemos escribir un modelo de regresión muestral correspondiente al modelo (1), así:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3)$$

donde los términos de error aleatorios  $\varepsilon_i$  se supone que tiene media 0, varianza constante y no están correlacionados.

El modelo (3), se puede representar matricialmente de la siguiente forma:

$$y = X\beta + \varepsilon$$

en donde:





$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$
$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

El método de mínimos cuadrados consiste en hallar el vector:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

que minimice la función:

$$S(\beta_0, \beta_1, \dots, \beta_k) = \sum_{i=1}^n \varepsilon_i^2 = (y - X\beta)'(y - X\beta).$$

con respecto a los parámetros  $\beta_0, \beta_1, \dots, \beta_k$ .

El estimador de mínimos cuadrados  $\hat{\beta}$  del vector de parámetros  $\beta$ , se calcula mediante la expresión:

$$\hat{\beta} = (X'X)^{-1}X'y \quad (4)$$

siempre que la matriz inversa  $(X'X)^{-1}$  exista<sup>1</sup>. La matriz  $X'X$  es invertible si sus columnas son vectores linealmente independientes.

<sup>1</sup> La notación A' simboliza la matriz transpuesta de A





Nota: el estimador de mínimos cuadrados  $\hat{\beta}$  es el mejor estimador lineal insesgado de  $\beta$ . Este resultado es conocido como teorema de Gauss - Markov.

Usando el estimador  $\hat{\beta}$  de  $\beta$ , se construye la ecuación de regresión múltiple estimada que se define como sigue:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \dots + \hat{\beta}_k X_k \quad (5)$$

A partir de la ecuación de estimación, calculamos el vector de los valores ajustados  $\hat{y}$  correspondientes a las respuestas observadas  $y_i$ , mediante:

$$\begin{aligned}\hat{y} &= X\hat{\beta} \\ &= Hy\end{aligned}$$

donde  $H = X(X'X)^{-1}X'$  es la matriz sombrero que representa la transformación lineal que convierte vectores de valores observados en vectores de valores estimados.

Los residuales  $e_i$  del modelo de regresión lineal múltiple se definen como las diferencias entre los valores observados  $y_i$  y los valores estimados  $\hat{y}_i$ , es decir,  $e_i = y_i - \hat{y}_i$ . Usando notación matricial, los  $n$  residuales se representan como:

$$\begin{aligned}e &= y - \hat{y} \\ &= (I - H)y\end{aligned}$$

donde  $I$  es la matriz identidad de orden  $n$ .

### Ejemplo

Vamos a calcular la ecuación de regresión estimada para los datos obtenidos de una encuesta de satisfacción aplicada a los pacientes de un hospital (ver tabla 1). La ecuación de estimación para esta situación, se puede escribir como sigue:

$$\hat{S} = \hat{\beta}_0 + \hat{\beta}_1 E + \hat{\beta}_2 G + \hat{\beta}_3 TP + \hat{\beta}_4 A$$





donde la variable S (Satisfacción) es la variable respuesta y las variables E (Edad), G (Gravedad), TP (Tipo-paciente) y A (Ansiedad), son las variables predictoras. El estimador de mínimos cuadrados  $\hat{\beta}$  se obtiene mediante la ecuación

$$\hat{\beta} = (X'X)^{-1}X'y$$

Ejemplo:

Por lo tanto, la ecuación de regresión estimada es:

$$\hat{S} = 143,8672 - 1,1172E - 0,5862G + 0,4149TP + 1,3064A$$

Esta ecuación nos permite predecir el grado de satisfacción de un paciente en función de las variables edad, gravedad, tipo de paciente y ansiedad.

Usando el programa R, podemos calcular el vector de valores estimados  $\hat{y}$  y el correspondiente vector de residuales  $e = Y - \hat{y}$  usando las funciones: fitted (nombre modelo lineal) y residuals (nombre modelo lineal).

