Capitulo X

ANALISIS DE REGRESION Y CORRELACION

El análisis de regresión consiste en emplear métodos que permitan determinar la mejor relación funcional entre dos o más variables concomitantes (o relacionadas), y el análisis de correlación, el grado de asociación de las mismas. Es decir; no sólo se busca una función matemática que exprese de que manera se realcionan, sino también con que presición se puede predecir el valor de una de ellas si se conoce los valores de las variables asociadas.

ANALISIS DE REGRESION

Una relacion funcional matemáticamente hablando, está dada por:

$$Y = f(x_1,...,x_n; \theta_1,...,\theta_m) \qquad (1)$$

donde:

Y : Variable respuesta (o dependiente)
x_i : La i-ésima variable independiente (i=1,...,n)
θ_j : El j-ésimo parámetro en la función (j=1,...m)
f : La función

Para elegir una relación funcional particular como la representativa de la población bajo investigación, usualmente se procede:

1) Una consideración analítica del fenómeno que nos ocupa, y
2) Un examen de diagramas de dispersión.

Una vez decidido el tipo de función matemática que mejor se ajusta (o representa nuestro concepto de la relación exacta que existe entre las variables) se presenta el problema de elegir un expresión particular de esta familia de funciones; es decir, se ha postulado una cierta función como término del verdadero estado en la población y ahora es necesario estimar los parámetros de esta función (ajuste de curvas).

Como los valores de los parámetros no se pueden determinar sin errores por que los valores observados de la variable dependiente no concuerdan con los valores esperados, entonces la ecuación (1) replanteada, estadísticamente, sería:

$$Y = f(x_1,...x_n;\theta_1,...,\theta_m) + \varepsilon \qquad (2)$$

donde ε respresenta el error cometido en el intento de observar la característica en estudio, en la cual muchos factores contribuyen al valor que asume ε.

REGRESION LINEAL SIMPLE

Cuando la relación funcional entre las variables dependiente (Y) e independiente (X) es una línea recta, se tiene una regresión lineal simple, dada por la ecuación

$$Y = \beta_o + \beta_1 X + \varepsilon$$

donde:

βo : El valor de la ordenada donde la línea de regresión

se intersecta al eje Y.
$\beta 1$: El coeficiente de regresión poblacional (pendiente de la línea recta)
ε : El error.

Supocisiones de la regresión lineal

1. Los valores de la variable independiente X son "fijos".
2. La variable X se mide sin error (se desprecia el error de medición en X)
3. Existen subpoblaciones de valores Y para cada X que están normalmente distribuidos.
4. Las variancias de las subpoblaciones de Y son todas iguales.
5. Todas las medias de las subpoblaciones de Y están sobre la misma recta.
6. Los valores de Y están nomalmente distribuidos y son estadísticamente independientes.

Las suposiciones del 3 al 6 equivalen a decir que los errores son aleatorios, que se distribuyen normalmente con media cero y variancia σ^2.

Estimación de parámetros

La función de regresión lineal simple es expresado como:

$$Y = ß_o + ß_1 X + \varepsilon \qquad (3)$$

la estimación de parámetros consiste en determinar los parámetros $ß_o$ y $ß_1$ a partir de los datos muestrales observados; es decir, deben hallarse valores como bo y b1 de la muestra, que represente a $ß_o$ y $ß_1$, respectivamente.

De la ecuación (3), para un x_i determinado, se tiene el correspondiente Y_i, y el valor del error ε_i sería $(Y_i - ß_o - ß_1 X_i)$

Empleando el método de los mínimos cuadrados, es decir minimizando la suma de cuadrados de los errores, se determinan los valores de b_o y b_1, así:

$$Q = \sum \varepsilon_i^2 = \sum \left(y_i - \beta_0 - \beta_1 x_i\right)^2$$

$$\frac{\partial Q}{\partial \beta_0} = 2\sum\left(y_i - \beta_0 - \beta_1 x_i\right)(-1) = 0$$

(4)

$$\frac{\partial Q}{\partial \beta_1} = 2\sum\left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) = 0$$

(5)

Al sistema formado por las ecuaciones (4) y (5) se les denomina ecuaciones normales.

