

Razonamiento Aproximado. Ingeniería Informática. 2do Año.

Tema IV: Correlación y Regresión.

L/T. Walpole (pág.389-506); L/T. Miller (pág.301-353)

Correlación y Regresión Lineal Simple.

Hasta ahora hemos tratado de establecer estimaciones o probar hipótesis para parámetros de una variables o para las diferencias entre muestras que resultan mediciones de una misma variable aleatoria. En estos momentos comenzaremos a analizar que pasaría si nuestro interés fuera establecer relaciones entre variable o investigar si existen.

Debe quedar claro que si existen relaciones entre variables la forma de representarla sería a través de funciones. Comenzaremos analizando los aspectos relacionados con la correlación, y aunque a continuación indicaremos varias medidas de asociación, trabajaremos básicamente con el *coeficiente de correlación lineal Simple de Pearson*.

Cuando las medidas descriptivas se emplean para estudiar dos o más variables, de modo conjunto, se denominan **medidas de asociación**.

Las medidas de asociación, según sus usos, se clasifican en **medidas de correlación** y **medidas de regresión**. La **correlación** es la técnica estadística que estudia el problema de medir la **intensidad o el grado de relación** que existe entre las variables que se investigan. Si la correlación se mide entre dos variables, se dice que es **simple** y cuando es entre tres o más, se llama **correlación múltiple**.

Para medir el grado de correlación entre las variables, se utilizan los **coeficientes de correlación**. Entre estos tenemos:

- 1.- El coeficiente de correlación phi.
- 2.- El coeficiente de correlación C de contingencia.
- 3.- El coeficiente de correlación T de Chuprov.
- 4.- El coeficiente de correlación r_c de contingencia.
- 5.- El coeficiente de correlación K de Cramer.
- 6.- El coeficiente de correlación punto biserial.
- 7.- El coeficiente de correlación de rangos.
- 8.- El coeficiente de correlación lineal simple de Pearson.
- 9.- El coeficiente de correlación concordancia de Kendall W

La **regresión**, por su parte, es la técnica estadística que estudia el problema de **encontrar la mejor función matemática** que describe el comportamiento conjunto de las variables que se investigan. Si la regresión se mide entre dos variables, se dice que es **simple** y cuando es entre tres o más, se denomina **regresión múltiple**.

A la función matemática que se utiliza en la regresión se le nombra **función o curva de regresión** y de acuerdo con el tipo de función utilizada, la regresión puede ser lineal (si la función lo es) o no lineal (si la función no lo es). Así, la función de regresión podrá ser una recta, una parábola u otra función cualquiera.

Consideraremos una muestra aleatoria de volumen N en la que, a cada uno de los elementos de esa muestra, se le han medido dos variables X e Y. En cada caso, especificaremos las escalas en las que se miden las variables, y si es o no necesario, tabular dichas variables de modo conjunto.

1.- El coeficiente de correlación phi

Este coeficiente se emplea cuando se busca la correlación entre dos variables que estén medidas, ambas, en escala nominal dicotómica. Para su cálculo requiere que se construya, con los datos de la muestra, una tabla simple de doble entrada. La tabla tendrá que tener dos filas (h=2) y de dos columnas (k=2), es decir, será del tipo 2X2: a las frecuencias absolutas conjuntas le llamaremos, respectivamente, A, B, C y D; tal y como se muestra a continuación:

X	Y		TOTAL
	1	2	
1	A	B	A+B
2	C	D	C+D
TOTAL	A+C	B+D	N

A partir de las frecuencias obtenidas aquí, se define el coeficiente phi:

$$\phi = \frac{|AD - BC|}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

A continuación nos referimos a las propiedades de este coeficiente, que son útiles para la interpretación práctica de él.

Propiedades de phi:

1.- El menor valor que puede tomar phi es cero.

2.- El mayor valor que puede tomar phi es uno.

Comentario: cuando phi es cero, indica que entre las variables **no existe relación**; en cambio, cuando es uno, significa que entre esas variables existe una **relación perfecta**.

Estas observaciones también son válidas para los coeficientes que siguen.

Ejemplo 1: En una muestra aleatoria de cuarenta estudiantes se observó el interés por el estudio (X) y el sexo (Y) de cada uno de ellos. Mida la correlación entre estas variables, a partir de los dados en la siguiente tabla:

Tabla #1

Interés por el estudio según

El sexo de los alumnos de la

la escuela Oscar Ortiz

Curso: 1995 -1996

X	Y		TOTAL
	M	F	
SÍ	10	10	20
NO	8	12	20
TOTAL	18	22	40

Fuente: Muestra investigada

Leyenda: X: Interés por el

estudio. Y: Sexo

Solución: tanto la variable X como la Y están medidas en escala nominal dicotómica. Se ha construido una tabla de doble entrada en la que h=2 y k=2; además A=10, B=10, C=8, D=12 y las frecuencias marginales son A+B=20, C+D=20, A+C=18 y B+D=22.

$A \cdot D = 10 \times 12 = 120$, $B \cdot C = 10 \times 8 = 80$ y $A \cdot D - B \cdot C = 120 - 80 = 40$.