$$\hat{y} = \begin{pmatrix} 55,85524 \\ 82,48064 \\ 88,11200 \\ 82,92132 \\ 46,56621 \\ 47,20912 \\ 30,69218 \\ 81,38880 \\ 93,13223 \\ 61,13073 \\ 65,66963 \\ 83,21994 \\ 86,84116 \\ 101,17345 \\ 83,27848 \\ 65,77176 \\ 49,73614 \\ 42,55412 \\ 47,66286 \\ 55,81267 \\ 51,18948 \\ 25,43453 \\ 61,08341 \\ 80,24865 \\ 68,83528 \end{pmatrix}, \quad e = \begin{pmatrix} 12,1447623 \\ -5,4806372 \\ 7,8879960 \\ -2,9213180 \\ -3,5662065 \\ -3,2091186 \\ -4,6921771 \\ 6,6112036 \\ -18,1322342 \\ -4,1307260 \\ -9,6696335 \\ 4,7800638 \\ 1,1588368 \\ 0,8265468 \\ 4,7215250 \\ 4,2282432 \\ 2,2638620 \\ 0,4458802 \\ -1,6628567 \\ 0,1873314 \\ 7,8105246 \\ 0,5654735 \\ -9,0834067 \\ 2,7513501 \\ 6,1647154 \end{pmatrix}$$





## Estimación de la varianza $\sigma^2$

De la misma manera que en el caso del modelo de regresión lineal simple, se estima la varianza del modelo usando la suma de cuadrados de los residuales. Así, un estimador insesgado de  $\sigma^2$  para un modelo de regresión lineal múltiple con  $p = k + 1$  parámetros es:

$$\hat{\sigma}^2 = \frac{SCE}{n - p} = \frac{1}{n - p} \sum_{i=1}^n e_i^2$$

En el ejemplo de los pacientes, la suma de cuadrados de los residuales es:

$$SCE = 1038,947$$

Además, hay  $n = 25$  observaciones y  $p = 5$  parámetros en el modelo, así una estimación puntual de  $\sigma^2$  es:

$$\hat{\sigma}^2 = \frac{1038,947}{25 - 5} = 51,94733$$

## Análisis de varianza

En el modelo de regresión múltiple, también es posible particionar la suma total de cuadrados (SCT) con  $n - 1$  grados de libertad en dos componentes: la suma de cuadrados de la regresión (SCR) con  $k$  grados de libertad y la suma de cuadrados del error (SCE) con  $n - p$ , es decir,

$$SCT = SCR + SCE$$

Continuando con nuestro ejemplo de los pacientes, tenemos que:

Fuente de variación	Suma de cuadrados	Grados de libertad
Regresión	SCR= 9739,293	$k = 4$
Residuales	SCE= 1038,947	$n - p = 20$
Total	SCT= 10778,24	$n - 1 = 24$





Para medir la bondad de ajuste de la ecuación de regresión lineal múltiple estimada empleamos el coeficiente de regresión múltiple  $R^2$  que se define como

$$R^2 = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT}$$

Este coeficiente de determinación se interpreta como la proporción de variabilidad observada en la variable respuesta que es explicada por el modelo de regresión lineal estimado en las variables predictoras bajo consideración.

#### Ejemplo

El coeficiente de determinación  $R^2$  del modelo de regresión estimado para explicar la relación entre la Satisfacción (S) de los pacientes de un hospital y las variables predictoras: Edad (E), Gravedad (G), Tipo de paciente (TP) y Ansiedad (A); es:

$$R^2 = \frac{9739,293}{10778,24} = 0,9036$$

Por lo tanto, el 90,36% de la variabilidad en el índice de satisfacción de los pacientes es explicada por el modelo de regresión lineal múltiple estimado en las variables predictoras: Edad (E), Gravedad (G), Tipo de paciente (TP) y Ansiedad (A).

#### Ejemplo

El coeficiente de determinación se ve afectado por el numero de variables predictoras involucradas en el modelo, es decir, al aumentar el número de variables aumenta el valor de  $R^2$ . Por esta razón, muchos analistas prefieren utilizar el coeficiente de determinación múltiple ajustado  $R^2_{adj}$ , para medir la bondad de ajuste del modelo estimado y hacer comparaciones entre modelos, el cual se define como:

$$R^2_{adj} = 1 - \frac{SCE / (n - p)}{SCT / (n - 1)}$$

Se observa que  $R^2_{adj}$  aumenta solo si se adicionan variables que reduzcan  $SCE/(n-p)$ . Así, el modelo que mide la satisfacción de los pacientes en un hospital en función de su edad,





gravedad, tipo de paciente y ansiedad tiene un coeficiente de determinación ajustado igual a 0,8843.

## Tema 4

### Inferencia en regresión múltiple

En este apartado vamos a mostrar como realizar pruebas de hipótesis y construir intervalos de confianza en regresión lineal múltiple.

#### Pruebas de significancia

En regresión lineal múltiple se usa la prueba F para probar la significancia global del modelo y la prueba t para probar la significancia de cada uno de los parametros del modelo individualmente.