Resolviendo las ecuaciones normales, se tiene:

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} = \frac{SPXY}{SCX}$$

donde:
b_0 : es el valor que representa (estimador) a ß$_0$
b_1 : es el valor que representa (estimador) a ß$_1$
SPXY : denota a la suma de productos de X con Y,
SCX : denota a la suma de cuadrados de X.

Luego, la ecuación de regresión es:

$$\hat{y} = b_0 + b_1 X$$

El coeficiente de regressión (b_1)

Está expresado en las mismas unidades de medida de la variable X. e indica el número de unidades que varía Y cuando se produce cambio en una unidad en X (pendiente de la recta de regresión).

Si b_1=0, se dice que no existe relación lineal entre las dos variables y que estas son independientes.

EJEMPLO: A continuación se desarrollara un ejemplo práctico que se irá explicando a través de los tópicos de regresión y correlación a tratarse.

Los datos de la siguiente tabla representan las alturas (X) y los pesos (Y) de varios hombres. Se escogieron las alturas de antemano y se observaron los pesos de un grupo de hombres al azar que tenian las alturas escogidas, resultando:

X(cm)	152	155	152	155	157	152	157	165	162	178	183	178
Y(kg)	50	61.5	54.5	57.5	63.5	59	61	72	66	72	84	82

Se asume que existe una relación funcional entre X e Y, obtener la ecuación de regresión.
Solución: En primer lugar se observa que Y=f(x), por tanto se asume que la variable altura (X) es independiente y la variable peso (Y) es la dependiente, luego se afirma que $Y = b_o + b_1X$. Para ello se efectúan los sgtes cálculos:

n = 12, $\sum X$ = 1946, $\bar{x}$ = 162.167, $\sum Y$ = 783, $\bar{y}$ = 65.25

SCX = $\sum X^2 - (\sum X)^2/12$ = 316986- (1946)²/12 = 1409.667

SPXY = $\sum XY - (\sum X)(\sum Y)/12$ = 128199.5-(1946x783)/12=1223

Luego, se calcula b_o y b_1:

b_o = 65.25-(0.8676)(162.167) = -75.446
b_1 = 1223/1409.667 = 0.8676

Por tanto, la ecuación buscada es:

$$\hat{y} = -75.446 + 0.8676X$$

El valor de b_1 = 0.8676 indica que por cada centímetro de aumento en la altura de los hombres, habrá un incremento ,en promedio, de 0.8676 kg en el peso de los mismos.

Fuentes de variación en la regresión lineal

Los cálculos de regresión pueden ser vistos como un proceso de partición de la suma total de cuadrados; así, gráficamente se tiene:

Grafico FIG 1

Se observa que la desviación total para un Yi en particular es igual a la suma de las desviaciones explicada e inexplicada, simbolicamente.

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

Luego,

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

SCT = SCR + SCE

SCT: Suma de cuadrados del total
SCR: Suma de cuadrados de la regresion
SCE: Suma de cuadrados residual

Suma de Cuadrados del Total (SCT), mide la dispersión (variación total) en los valores observados de Y. Este término se utiliza para el cálculo de la variancia de la muestra.

Suma de Cuadrados explicada (Suma de Cuadrados debido a la Regresión, SCR) mide la variabilidad total en los valores observados de Y en consideración a la relación lineal entre X e Y. Suma de Cuadrados inexplicada (Suma de Cuadrados del Error, SCE) mide la dispersión de los valores Y observados respecto a la recta de regresión Y (es la cantidad que se minimiza cuando se obtiene la recta de regresión).