$(A+B)(C+D)A+C)(B+D) = 20 \times 20 \times 18 \times 22 = 158400$ y la raíz cuadrada de este resultado es: 397.994974843, por tanto, el coeficiente es:

$$\phi = \frac{40}{397.994974843} = .100503781526 \approx .10$$

. Existe una correlación muy baja entre las variables interés y sexo.

2.- El coeficiente de correlación C de contingencia

Una limitación que tiene phi es que solo se puede utilizar en tablas del tipo 2X2, por tal motivo no siempre es aplicable. En ocasiones, podremos utilizar el coeficiente C de contingencia.

Este coeficiente se emplea cuando se busca la correlación entre dos variables que estén medidas, ambas, en escala nominal, pero no necesariamente una y la otra tienen que ser dicotómicas. Para su cálculo requiere que se construya, con los datos de la muestra, una tabla simple de doble entrada de h filas (h=2) y de k columnas (k=2). Además, h y k pueden ser iguales o no.

A partir de la tabla hXk construida, el coeficiente C se define por:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}, \text{ donde } \chi^2 = N \left(\sum_{i=1}^h \sum_{j=1}^k \frac{N_{ij}^2}{N_{i.} N_{.j}} - 1 \right)$$

En resumen, para calcular el valor de C, seguiremos los siguientes pasos:

1.- Construir la tabla bivariada con h filas y k columnas y obtener las frecuencias observadas conjuntas N_{ij} ; así como, las marginales $N_{i.}$ y $N_{.j}$.

Propiedades de C:

- 1.- El menor valor que puede tomar C es cero.
- 2.- El mayor valor que puede tomar C es siempre menor que uno.
- 3.- En el caso en que h=k, el mayor valor que podrá tomar C es

$$\sqrt{\frac{k-1}{k}} \text{ o } \sqrt{\frac{h-1}{h}}, \text{ por ser } h=k.$$

Comentario: la razón de que C nunca alcance el valor de uno, es su gran limitación, ya que, en ningún caso se podrá llegar a saber si entre las dos variables existe una relación perfecta.

Ejemplo 2: Utilice los datos del ejemplo 1 para calcular el coeficiente C.

Solución: como ya se dijo, estamos ante dos variables medidas en escala nominal dicotómica; si estas fueran politómicas, también se podría aplicar el coeficiente C, no así el ϕ . La tabla está dada en el ejemplo 1: $N_{11}=10$, $N_{12}=10$, $N_{21}=8$ y $N_{22}=12$; además, $N_{1.}=20$, $N_{2.}=20$, $N_{.1}=18$ y $N_{.2}=22$.

Calculemos ahora el valor de χ^2 :

N_{ij}	N_{ij}^2	$N_{i.}$	$N_{.j}$	$N_{i.} N_{.j}$	$N_{ij}^2 / N_{i.} N_{.j}$
10	100	20	18	20X18=360	.277777777778
10	100	20	22	20X22=440	.227272727273
8	64	-	-	20X18=360	.177777777778
12	144	-	-	20X22=440	.327272727273
					1.0101010101

$$1.0101010101-1=.0101010101, \chi^2=40 \times .0101010101=.404040404$$

$$\chi^2+N=.404040404+40=40.4040404$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = \sqrt{\frac{.4040404}{40.4040404}} \approx \sqrt{.0099999} \approx .099$$

El mayor valor de C, teóricamente, aquí es $\sqrt{\frac{2-1}{2}} = .707106781187$, como C=.099 se aleja considerablemente de este valor, se puede decir que la relación entre estas variables es muy baja.

3.- El coeficiente de correlación T de Chuprov:

Atendiendo al hecho de que C nunca puede tomar el valor de uno, en ocasiones podremos utilizar el coeficiente T de Chuprov. Este coeficiente se emplea cuando se busca la correlación entre dos variables que estén medidas, ambas, en escala nominal, pero no necesariamente una y la otra tienen que ser dicotómicas. Para su cálculo requiere que se construya, con los datos de la muestra, una tabla simple de doble entrada de h filas (h=2) y de k columnas (k=2). Además, h y k pueden ser iguales o no.

A partir de la tabla hXk construida, el coeficiente T se define por:

$$T = \sqrt{\frac{\chi^2}{N\sqrt{(h-1)(k-1)}}}, \text{ donde } \chi^2 \text{ fue dado antes.}$$

Propiedades de T:

- 1.- El menor valor que puede tomar T es cero.
- 2.- El mayor valor que puede tomar T siempre será menor o igual que uno.
- 3.- Solo en el caso en que h=k, el coeficiente T, podrá tomar el valor de uno.

Comentario: la razón de que T solo pueda alcanzar el valor de uno cuando h=k, es su gran limitación, ya que, su utilización se restringe a casos particularidades. Cuando h≠k no se puede llegar a saber si entre las dos variables existe una relación perfecta, ni siquiera se podrá conocer si esa relación es o no "alta", pues no se no existe un "extremo" superior para este coeficiente en estos casos.

Ejemplo 3: Utilice los datos del ejemplo 1 para calcular el coeficiente T.

