Modelo de regresión y sus aplicaciones en la administración

HERNAN BEJARANO BARRERA*

INTRODUCCION

Los métodos estadísticos suelen constituirse en el soporte apropiado para obtener la mejor y más oportuna información de los elementos bajo investigación y llegar a una decisión en condiciones de riesgo e incertidumbre. Entre dichos métodos se encuentran los referentes al estudio de los elementos mediante Modelos de Regresión. Se pretende con este escrito proporcionar los conceptos y metodología básica necesaria para examinar técnicas que permitan ajustar una ecuación de algún tipo al conjunto de datos dado, con el fin de obtener una ecuación de predicción razonablemente precisa y que proporcione un modelo teórico que no está disponible.

I- MODELO GENERAL DE REGRESION LINEAL

Sean K + 1 variables de las cuales K son independientes ($X_1, X_2, ..., X_K$) denominados de estímulo y la otra variable Y dependiente o la respuesta.

La ecuación:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots$$

$$+ \beta_K X_{Ki} + U_i \quad \text{para todo.}$$

$i = 1, 2, ..., n$ se llama modelo general de regresión lineal, donde

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_K$$

son constantes lineales desconocidas y cada

$$\beta_j, j = 1, 2, ..., K$$

representa el cambio en la respuesta promedio para un cambio igual a una unidad de la correspondiente variable de predicción X_j, cuando todas las demás están constantes. Además $U_1, U_2, ..., U_n$ son variables aleatorias no observables con valor esperado cero y con varianza constante desconocida (Var).

$$E(U_i) = 0$$

$$Var(U_i) = Var$$

* Estadístico Universidad Nacional
Coordinador Area Cuantitativa Facultad Ingeniería de Sistemas
EAN

II- NOTACION MATRICIAL DEL MODELO GENERAL DE REGRESION LINEAL

Dada una muestra de n observaciones $Y_1, Y_2, \ldots, Y_n$, entonces el modelo genera el siguiente sistema:

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_K X_{K_1} + U_1$$

$$Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \ldots + \beta_K X_{K_2} + U_2$$

$$Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots + \beta_K X_{Kn} + U_n$$

Este sistema de n ecuaciones plantea la siguiente forma matricial:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & X_{31} & X_{K1} \\ 1 & X_{12} & X_{22} & X_{32} & X_{K2} \\ 1 & X_{1n} & X_{2n} & X_{3n} & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_K \end{bmatrix} + \begin{bmatrix} U_1 \\ U_2 \\ U_n \end{bmatrix}$$

$(n \times 1) \quad n \times (K+1) \quad (K+1) \times 1 \quad (n \times 1)$

Que puede expresarse como:

$$Y = X\beta + U$$

Donde:

Y = Es un vector aleatorio cuyos elementos son observables.

X = Es una matriz de valores fijos o no aleatorios que en el modelo se consideran constantes o predeterminados.

β = Es un vector de elementos constantes (escalares) lineales desconocidos.

U = Es un vector de elementos independientes no observables.

III- SOLUCION DEL MODELO GENERAL DE REGRESION LINEAL

Para obtener los estimadores de mínimos cuadrados de los parámetros $\beta_0, \beta_1, \ldots, \beta_K$ se generalizará un conjunto de datos consistentes en n observaciones.

El método de mínimos cuadrados considera la desviación del valor observado Y_i con respecto a su valor estimado $\hat{Y}_i$ y determina los valores de β_j que minimizan la suma de los cuadrados de estas desviaciones.

Sea $E_i = (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \beta_K X_{Ki})$

la iésima desviación.

$$\Rightarrow \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_2 - \hat{\beta}_0 - \beta_1 X_{1i} - \beta_K X_{Ki})^2$$

es la suma de los cuadrados de los errores.

Los estimadores de mínimos cuadrados de β se obtienen diferenciando la suma de cuadrados de los errores con respecto a cada β_j en forma parcial e igualando a cero. Al hacer las simplificaciones y operaciones necesarias se calculan las siguientes ecuaciones normales:

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \Sigma X_{1i} + \beta_2 \Sigma X_{2i} + \ldots + \beta_K \Sigma X_{Ki}$$

$$\sum_{i=1}^{n} X_{1i} Y_i = \beta_0 \Sigma X_{1i} + \beta_1 \Sigma X_{1i}^2 + \beta_2 \Sigma X_{1i} X_{2i} + \ldots \beta_K \Sigma X_{Ki} X_{1i}$$

$$\vdots \qquad\qquad\qquad\qquad\qquad\qquad \vdots$$

$$\sum_{i=1}^{n} X_{Ki} Y_i = \beta_0 \Sigma X_{Ki} + \beta_1 \Sigma X_{1i} X_{Ki} + \beta_2 \Sigma X_{2i} X_{Ki} + \ldots + \beta_K \sum_{i=1}^{n} X_{Ki}^2$$