Análisis de Variancia para la regresión lineal simple

Cuando cada partición se asocia a una porción correspondiente del total de grados de libertad, la técnica es conocida cono ANALISIS DE VARIANCIA (ANVA), que generalmente se presenta en un cuadro de la siguiente forma:

Cuadro ANVA				
F. de V.	G.L.	SC.	CM.	Fc
Regresión	1	b_1SPXY	b_1SPXY	CMR/CME
Error	n-2	$\Sigma(Y_i-Y_i)^2$	SCE/(n-2)	
Total	n-1	SCT		

La prueba estadística es F y evalua las hipótesis:

H_p : No existe una regresión lineal entre X e Y
H_a : Existe regresion lineal de Y en función de X

Para el ejemplo planteado efectuar el ANVA.

Cálculo de las sumas de cuadrados:

SCTotal = SCT = $\Sigma Y^2-(\Sigma Y)^2/n$ = 52297 - $(783)^2/12$ = 1206.25

SCRegresión = SCR = b_1SPXY = (0.8676)(1223) = 1061.0748

SCError = SCE = SCT - SCR = 1206.25 -1061.0748 = 145.1752

Càlculo de la variancia residual o del error:

S^2 = (SCT - SCR)/(N-2) = 145.1752/10 = 14.5175.

esto nos indica que la variabilidad de los pesos de los hombres es 14.5175 kg^2 sin tener en cuenta el efecto de las alturas (X) sobre los pesos (Y); es decir, mide la variabiliad de Y una vez descontado el efecto de X, siendo menor que la variancia de Y.

Cuadro ANVA				
F. de V.	G.L.	SC.	CM.	Fc
Regresión	1	1061.0748	1061.0748	73.089 **
Error	10	145.1752	14.5175	
Total	11	1206.2500		

El valor $F_{0.01}(1,10)$ = 10. ; como Fc > $F_{\square}$, la regresión es altamente siginifactiva.

INTERVALOS DE CONFIANZA

En muchos casos es de interés conocer entre que valores se encuentra el coeficiente de regresión de la población $ß_1$ para un cierto grado de confianza fijada, este procedimiento permite hallar los valores llamados límites de confianza, así:

$b_1 - t_0 S_{b1} \leq ß_1 \leq b_1 + t_o S_{b1}$

donde: t_0 es el valor "t" tabular al nivel de significación $\square$ y n-2 grados de libertad ($t_0 = t_{\square,n-2}$).

$S^2_{b1} = S^2_E/SCX$ = CME/SCX (obtenido del cuadro ANVA) es la variancia estimada del coeficiente de regresión.

También es de interés determinar el intervalo de confianza de $\mu_{y/x}$, para un valor asumido de X_i, que se calcula con la expresión:

$$\hat{y} - t_0 S_y \leq \mu_{y/x} \leq \hat{y} + t_0 S_y$$

donde $t_o=t_{\square,n-2\ gl.}$ y S^2_Y=CME(1/n + $(X_i- X)^2$/SCX).

Nota: Puede observarse que la variancia de la línea de regresión $S^2_{\hat{y}}$ irá incrementándose conforme X_i se aleja de X.

EJEMPLO.

Calcular los límites de confianza para el coeficiente de regresión $ß_1$ y de $\mu_{y/x}$ para X= 185 , al 95% de confianza.

a) Para $ß_1$: cálculos previos:

S^2_{b1} = CME/SCX = 14.5175/1409.667 =0.010298.

así, S_{b1}=0.1015 y t_o=2.228. Luego,

$0.8676-(2.228)(0.101479) \leq ß_1 \leq 0.8676+(2.228)(0.101479)$

$0.6415 \leq ß_1 \leq 1.0937$

b) Para $\mu_{y/x}$, donde X=185:

$\hat{y} = -75.446 + 0.8676(185) = 85.06,$ y

$S^2_{\hat{y}}$= 14.5175(1/12 -(185 -162.167)²/1409.667) = 6.57889

así, $S_{\hat{y}}$= 2.5649. Luego, aplicando la fórmula dada

anteriormente se tiene:

$79.3454 \leq \mu_{y/x} \leq 90.7746$.

Este intervalo de confianza nos indica que si las tallas fuesen de 185 cm, existe el 95% de probabilidad que los valores del intervalo encierren el verdadero promedio.