Solución: como ya se dijo, estamos ante dos variables medidas en escala nominal dicotómica; si estas fueran politómicas, también se podría aplicar el coeficiente T, no así el ϕ . La tabla está dada en el ejemplo 1 y el valor de $\chi^2=40.404040404$ fue obtenido en el ejemplo 2.

$$(h-1)(k-1)=(2-1)(2-1)=1 \text{ y } \sqrt{1}=1; N\sqrt{(h-1)(k-1)}=40(1)=40;$$
$$\frac{40.404040404}{40} = .010101010101; \sqrt{.010101010101} = .0100503781526 \approx .10$$

Como en este caso h=k, teóricamente T hubiese podido ser uno, por tanto, la relación entre X e Y es muy baja, ya que el valor de este coeficiente obtenido antes, está muy cerca de cero.

4.- El coeficiente de correlación r_c de contingencia

Un nuevo coeficiente que, en algunos casos se puede utilizar como alternativa de los anteriores, es el coeficiente r_c de contingencia. Este coeficiente se emplea cuando se busca la correlación entre dos variables que estén medidas, ambas, en escala nominal, pero no necesariamente una y otra variables tienen que ser dicotómicas. Para su cálculo requiere que se construya, con los datos de la muestra, una tabla simple de doble entrada de h filas (h=2) y de k columnas (k=2), pero necesariamente h y k tienen que ser iguales.

A partir de la tabla hXk construida (h=k), el coeficiente r_c se define por:

$$r_c = \sqrt{\frac{\chi^2}{N(k-1)}}, \text{ donde } \chi^2 \text{ fue dado antes.}$$

Propiedades de r_c :

- 1.- El menor valor que puede tomar r_c es cero.
- 2.- El mayor valor que puede tomar r_c es uno.

Comentario: cuando r_c es cero, indica que entre las variables **no existe relación**; en cambio, cuando es uno, significa que entre esas variables existe una **relación perfecta**. La gran limitación de este coeficiente es que h tiene que ser igual a k.

Ejemplo 4: Utilice los datos del ejemplo 1 para calcular el coeficiente r_c .

Solución: como ya se dijo, estamos ante dos variables medidas en escala nominal dicotómica; si estas fueran politómicas, también se podría aplicar el coeficiente, pero solo si $h=k$. La tabla está dada en el ejemplo 1 y el valor de $\chi^2=.40404040404$ fue obtenido en el ejemplo 2.

$$N(k-1)=40(2-1)=40(1)=40; \quad r_c = \sqrt{\frac{\chi^2}{N(k-1)}} = \sqrt{\frac{.40404040404}{40}} = .100503781526.$$

Como en este caso $h=k$, se pudo utilizar este coeficiente, por tanto, la relación entre X e Y es muy baja, ya que el valor de r_c obtenido antes, está muy cerca de cero.

5.- El coeficiente de correlación K de Cramer

Un coeficiente que elimina las limitaciones de todos los anteriores es el K de Cramer. Este coeficiente se emplea cuando se busca la correlación entre dos variables que estén medidas, ambas, en escala nominal, pero no necesariamente una y la otra tendrán que ser dicotómicas. Para su cálculo requiere que se construya, con los datos de la muestra, una tabla simple de doble entrada de h filas ($h=2$) y de k columnas ($k=2$). Además, h y k pueden ser iguales o no.

A partir de la tabla $h \times k$ construida, el coeficiente K se define por:

$$K = \sqrt{\frac{\chi^2}{N \min(k-1, h-1)}}, \text{ donde } \chi^2 \text{ fue dado antes.}$$

Propiedades de K:

1.- El menor valor que puede tomar K es cero.

2.- El mayor valor que puede tomar K es uno.

Comentario: cuando K es cero, indica que entre las variables **no existe relación**; en cambio, cuando es uno, significa que entre esas variables existe una **relación perfecta**.

Ejemplo 5: Utilice los datos del ejemplo 1 para calcular el coeficiente K.

Solución: como ya se dijo, estamos ante dos variables medidas en escala nominal dicotómica; si estas fueran politómicas, también se podría aplicar el coeficiente, sin ninguna limitación. La tabla está dada en el ejemplo 1 y el valor de $\chi^2=.40404040404$ fue obtenido en el ejemplo 2.

En este caso $h=k=2$, por ello $h-1=2-1=1$ y $k-1=2-1=1$, el menor valor de estas dos restas es 1: $\min(h-1, k-1)=1$, por tanto, $N \min(h-1, k-1)=40(1)=40$.

$$K = \sqrt{\frac{.40404040404}{40}} = .100503781526: \text{ aquí podemos realizar una interpretación similar a la del ejemplo anterior.}$$

6.- El coeficiente de correlación biserial puntual:

El coeficiente de correlación biserial puntual, también llamado coeficiente de correlación punto biserial, se emplea cuando se busca la correlación entre **dos variables, una de ellas medida en escala métrica y la otra, en escala nominal dicotómica**. Para su cálculo no requiere que se hayan tabulado previamente las variables; aunque también, primero se pueden tabular estas de modo conjunto, y después calcular dicho coeficiente. La tabla que se confeccione puede ser simple o de agrupación y tendrá h filas ($h=2$) y $k=2$ columnas. Nosotros trataremos el caso en el que las variables no se hayan tabulado.

Denotaremos por r_{bp} el coeficiente de correlación punto biserial. Consideremos una muestra de volumen N, en la que se han medido, en escala métrica la variable X y en escala nominal dicotómica la variable Y. Sean $X_1, X_2, X_3, \dots, X_N$ y $Y_1, Y_2, Y_3, \dots, Y_N$, respectivamente, los valores de X y de Y (aquí las Y_i solo toman dos valores diferentes). Llamémosle P a la proporción de elementos de la muestra que corresponden a una de esas dos categorías; por tanto, en la otra categoría tendremos una proporción igual $1-P$.