Empleando la notación matricial obtenemos:

$$\begin{bmatrix} \Sigma Y_i \\ \Sigma X_{1i} Y_1 \\ \Sigma X_{Ki} Y_i \end{bmatrix} \begin{bmatrix} n & \Sigma X_{1i} & \Sigma X_{Ki} \\ \Sigma X_{1i} & \Sigma X_{1i}^2 & \Sigma X_{1i} X_{Ki} \\ \Sigma X_{ki} & \Sigma X_{1i} X_{Ki} & \Sigma X_{Ki}^2 \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_K \end{bmatrix}$$

En la forma general queda el sistema como:

$$X'Y = (X'X)\beta$$

Donde $X'X$ es una matriz simétrica y si tiene inversa la solución del modelo es:

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

Que es válida para cualquier tipo de modelo lineal o no lineal.

IV- SUMA DE CUADRADOS

Una vez obtenida la solución del modelo se pueden plantear la descomposición de la suma de cuadrados en términos de total, de variaciones explicadas y variaciones no explicadas por el modelo de regresión o debidas al error.

a. Suma de cuadrados totales (SCT)

$$SCT = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = Y'Y - n(\bar{Y})^2$$

b. Suma de cuadrados de las variaciones explicadas por el modelo de regresión.

$$SCE_X = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \beta^1 (X^1 Y) - n\bar{Y}^2$$

c. Suma de cuadrados de las variaciones no explicadas o residuales por el modelo o debidas al error.

$$SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = Y^1 Y - \beta^1 (X^1 Y)$$

d. Además se tiene que:

$$SCT = SCE_X + SCR$$

A partir de las anteriores sumas expresadas en notación matricial podemos plantear una primera parte referente a:

1. Coeficiente de Determinación R^2

Mide el porcentaje de la información que es explicada por el modelo de regresión. Se define como:

$$R^2 = \frac{SCE_X}{SCT}$$

y presenta un rango de variación dado por:

$$0 \leq R^2 \leq 1$$

donde el R^2 permite validar el modelo de regresión.

Si R^2 es un valor cercano a uno nos indica que el modelo está explicando un alto porcentaje de la información, en caso contrario significa que la ecuación de regresión utilizada no es la adecuada para la predicción.

2. Análisis de Varianza

Con esta prueba estadística se quiere comprobar si todas las variables (parámetros de regresión) en conjunto no inciden en el modelo propuesto, es decir, parámetros igual a cero, o, si inciden en forma conjunta. Las hipótesis a probar es:

$$H_0 : \beta_1 = \beta_2 = \beta_K = 0$$

$$H_1 : \beta_j = 0 \text{ para algún } j = 1, 2, \ldots, K.$$

Para establecer la prueba estadística empleamos la siguiente tabla de análisis de varianza:

FUENTE DE VARIACION	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADO MEDIO	PRUEBA F
Debido a la regresión.	$(K+1)-1$	SCE_X	$\frac{SCE_X}{K}$	$F_C = \frac{CME_X}{CMR}$
Debido a los residuos.	$n-(K+1)$	SCR	$\frac{SCR}{n-(K+1)}$	
TOTAL	$n-1$	SCT		

La prueba

$$F_c = \frac{\text{Cuadrado medio explicado}}{\text{Cuadrado medio del error}}$$

Como toda hipótesis tiene que contrastarse con un valor tabulado, planteamos la siguiente región crítica:

$$P_r (F_e > F_{T_1}) = \alpha/2$$

$$P_r (F_e < F_{T_2}) = \alpha/2$$

Si se cumple una de las anteriores condiciones, se rechaza la hipótesis nula H_0, con un riesgo de tipo I dado por α.

Los valores tabulados se determinaron de la siguiente forma para un $\alpha = 0{,}05$.

$$F_{T_1} (K; n-(K+1); 0{,}975)$$

$$F_{T_2} (K, n-(K+1; 0{,}025) = \frac{1}{F_{T_2} (n-(K+1); K; 0{,}975}$$

3. Intervalo de Confianza para Y

Se desea estimar el valor de Y dependiendo de los valores $X_1, X_2, ..., X_K$ con un nivel alto de confiabilidad. Para esto es necesario calcular el error de la estimación o su varianza que corresponde al cuadrado medio del error, obtenido en el análisis de varianza.