PRUEBAS DE HIPOTESIS

Se plantea los siguientes casos:

a) Cuando ß=0 (Prueba de Independencia); es decir, si la variable Y es independiente de la variable X. Esto equivale a plantear la hipótesis Hp: $ß_1$=0, y mediante la prueba F comparar la F calculada (Fc) con la F tabular (Fo), donde Fc=CMR/CME y Fo=F□(1,n-2 gl). Si Fc>Fo, se rechaza la hipóteis planteada y se concluye que Y depende de X.

b) Cuando $ß_1$ tiene un valor específico, digamos $ß_{10}$; es decir; Hp: $ß_1$=$ß_{10}$. En este caso se usa el estadístico t para probar esta hipòtesis, se calcula el valor de t:

$$t_c = \frac{b1 - \beta_{10}}{S_{b1}} = \frac{b1 - \beta_{10}}{\sqrt{\frac{CMe}{SCx}}}$$

Si $t_c > t_0$ se rechaza la hipótesis planteada, donde t_0 es el valor de la tabla al nivel α y n-2 gl.

EJEMPLO:

Probar si el peso de los hombres es independiente de sus alturas, También probar si por cada cm. de altura en cada hombre el peso aumenta en 1.2 kg.

caso (a): ¿Son X y Y independientes?. Las hipótesis son:

Hp: $ß_1 = 0$
Ha: $ß_1 \neq 0$

Aplicando las fórmulas dadas se tiene:

Fc= 1061.0748/14.5175 = 73.089.

Las F tabulares a 0.05 y 0.01 son:

Fo = 4.96 y Fo = 10.04, respectivamente.

Luego, comparando para ambos valores se tiene que Fc > Fo. Por lo tanto se concluye en que la influencia de X sobre Y es directa y no se debe al azar (es decir, Y depende de X).

caso (b): Se tiene Hp: $ß_1 = 1.2$
Ha: $ß_1 \neq 1.2$

t_c = (0.8676-1.2)/0.101479 = -3.27.

Como t tabular es t_o= -2.228 (□=0.05 y gl. = 10) el valor de t_c cae en la zona de rechazo de la Hp ($t_c < t_o$ =-2.228), por lo tanto se concluye que por cada cm adicional el la altura no hay aumento de 1.2 kg.

PREDICCION

Hallada la ecuación de regresión puede darse uso en los siguientes casos:

a) Predecir el valor probable de Y dado un valor particular de X.
b) Estimar el valor desconocido de X asociado a un valor observado de Y.
c) Construir un intervalo de predicción para un valor predicho de Y.

Para los casos (a) y (b), se identifican los valores de las variables y se rremplazan en la ecuación $Y=b_o+b_1X$.

Así por ejemplo:

Suponga que esta interesado en conocer $\hat{y}_p$ (estimado de la media poblacional Y para un valor predicho Xp no considerado en la muestra). Este valor se obtiene de $Y=b_o+b_1X$; así, si Xp=160 entonces $\hat{y}_p = 63.37 Kg.$, este valor debe interpretarse como el estimado del peso promedio si se tuviesen varias alturas de 160 cm.

Para el caso (c), el intervalo de confianza para la predicción es

$$\hat{y}_p - t_0 S_{yp} \leq \mu_{y/x} \leq \hat{y}_p + t_0 S_{yp}$$

donde t_0 es t tabular a nivel α y grados de libertad de n-2.

$$Sy_p = \sqrt{CMe\left(1 + \frac{1}{n} + \frac{(xp - \bar{x})^2}{SCx}\right)}$$

Luego, el intervalo de confianza para la predicción hallada al 95 % de confianza es:

$$54.5207 \leq \mu_{y/x} \leq 72.2193$$

Significa, si se tuviesen muchos hombres de 160 cm, existe el 95% de probabilidad de que el intervalo de confianza [54.5207 , 72.2193], encierre el verdadero promedio de los pesos.