Por otro lado, sea S la desviación estándar de la variable medida en escala métrica. Agrupemos los datos de la variable métrica en dos subgrupos según los valores de la variable Y , y calculemos, también para variable X , las medias aritméticas correspondientes a cada uno de los dos subgrupos formados.

Sean \bar{X}_P y \bar{X}_Q las medias de la variable X de los subgrupos que tienen proporciones P y $1-P$, respectivamente.

$$r_{bp} = \frac{(\bar{X}_P - \bar{X}_Q)\sqrt{P(1-P)}}{S}$$

Con esto definimos el coeficiente biserial puntual por:

Propiedades de r_{bp} :

- 1.- El menor valor que puede tomar r_{bp} es menos uno.
- 2.- El mayor valor que puede tomar r_{bp} es uno.

Comentario: cuando r_{bp} es cero, indica que entre las variables **no existe relación**; en cambio, cuando es uno o menos uno, significa que entre esas variables existe una **relación perfecta**. El signo negativo de un valor de r_{bp} indica que la relación entre las variables es inversa. Por otro lado, si el signo de r_{bp} es positivo indica que la relación entre las variables es directa.

Ejemplo 6: En una muestra aleatoria de seis estudiantes se observó el peso en kilogramos (X) y el sexo (Y) de cada uno de ellos. Mida la correlación entre estas variables, a partir de los siguientes datos:

Alumnos	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆
X	62	61	61	58	64	66
Y	M	F	F	F	M	M

Solución: X está medida en escala de razones y Y en nominal dicotómica. La desviación estándar de la variable X es $S=2.75680975042$. Al dividir los elementos de la muestra en dos subgrupos, según el sexo, tenemos que los alumnos A_1 , A_5 y A_6 integran un subgrupo (masculinos) y los alumnos A_2 , A_3 y A_4 integran el otro subgrupo (femenino). La proporción de alumnos masculinos en la muestra es $P=3/6=.5$, mientras que la de femenino es $1-P=1-.5=.5$.

$$\bar{X}_P = \frac{62 + 64 + 66}{3} = 64kg$$

La media del subgrupo de los alumnos es

$$\bar{X}_Q = \frac{61 + 61 + 58}{3} = 60kg$$

La media del subgrupo de las hembras es

$$P(1-P)=.5(1-.5)=.5 \times .5=.25 \text{ y } \sqrt{.25}=.5. \quad \bar{X}_P - \bar{X}_Q = 64 - 60 = 4, \text{ por tanto, } (\bar{X}_P - \bar{X}_Q)\sqrt{P(1-P)} = 4 \times .5 = 2.$$

De aquí se tiene que $r_{bp} = 2/2.75680975042 = .72547625011$. Entre el peso y el sexo de los alumnos de esta muestra existe una relación directa y alta.

7.- El coeficiente de correlación de rangos

El coeficiente de correlación de rangos, también llamado coeficiente de correlación de Spearman, se emplea cuando se busca la correlación entre **dos variables que estén medidas en escala ordinal o métrica**; es decir, las dos variables pueden estar medidas en escala métrica, las dos en escala ordinal o una en escala métrica y la otra en ordinal.

Denotaremos por r_s el coeficiente de correlación de Spearman y para una muestra de volumen N , en la que se han medido, en escala ordinal o métrica, las variables X e Y con los valores $X_1, X_2, X_3, \dots, X_N$ y $Y_1, Y_2, Y_3, \dots, Y_N$, respectivamente, se determinan las diferencias d_i entre cada valor de X y su correspondiente valor de Y : $d_i = X_i - Y_i$.

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

Y_i , con ello se define este coeficiente por:

Propiedades de r_s :

- 1.- El menor valor que puede tomar r_s es menos uno.
- 2.- El mayor valor que puede tomar r_s es uno.

Comentario: cuando r_s es cero, indica que entre las variables **no existe relación**; en cambio, cuando es uno o menos uno, significa que entre esas variables existe una **relación perfecta**. El signo negativo de un valor de r_s indica que la relación entre las variables es inversa: a valores altos de una variable corresponderán valores bajos de la otra. Por otro lado, si el signo de r_s es positivo indica que la relación entre las variables es directa: a valores altos de una variable corresponderán, también, valores altos de la otra.

Ejemplo 7: En una muestra aleatoria de siete estudiantes se observaron las calificaciones en Matemática (X) y en Física (Y), ambas en puntos, de cada uno de ellos. A partir de los siguientes datos obtenga el coeficiente de correlación de rangos:

Alumnos	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
X	98	97	90	88	84	86	83
Y	97	96	96	96	88	88	78

Solución: ambas variables X e Y están medidas en escala de intervalo (métrica), por tanto, es posible utilizar el coeficiente de correlación de Spearman. Los rangos asignados a los valores de cada una de estas variables por separado; así como, las diferencias entre ellos y su cuadrado se muestran a continuación:

		Rangos de		Dif.	
X	Y	X	Y	d_i	d_i^2
98	97	7.0	7.0	.0	.00
97	96	6.0	5.0	1.0	1.00
90	96	5.0	5.0	.0	.00
88	96	4.0	5.0	-1.0	1.00
84	88	2.0	2.5	-.5	.25
86	88	3.0	2.5	.5	.25
83	78	1.0	1.0	.0	.00
-	-	-	-	-	$\sum d_i^2 = 2.50$

$$6 \sum d_i^2 = 6(2.50) = 15, (N^2 - 1) = (7^2 - 1) = 49 - 1 = 48, N(N^2 - 1) = 7(48) = 336,$$

$$6 \sum d_i^2 / [N(N^2 - 1)] = 15 / 336 = .0446428571429,$$

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} = 1 - .0446428571429 = .955357142857$$

Entre las calificaciones de Matemática y Física de los alumnos de esta muestra existe una relación directa y alta.

8.- El coeficiente de correlación lineal simple de Pearson

El coeficiente de correlación lineal simple, también llamado coeficiente de correlación de Pearson, se emplea cuando se busca la correlación entre **dos variables que estén medidas, ambas, en escala métrica**. Para su cálculo no requiere que se hayan tabulado previamente las variables; aunque también, primero se pueden tabular estas de modo conjunto, y después calcular dicho coeficiente. La tabla que se confeccione puede ser simple o de agrupación y tendrá h filas (h=2) y k columnas (k=2). Nosotros trataremos el caso en el que las variables no se hayan tabulado.

Antes de calcular este coeficiente es conveniente representar las variable X e Y que se investigan, en un sistema de coordenadas cartesianas. Cada par de valores (X_i, Y_i) se ploteará como un punto aislado, por lo que se obtendrá una "nube de puntos" que se denomina **diagrama de dispersión**.

Este diagrama es útil porque ofrece **a priori** una información sobre el comportamiento conjunto de los datos de la muestra, específicamente, en él se visualiza a qué función matemática se ajustan los datos de esa muestra.

Denotaremos por r el coeficiente de correlación de Pearson y para una muestra de volumen N , en la que se han medido, en escala métrica, las variables X e Y con los valores $X_1, X_2, X_3, \dots, X_N$ y $Y_1, Y_2, Y_3, \dots, Y_N$, respectivamente, se tiene que:

$$r = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{[N \sum X_i^2 - (\sum X_i)^2][N \sum Y_i^2 - (\sum Y_i)^2]}}$$

Propiedades de r :

- 1.- El menor valor que puede tomar r es -1.
- 2.- El mayor valor que puede tomar r es uno.

Comentario: cuando r es cero, indica que entre las variables **no existe relación lineal**; en cambio, cuando es uno, independientemente del signo positivo o negativo que tenga este número, significa que entre esas variables existe una **relación lineal perfecta**. El signo negativo de un valor de r indica que la relación entre las variables es inversa: a valores altos de una variable corresponderán valores bajos de la otra. Por otro lado, si el signo de r es positivo indica que la relación entre las variables es directa: a valores altos de una variable corresponderán, también, valores altos de la otra.

9.- El coeficiente de concordancia de Kendall W (correlación entre más de dos variables)

Si en el ejemplo anterior, además de las asignaturas de Matemática y de Física se incluyera la de Química, entonces estaríamos ante una situación en la que se desea medir la correlación entre tres variables ($h=3$). Esta medición no se puede realizar con los coeficientes anteriores, aunque el procedimiento para el cálculo del coeficiente es similar al anterior: las variables deben estar medidas en escala métrica u ordinal.

Los valores de cada una de estas variables se transforman en rangos de modo independiente una de otras, a continuación se determinan la suma de los rangos que le corresponden a cada alumno (R_j) y el promedio de estas sumas (media). Seguidamente, se determina la suma del cuadrado de las diferencias de cada rango con respecto

$$W = \frac{12S}{h^2(N^3 - N)}$$

a la media de ellos, a lo cual llamaremos S ; con esto se tiene que:

Propiedades de W :

—Siempre será un valor entre 0 y 1: cero indica que no hay concordancia entre las variables, mientras que uno indica concordancia perfecta.

—La razón por la cual W no puede ser negativo está dada en que entre más de dos variables no pueden existir desacuerdos discrepantes totalmente: por ejemplo si X y Y están en desacuerdo, y a la vez, X está en desacuerdo con Z , necesariamente entre Y y Z hay concordancia.

Ejemplo: calcular el coeficiente W para los siguientes datos, de una muestra aleatoria de 7 alumnos, donde X : notas de Matemática,

Y : notas de Física y Z : notas de Química.

Alumnos	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
X	98	97	90	88	84	86	83
Y	97	96	96	96	88	88	78
Z	91	98	92	88	87	96	85

Solución: $N=7$ (tamaño de la muestra o total de alumnos).

$h=3$ (total de variables)

De aquí, se tiene que $h^2(N^3 - N) = 3^2(7^3 - 7) = 9(343 - 7) = 9(336) = 3024$ (que es el denominador de W (ver la fórmula).