El error de estimación $\partial\, Y/X$ se obtiene a partir de la siguiente relación:

$$\partial\, Y/X = \sqrt{\frac{SCR}{n-(K+1)}}$$

donde K +1 es el total de variables analizadas.

El intervalo de predicción es:

$$P_r\ (\hat{Y}_i \pm Z\ \ \partial\, Y/X_{ij}) = 1-\alpha\ .$$

Dentro de un análisis de regresión también podemos analizar en forma independiente el comportamiento de cada uno de los parámetros β_j.

Para esto necesitamos diseñar una nueva matriz que es la matriz de varianzas y covarianzas para los β_j que nos mide la variación y la variación conjunta entre dos elementos de β_j.

Por lo tanto se tiene que:

$$(Var-Cov)\ (\beta) = \partial^2\, Y/X\ (X'\,X)^{-1}$$

lo cual nos da una matriz con los siguientes elementos:

$$(Var-Cov)\ (\beta) = \begin{array}{c} \\ \beta_0 \\ \beta_1 \\ \beta_K \end{array} \begin{array}{c} \begin{array}{cccc} \beta_0 & \beta_1 & & \beta_k \end{array} \\ \begin{bmatrix} \partial^2_{\beta_0} & \partial_{(\beta_0,\beta_1)} \ldots\ldots & \partial_{(\beta_0,\beta_K)} \\ \vdots & \partial^2_{\beta_1} \ldots\ldots & \partial^2_{(\beta_1,\beta_k)} \\ \partial_{(\beta_K\beta_0)} \ldots\ldots & & \partial^2_{\beta_K} \end{bmatrix} \end{array}$$

Esta matriz es simétrica y sobre la diagonal principal se encuentran las varianzas de los valores β_j y en las otras celdas las variaciones conjuntas o covarianzas de β.

Ahora mencionamos las aplicaciones que podemos obtener del empleo de esta matriz.

4. Prueba de la Hipótesis de que uno de los Parámetros β_j es igual a cero.

Esta prueba permite comprobar en forma independiente si una variable X_j incide en forma separada en el comportamiento de la variable dependiente Y.

Podemos plantear la siguiente hipótesis:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Como toda prueba de hipótesis necesita una pureba estadística, para este caso se ha diseñado la siguiente:

$$Z_c = \frac{\hat{\beta}_j - \beta_j}{\partial\beta_j}$$

Toda prueba estadística debe contrastarse con un valor tabulado y se ha diseñado la siguiente región crítica:

$$P_r\ (Z_c > Z_T) = \alpha/2$$

$$P_r\ (Z_c < Z_T) = \alpha/2$$

Para un riesgo del 5% y empleando la distribución normal con un valor $Z_T = \pm\ 1{,}96$ y teniendo en cuenta que si se cumple una de las dos condiciones anteriores se rechaza la hipótesis nula (H_0) lo cual significa que la variable que estamos contrastando, si incide en el comportamiento de la variable dependiente.

5. Un Intervalo de Confianza para un Parámetro de Regresión β_j

El intervalo que vamos a definir nos permite establecer los límites en que puede variar la contribución de una variable X_j a la respuesta de una variable dependiente Y con un alto grado de seguridad.

$$P_r(\hat{\beta}_j \pm Z\ \ \partial\beta_j) = 1-\alpha$$

Donde $\hat{\beta}_j$ es el valor del estimador en el modelo, $\partial\beta_j$ el error estandar de $\hat{\beta}_j$ y Z el desvío estandar de la distribución normal que corresponde a 1,96 para una confiabilidad del 95%.

6. Correlación

Otra medición importante en el análisis de regresión es la correlación que nos permite establecer el

grado de asociación o relación que presentan las variables en forma de parejas o para cualquier conjunto de variables.

a. Correlación múltiple

Esta medición permite establecer en qué porcentaje están asociadas las K variables independientes y la variable dependiente Y.

Se calcula esta medida como la raíz cuadrada del coeficiente de determinación, es decir,

$$R = \sqrt{R^2}$$

b. Correlación simple R

Esta medida define el grado de relación entre un par de variables. Para esto se ha establecido una matriz de correlación R, simétrica que se presenta en forma matricial como

$$R = \begin{array}{c} \\ \beta_0 \\ \beta_1 \\ \\ \beta_K \end{array} \begin{array}{c} \begin{array}{cccc} \beta_0 & & \beta_1 & \beta_K \end{array} \\ \begin{bmatrix} 1 & \partial(\beta_0,\beta_1) & \partial(\beta_0,\beta_K) \\ & 1 & \\ \partial(\beta_K,\beta_0) & & 1 \end{bmatrix} \end{array}$$

Para calcular esta matriz se emplea como referencia la matriz de varianzas covarianzas de β_j.