La prueba de significancia global del modelo se plantean las siguientes hipótesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

$H_1$  : Por lo menos uno de estos parámetros es distinto de cero

Para esta prueba, se usa el siguiente estadístico de prueba:

$$F_0 = \frac{CMR}{CME}$$

donde  $CMR = SCR/k$  y  $CME = SCE/(n-p)$ . Este estadístico de prueba tiene una distribución F, con  $k$  y  $n-p$  grados de libertad en el numerador y en el denominador respectivamente, cuando  $H_0$  es verdadera.

Se espera que cuando  $H_0$  es verdadera, entonces el valor calculado del estadístico de prueba  $f_0$  sea un valor “pequeño”. Así, que para un nivel de significancia  $\alpha$  rechace  $H_0$  si





$f_0 > f_{\alpha, k, n-p}$  donde  $f_{\alpha, k, n-p}$  es el quantil de la distribución F con grados de libertad  $k$  y  $n-p$  que deja un área igual a  $\alpha$  a su derecha, o si valor-p  $\leq \alpha$ .

En conclusión, la hipótesis nula es rechazada, si al menos una de las variables predictoras en el modelo esta linealmente relacionada con la variable respuesta.

Para probar la significancia del modelo de regresión lineal en el caso del ejemplo de los pacientes bajo consideración, se plantean las siguientes hipótesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$H_1$  : Por lo menos uno de estos parámetros es distinto de cero

El valor calculado del estadístico de prueba F0 es:

$$f_0 = \frac{9739,293 / 4}{1038,947 / 20} = 46,87$$

Como el valor-p =  $6,951 \times 10^{-10}$ , esto muestra que podemos rechazar con un nivel de significancia muy bajo la hipótesis nula. Se concluye que por lo menos alguna variable predictora esta relacionada linealmente con respecto a la variable respuesta.

## Inferencia sobre los parámetros del modelo individualmente

Las inferencias sobre cada parámetro  $\beta_i$  están basadas en el estadístico:

$$T = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)}$$

que tiene una distribución t con  $n - p$  grados de libertad.

Específicamente, para probar la significancia del parámetro  $\beta_i$  se plantean las siguientes hipótesis:

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$





Si asumimos que la hipótesis nula es verdadera, estas hipótesis se prueban usando el estadístico de prueba:

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

La hipótesis nula  $H_0$  es rechazada si  $|t_0| > t_{\alpha/2, n-p}$  o si valor-p  $\leq \alpha$ , para un determinado nivel de significancia  $\alpha$  dado.

Por otra parte, un intervalo de confianza de  $(1 - \alpha)100\%$  para el parámetro  $\beta_i$  es:

$$\hat{\beta}_i - t_{\alpha/2, n-p} se(\hat{\beta}_i) \leq \beta_i \leq \hat{\beta}_i + t_{\alpha/2, n-p} se(\hat{\beta}_i)$$

## Tema 5

### Intervalos de confianza para la respuesta media e intervalos de predicción

La ecuación de regresión estimada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

se puede utilizar para hacer una estimación puntual de la respuesta media,  $y^* = E(y)$ , para los valores particulares  $x_1^*, x_2^*, \dots, x_k^*$  de las variables predictoras  $x_1, x_2, \dots, x_k$ , respectivamente. Denotaremos esta estimación como  $\hat{y}^*$  y su error estándar como  $se(\hat{y}^*)$ .

Por lo tanto, un intervalo de confianza de  $(1 - \alpha)100\%$  para la respuesta media en el punto  $x_1 = x_1^*, x_2 = x_2^*, \dots, x_k = x_k^*$  en un modelo de regresión lineal múltiple está dado por:

$$\hat{y}^* - t_{\alpha/2, n-p} se(\hat{y}^*) \leq y^* \leq \hat{y}^* + t_{\alpha/2, n-p} se(\hat{y}^*)$$





## Bibliografía

- Montgomery D., Peck E. y Vining G. (2006). Introducción al análisis de regresión lineal. Tercera edición. México: CECSA
- Kutner M., Nachtsheim C., Neter J. y Li W. (2005). Applied Linear Statistical Models. Quinta edición. New York: McGraw-Hill.
- Sheather S. (2009). A Modern Approach to Regression with R. New York: Springer.
- Weisberg S. (2005). Applied Linear Regression. Tercera edición. New Jersey: Wiley.