Para calcular S , primero se transforman en rangos los datos de cada variable:

Alumnos	A ₁	A ₂	A ₃	A ₄	A ₅	A ₆	A ₇
X	7	6	5	4	2	3	1
Y	7	5	5	5	2.5	2.5	1
Z	4	7	5	3	2	6	1
R_j	18	18	15	12	6.5	11.5	3

(suma de los rangos)

$$\bar{R} = \frac{\sum R_j}{N} = \frac{18+18+15+12+6.5+11.5+3}{7} = 12$$

media de los rangos:

$$S = (18-12)^2 + (18-12)^2 + (15-12)^2 + (12-12)^2 + (6.5-12)^2 + (11.5-12)^2 + (3-12)^2$$

$$S = 6^2 + 6^2 + 3^2 + 0^2 + (-5.5)^2 + (-.5)^2 + (-9)^2$$

$$S = 36 + 36 + 9 + 0 + 30.25 + 0.25 + 81 = 183.5$$

$$W = \frac{12(183.5)}{3024} = 0.728174603175 \quad (\text{Interprete el resultado})$$

Observación: este coeficiente se puede emplear cuando se midan tres o más variables. Además, el se suele emplear cuando en la investigación se utiliza el criterio de expertos (jueces) para medir el grado de concordancia entre estos. En tal caso, las variables serían los jueces y la muestra los ítems que se den a los expertos para que ellos ofrezcan su criterio.

Comenzaremos analizando algunos elementos referentes a el coeficiente de correlación lineal de Pearson ρ .

Propiedades:

- 1) $-1 < \rho < 1$.
- 2) Si $\rho = 0$ se dice que las variables están incorrelacionadas.
- 3) Si las variables son independientes $\rho = 0$.
- 4) $\rho = 0$ no implica independencia.
- 5) Si $\rho_{xy} = 1$, entonces $Y = ax + b$; $a > 0$ y viceversa, si $Y = ax + b$; $a > 0$ entonces $\rho_{xy} = 1$.
- 6) Si $\rho_{xy} = -1$, $Y = ax + b$ y $a < 0$ y viceversa.

Relaciones entre variables.

Análisis de Regresión: Estudio de la dependencia de una variable (variable dependiente) de una o más variables (variables independientes o explicativas) con el objetivo de predecir o estimar el valor medio poblacional de la primera en términos de valores conocidos de las segundas.

Así, se llama modelo de regresión a un modelo que expresa el **valor esperado** de una variable aleatoria Y , llamada *variable dependiente* en función de una ó más *variables independientes* X_1, X_2, \dots, X_k .

$$E(Y / X_1, X_2, \dots, X_k) = f(X_1, X_2, \dots, X_k)$$

Se supone que las variables independientes X_1, X_2, \dots, X_k son *variables controladas* (no aleatorias, con valores fijados), mientras que la variable dependiente Y es una variable aleatoria cuyos valores *se observan* para los valores de las variables independientes *fijados*, aunque realmente, sobre todo en fenómenos económicos, tal supuesto no se cumple, teniendo las variables independientes carácter aleatorio también.

(regresión no implica causación)

Clasificación de modelos

Según el número de variables independientes se clasifican los modelos en modelos de regresión *simple*, *doble* ó *múltiple*.

Según la expresión funcional f el modelo se denomina *lineal* ó *no lineal*.

Esta linealidad se refiere a linealidad en los parámetros, es decir, la función f debe ser tal que :

$$f_{\alpha\beta_1+\lambda\beta_2}(X_1, X_2, \dots, X_k) = \alpha f_{\beta_1}(X_1, X_2, \dots, X_k) + \lambda f_{\beta_2}(X_1, X_2, \dots, X_k)$$

Modelo de Regresión Lineal Simple

Así, se llama Modelo de Regresión Lineal Simple al modelo

$$E(Y / X) = \beta_1 + \beta_2 X \Leftrightarrow Y = \beta_1 + \beta_2 X + \varepsilon \Leftrightarrow Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

Diagrama de dispersión

Una primera vía para comenzar a estudiar la relación entre las variables es la vía gráfica a través del llamado *diagrama de dispersión*.

Coefficiente de correlación lineal simple de Pearson

En todo lo anterior hemos supuesto que X es una variable “controlada”, es decir, cuyos valores son fijados, observándose entonces el valor correspondiente de Y.

Aún cuando este no sea el caso, es decir, siendo X e Y variables aleatorias, puede calcularse el *coeficiente de correlación lineal* r , que mide el grado y sentido de la relación lineal entre X e Y

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} \quad -1 \leq r \leq 1$$

Estimación de los parámetros

El método que se verá de estimación de los parámetros es el de *mínimos cuadrados*.

La idea fundamental consiste en tomar como valores estimados de los parámetros a los valores $b_1 = \hat{\beta}_1$ y $b_2 = \hat{\beta}_2$ tales que hagan mínima la suma de cuadrados de las desviaciones de los valores reales de los estimados (*residuos*):

$$e_i = Y_i - \hat{Y}_i ; \min \sum e_i^2 = \min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - b_1 - b_2 X_i)^2.$$

Cómo resultado de la aplicación del método al modelo de regresión lineal simple se obtienen las siguientes expresiones de cálculo para b_1 y b_2 .

$$b_2 = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} ; b_1 = \bar{Y} - b_2 \bar{X}$$

Ecuación de regresión estimada

Se llama ecuación de regresión estimada a:

$\hat{Y} = b_1 + b_2 X$. La interpretación de \hat{Y} sería la de un estimado del valor medio o esperado de Y para el valor dado de X.