• La correlación para los parámetros β_i y β_j se define como:

$$\beta_i, \beta_j = \frac{\text{Cov}(\beta_i, \beta_j)}{\sqrt{\partial^2 \beta_i, \partial^2 \beta_j}}$$

Estos coeficientes nos indican si en las variables hay alta asociación (relación directa) o alta disociación (relación inversa).

Ejemplo de un caso

Un analista de una compañía manufacturera desea explicar las variaciones que han ocurrido en el costo de manufactura por unidad del producto (Y) en función del nivel de producción X_1 como porcentaje de la capacidad fija y un índice de los costos de mano de obra y materia prima. La información corresponde al período de 12 meses del año 1987.

PERIODO	COSTO	NIVEL PRODUCCION	INDICE COSTO
1	3,65	85	80
2	4,22	78	93
3	4,29	82	107
4	5,43	64	115
5	6,62	50	130
6	5,71	62	128
7	5,09	70	116
8	3,99	90	92
9	4,08	94	94
10	4,38	100	110
11	4,28	104	115
12	4,42	82	117

Ajustaremos un modelo de la forma:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$$X = \begin{bmatrix} 1 & 85 & 80 \\ 1 & 78 & 93 \\ 1 & 82 & 107 \\ 1 & 64 & 115 \\ 1 & 50 & 130 \\ 1 & 62 & 128 \\ 1 & 70 & 116 \\ 1 & 90 & 92 \\ 1 & 94 & 94 \\ 1 & 100 & 110 \\ 1 & 104 & 115 \\ 1 & 82 & 117 \end{bmatrix}$$

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 85 & 78 & 82 & 64 & 50 & 62 \\ 80 & 93 & 107 & 115 & 130 & 128 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 70 & 90 & 94 & 100 & 104 & 82 \\ 116 & 92 & 94 & 110 & 115 & 117 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 12 & 961 & 1297 \\ & 79849 & 102414 \\ & & 142777 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{64516342}$$

$$\begin{bmatrix} 911973280 & -4377739 & -5144299 \\ & 31115 & 17449 \\ & & 34667 \end{bmatrix}$$

$$Y = \begin{bmatrix} 3,65 \\ 4,22 \\ 4,29 \\ 5,43 \\ 6,62 \\ 5,71 \\ 5,09 \\ 3,99 \\ 4,08 \\ 4,38 \\ 4,28 \\ 4,42 \end{bmatrix} \qquad X^t Y = \begin{bmatrix} 56,16 \\ 4368,21 \\ 6191,6 \end{bmatrix}$$

La solución del modelo es:

$$\beta = (X^t X)^{-1} (X^t Y) = \begin{bmatrix} 3,75245 \\ -0,02945 \\ 0,0304 \end{bmatrix} \begin{matrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{matrix}$$

Entonces el modelo se plantea de la forma:

$$\hat{Y} = 3,75245 - 0,02945X_1 + 0,0304X_2$$

Estos coeficientes significan:

$\hat{\beta}_0 =$ 3,75, es el costo promedio de manufactura por unidad de producto, sin tener en cuenta otros indicadores.

$\hat{\beta}_1 =$ −0,02945, es la disminución del costo al cambiar en un punto el nivel de producción.

$\hat{\beta}_2 =$ 0,0304, es el aumento del costo al cambiar en un punto el índice de costos.

Una vez obtenido el modelo lo validamos por medio del coeficiente de determinación.

Calculamos las sumas de cuadrados y obtenemos los siguientes datos:

$$SCT = 270,9622 - 12\left(\frac{56,16}{12}\right)^2 = 8,1334$$

$$SCE_X = 270,3184 - 12\left(\frac{56,16}{12}\right)^2 = 7,4896$$

$$SCR = 270,9622 - 270,3184 = 0,6438$$

Por lo tanto:

$$R^2 = \frac{7,4896}{8,1334} = 0,9208$$

Significa que el modelo explica el 92,08%, lo cual es muy representantivo.

Para este conjunto de datos se usaron dos modelos adicionales que son:

$$Y = \beta_0 X_1^{\beta_1} X_2^{\beta_2}$$

con valores $Y = 2,0354\, X_1^{0,4685} X_2^{0,6118}$

y con $R^2 = 83,43\%$

y el modelo $Y = \beta_0 \beta_1^{X_1} \beta_2^{X_2}$

con valores $Y = 3,5082(0,9946)^{X_1} (1,0066)^{X_2}$

con un $R^2 = 93,41\%$

locual indica que este modelo es más representativo que el encontrado en forma lineal, pero haremos las aplicaciones con nuestro primer modelo.