ANALISIS DE CORRELACION

El análisis de correlación consiste en emplear métodos que permitan medir el grado o intensidad de asociación entre dos o más variables. El concepto de correlación está estrechamente vinculado al concepto de regresión, pues, para que una ecuación de regresión sea razonable los puntos muestrales deben estar ceñidos a la ecuación de regresión; además el coeficiente de correlación debe ser:

- grande cuando el grado de asociación es alto, y pequeño cuando es bajo
- independiente de las unidades en que se miden las variables.

CORRELACION LINEAL SIMPLE.

El coeficiente de correlación (r) es un número que indica el grado o intensidad de asociación entre las variables X e Y. Su valor varía entre -1 y +1; esto es:

$-1 \leq r \leq 1.$

Si r=-1, la asociación es perfecta pero inversa; es decir, a valores altos de una variable le corresponde valores bajos a la otra variable, y viceversa.

Si r=+1, también la asociación es perfecta pero directa.

Si r=0, no existe asociación entre las dos variables.

Luego puede verse que a medida que r se aproxime a -1 ó +1 la asociación es mayor, y cuando se aproxima a cero la asociación disminuye o desaparece.

El coeficiente de correlación está dada por:

$$r = \sqrt{\frac{SPxy}{(SCx)(SCy)}}$$

Así, para el ejemplo planteado:

$$r = \sqrt{\frac{1223}{(1409.667)(1206.25)}} = 0.9381$$

Este valor nos indica que hay un alto grado de asociación entre las variables altura y peso, y la relación es directa (signo positivo de r).

COEFICIENTE DE DETERMINACION

De la descomposición de la suma de cuadrados total, se obtuvo:

SCT = SCR + SCE

dividiendo ambos miembros por la SCT, se tiene:

1 = SCR/SCT + SCE/SCT

de este resultado, se define el COEFICIENTE DE DETERMINACION de la muestra, denotada por r^2, como:

r^2 = 1 - SCE/SCT = SCR/SCT
r^2 = SC explicada/SC total
r^2 = error explicado/error total

Como SCR $\leq$ SCT, se deduce que $0 \leq r^2 \leq 1$.

Interpretación de r^2:

Puede interpretarse desde 3 aspectos:

a) Como una medida de mejora debido a la línea de regresión. Aquí, r^2 proporciona la reducción relativa de la SCT (error total).

Si r^2= 0 decimos que no hay reducción en la SCT; es decir no hay mejora debido al ajuste de la línea de regresión, lo que significa que:

$$\text{Error\_Explicado} = \sum (\hat{y}_i - \bar{y})^2 = 0$$

Gráficamente, se observa que la línea de regresión es horizontal y coincidente con Y.

Si r^2=1, decimos que ha habido una reducción del 100% en el error total, o sea:

$$\text{Error\_total} = \sum (y_i - \hat{y}_i)^2 = 0$$

Gráficamente, todos los puntos del diagrama de dispersión caen sobre la línea de regresión no horizontal.

b) Como medida de grado de ajuste.

Si r^2=1, los puntos Y_i caen todos sobre la línea de regresión.

Si r^2=0, los puntos son esparcidos y la línea de regresión resulta horizontal.

En conclusión, cuando mayor es el grado de ajuste de la línea de regresión a los puntos, el valor de r^2 se acerca a 1.

c) Como el grado de linealidad de dispersión de los puntos. Si r^2 se aproxima al volor uno, la dispersión de puntos se parece a una línea recta.

Si r^2 se acerca al valor cero, la dispersión no se parece a una línea recta.

EJEMPLO. Del caso planteado. Hallar e interpretar r^2.

Según los calculos,

$$r = \frac{1223}{\sqrt{(1409.667)(1206.25)}} = 0.9381$$

entonces $r^2=0.88004$; $(1-r^2)=0.12996$

Indica que el 88.004% de los cambios en los pesos se asocia a los cambios en las alturas (tallas), resultando, 12.996% de variabilidad que no es explicada por la regresión.