Análisis de varianza en la regresión

En caso de no existir dicha relación o ser muy débil no tendría ninguna ventaja usar esta última.

Puede analizarse dicho problema a través de la descomposición de la variación total de las Y's en dos partes:

- una debida a las diferencias entre los valores de Y correspondientes a distintas X's y causada por la relación de dependencia entre ambas variables y además a los factores aleatorios.
- otra debida a la influencia de factores aleatorios. Es decir, aún para valores de Y correspondientes al mismo valor de X existen diferencias debidas a los factores aleatorios.

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

A estas sumas se les denomina *Suma de cuadrados total (SCT)*, *Suma de cuadrados del Error ó Residual (SCE)* y *Suma de cuadrados de la regresión (SCReg)* respectivamente y tienen las siguientes fórmulas de cálculo:

$$SCT = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \rightarrow n - 1 \text{ g.l.}$$

$$SCReg = b_2 \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] \rightarrow 1 \text{ g.l.}$$

$$SCRes = SCT - SCReg \rightarrow n - 2 \text{ g.l.}$$

F.Variación	S.C	g.l	C.M	F
Total	SCT	n-1	-	
Regresión	SCReg	1	$CM\ Reg = \frac{SC\ Reg}{1}$	$\frac{CM\ Reg}{CME}$
Error	SCE	n-2	$CME = \frac{SCE}{n-2}$	

Coefficiente de determinación

La parte de la variación total que es “explicada” por la regresión, o sea, por la influencia de X sobre Y es SCReg, la proporción que esta representa de la variación total es el llamado *coeficiente de determinación*

$$R^2 = \frac{SC\ Reg}{SCT} \text{ (se acostumbra a dar en porciento)}$$

Inferencias acerca de la regresión

¿Qué inferencias son de interés sobre la regresión?

- ❖ Las estimaciones por intervalo de los parámetros del modelo.
- ❖ Pruebas de hipótesis sobre estos. En particular es de interés el problema referente a la significación o no de la influencia que ejerce X sobre Y, la determinación de si es significativamente mejor la estimación del valor esperado de Y a través de la ecuación de regresión que por medio de \bar{Y} .

Prueba F y T.

$$H_0: \beta_2 = 0 \text{ en } E(y/x) = \beta_1 + \beta_2 x \quad (b_1 + b_2 x \text{ no es mejor que } \bar{y}).$$

$$H_1: \beta_2 \neq 0 \text{ en } E(y/x) = \beta_1 + \beta_2 x \quad (b_1 + b_2 x \text{ es mejor que } \bar{y}).$$

Puede verse que dados los supuestos asumidos el cociente $F = \frac{CM\ Reg}{CME}$ bajo la hipótesis $H_0: \beta_2 = 0$ tiene distribución **F de Fisher** con 1 g.l en el numerador y n-2 g.l en el denominador, por tanto:

$$\text{Región crítica o de rechazo: } F > F_{\alpha}(1; n-2) \Leftrightarrow |t| > t_{\alpha/2}(n-2) \text{ con } t = \sqrt{F}$$

Debe enfatizarse que el valor numérico del coeficiente que acompaña a la variable independiente no expresa directamente lo significativo o no de su influencia sobre Y. Es decir, un valor “pequeño” no implica poca influencia y viceversa. Ello depende de las unidades de medida y de la desviación típica del estimador.

Veamos cuáles son los supuestos teóricos:

- El modelo de regresión es lineal en sus parámetros.
- Los valores de X son no estocásticos.
- El valor medio o promedio de u_i es igual a cero. $E(u_i | X_i) = 0$.
- Homocedasticidad o igualdad de varianzas de u_i .

$$\text{var}(u_i | X_i) = E[u_i - E(u_i | X_i)]^2 = E(u_i^2 | X_i) = \sigma^2$$
- No existe autocorrelación entre las u. $\text{cov}(u_i, u_j | X_i, X_j) = 0$.
- Cero covarianza entre u_i y X_i . $\text{cov}(u_i, X_i) = 0$.
- El número de observaciones n debe ser mayor al número de parámetros a estimar.
- El modelo de regresión está correctamente especificado.
- No existe multicolinealidad.

Estimación por intervalos

Pueden construirse estimaciones por intervalos de los parámetros del modelo, así como de $E(Y/X)$ y Y/X con su interpretación usual de dar una medida del error probable de la estimación puntual.

Las expresiones de cálculo serían respectivamente:

$$(1) b_2 \mp t_{\frac{\alpha}{2}}(n-2) \cdot \sqrt{\frac{CME}{\sum X_i^2}} \quad (2) b_1 \mp t_{\frac{\alpha}{2}}(n-2) \cdot \sqrt{CME \frac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}}$$

$$(3) \hat{Y}_0 \mp t_{\frac{\alpha}{2}}(n-2) \sqrt{CME \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)} E(Y/X)$$

$$(4) \hat{Y}_0 \mp t_{\frac{\alpha}{2}}(n-2) \sqrt{CME \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)}$$

Observemos que:

- La longitud de cada uno de los intervalos es menor mientras menor sea CME, es decir, mientras “mejor” sea el ajuste de los datos al modelo.
- En (3) y (4) la estimación será también más precisa mientras el valor de la variable independiente para el que se estimará (X_0) esté más cercano al valor medio de los valores fijados de la variable independiente y lo será menos para valores más “extremos”. En particular no es correcto utilizar la ecuación de regresión estimada ni las estimaciones por intervalos correspondientes para valores de X fuera del rango fijado de valores experimentales

Transformación de modelos

Existen numerosos modelos de regresión que aunque inicialmente no poseen la forma $E(Y/X) = \beta_1 + \beta_2 f(X)$, pueden ser transformados a esta.