Deseamos estimar el costo de manufactura para un nivel de producción de $X_1 = 70$ y un índice de costos $X_2 = 116$, con una confiabilidad del 95%.

Para esto calculamos el error estandar de la estimación.

$$\partial Y/X = \sqrt{\frac{0,6438}{12-3}} = 0,2675$$

En base a esto obtenemos los siguientes estimados

$$5,21735 \pm 1,96(0,2675) =$$
$$4,69305 \leqslant Y \leqslant 5,74165 \$$$

Esto nos indica que en el mes 7 los costos observados fueron de 5,09, los cuales están dentro del intervalo.

Ahora vamos a probar la incidencia o no que tienen estos índices en conjunto con relación al costo de manufactura.

Planteamos nuestras hipótesis de trabajo:

$H_0 : \beta_1 = \beta_2 = 0$

$H_1 : \beta_1 \neq \beta_2 \neq 0$

Diseñamos la prueba estadística F:

F de V	GL	SC	CM	
Debido a la regresión.	3 – 1 = 2	7,4896	$\frac{7,4896}{2} = 3,7448$	$F = \frac{3,7448}{0,0715} = 52,37$
Debido a residuos.	12 – 3 = 9	0,6438	$\frac{0,6438}{9} = 0,0715$	
TOTAL	12 – 1 = 11	8,1334		

La región crítica está dada por:

$P_r\ (F_C > F_T) = 0,05$ donde

$F_T\ (2;9;0,975) = 5,71$

Como 52,37 es mayor que 5,71 rechazamos H_0 en forma altamente significativa, es decir, que estos dos índices de producción y de costo tienen mucha importancia (Incidencia) en el costo de manufactura de una unidad de producción.

En forma separada vamos a comprobar estos índices si afectan el costo de manufactura. Empleamos la matriz de varianzas covarianzas. Entonces:

$$(\text{Var} - \text{Cov})\beta = \begin{array}{c} \\ Y \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y & X_1 & X_2 \end{array} \\ \begin{bmatrix} 1,0107 & -0,00485 & -0,0057 \\ & 0,000035 & 0,000019 \\ & & 0,000038 \end{bmatrix} \end{array}$$

Verifiquemos si el nivel de producción tiene incidencia en el costo o no. Planteamos las hipótesis:

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

La prueba estadística es:

$$Z = \frac{-0,02945 - 0}{\sqrt{0,000035}} = -4,978$$

Como –4,978 es menor que –1,96 rechazamos H_0 en forma significativa, es decir, que el nivel de producción afecta de una manera representativa el costo de manufactura.

Comprobamos para el índice de costo

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

Con prueba estadística:

$$Z = \frac{0,0304}{\sqrt{0,000038}} = 4,93$$

Como 4,93 es mayor que 1,96 rechazamos H_0 en forma significativa, es decir, que el nivel de costos afecta de una manera representativa el costo de manufactura.

También podemos analizar los cambios que ocurren en el costo de manufactura al variar en 1 punto los índices de producción y de costos. Para el primero sería:

$-0,02945 \pm 1,96\sqrt{0,000035} =$

$-0,041 \leq \beta_1 \leq -0,01785$

y el otro sería

$0,0304 \pm 1,96\sqrt{0,000038} = 0$

$0,0183 \leq \beta_2 \leq 0,0425$

Estos intervalos indican cuánto pueden aumentar o disminuir por punto de índice el costo de manufactura.

Finalmente se ve el grado de correlación de las variables.

En conjunto están muy asociadas estas tres variables, o sea, el 0,9595. En forma de parejas obtenemos la matriz de correlación, R.

$$R = \begin{array}{c} \\ Y \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y & X_1 & X_2 \end{array} \\ \begin{bmatrix} 1 & -,8154 & -0,9198 \\ & 1 & 0,521 \\ & & 1 \end{bmatrix} \end{array}$$

Como vemos

$\sqrt[R]{(Y_1 X_1)} = 81,54\%$ y

$\sqrt[R]{(YX_2)} = -91,98\%$

indican que hay un alto grado de disociación entre costo de manufactura y los índices de producción y de costos.

con este escrito se quiere demostrar las aplicaciones de los métodos estadísticos en el campo de la administración.

En otros escritos posteriores se presentarán otros usos de los métodos estadísticos.