Por ejemplo:

$$\diamond Y = e^{\beta_1 + \beta_2 X + U}$$

puede ser transformado como : $\ln Y = \beta_1 + \beta_2 X + U \Leftrightarrow Y' = \beta_1 + \beta_2 X + U$.

$$\diamond Y = \frac{1}{\beta_1 + \beta_2 X + U} \Leftrightarrow \frac{1}{Y} = \beta_1 + \beta_2 X + U \Leftrightarrow Y' = \beta_1 + \beta_2 X + U$$

Bastaría calcular los valores de $Y' = \ln Y$ ó $Y' = \frac{1}{Y}$ y usar las mismas expresiones de cálculo para b_2 y b_1 , utilizando los valores de X e Y'

Modelo de Regresión Lineal Múltiple

Consideraremos ahora un problema con más de una variable independiente. Supongamos n valores observados de la variable dependiente Y para valores fijados de las variables independientes.

$$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + \dots + \beta_{k+1} X_k + U$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{n \times 1} \quad X = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix}_{n \times (k+1)} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}_{n \times 1} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{k+1} \end{pmatrix}_{(k+1) \times 1}$$

$$Y = X\beta + \varepsilon$$

Vector de estimadores mínimo-cuadráticos

$$\hat{\beta} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{k+1} \end{pmatrix}$$

$$\min \sum e_i^2 ; \sum e_i^2 = e'e$$

Al minimizar $e'e$ (hallar $\hat{\beta}$ que minimice $e'e$) se obtiene la ecuación matricial $(X'X)^{-1}\hat{\beta} = X'Y$ cuya solución es $b = (X'X)^{-1}X'Y$

Prueba F total

Análogamente al caso de una variable independiente, es de interés la realización de pruebas de hipótesis sobre los coeficientes que acompañan a las variables independientes, para determinar si la influencia de estas en la variable dependiente es significativa.

$$H_o : \beta_2 = \beta_3 = \dots = \beta_{k+1} = 0$$

Ahora bien, la alternativa NO es que todos los coeficientes son diferentes de 0, sino

$$H_1 : \text{Al menos un } \beta_j \neq 0; j = 2, \dots, k+1$$

Las expresiones matriciales correspondientes a las sumas de cuadrados y la tabla ANVA serían

$$SCT = \sum (Y_i - \bar{Y})^2 = Y'Y - n\bar{Y}^2 \quad SC \text{ Reg} = \hat{\beta}'X'Y - n\bar{Y}^2 \quad SCE = SCT - SC \text{ Reg}$$

Tabla ANVA

F.Variación	S.C	g.l	C.M	F
Total	SCT	n-1	-	
Regresión	SCReg	k	$CM \text{ Reg} = \frac{SC \text{ Reg}}{k}$	$\frac{CM \text{ Reg}}{CME}$
Error	SCE	n-k-1	$CME = \frac{SCE}{n-k-1}$	

Nótese que el caso de una variable independiente se obtiene de la tabla anterior como caso particular para $k = 1$. De modo similar al caso simple, el coeficiente de determinación expresa el grado de ajuste del modelo a los datos

$$R^2_{\text{múltiple}} \longrightarrow R^2_{X_1, X_2, \dots, X_k} = \frac{SC \text{ Reg}_{X_1, X_2, \dots, X_k}}{SCT}$$

No obstante, ante el rechazo de H_o solo se conoce que al menos una de las variables independientes tiene una influencia significativa en $E(Y)$, no su número ni su identificación.

Tabla ANVA en función de R^2 .

F.Variación	S.C	g.l	C.M	F
Total	SCT	n-1	-	
Regresión	$R^2 \cdot SCT$	k	$\frac{R^2 \cdot SCT}{k}$	$\frac{R^2(n-k-1)}{(1-R^2) \cdot k}$
Error	$(1-R^2) \cdot SCT$	n-k-1	$CME = \frac{(1-R^2) \cdot SCT}{n-k-1}$	

Es claro que no puede analizarse la influencia de una variable independiente sobre Y sin tener en cuenta la influencia de las otras, una parte de la variación de Y explicada por una variable puede ser también explicada por otra.

Dóctimas parciales.

Veamos el caso de 2 variables independientes para concretar estas ideas:

Supongamos que ya se ha determinado que la variable X_1 tiene una influencia significativa en Y y que por tanto hemos decidido incluirla en el modelo. La pregunta sería: ¿es necesario o útil incluir también a X_2 ? Para ello sería necesario medir la variación que explica X_2 que no ha sido explicada por X_1 .

$$SC \text{ Reg}_{X_2/X_1} = SC \text{ Reg}_{X_1 X_2} - SC \text{ Reg}_{X_1} = SCE_{X_1} - SCE_{X_1 X_2}$$

Si esta variación “adicional” que explica X_2 es significativa entonces se incluiría también a X_2 , de lo contrario sólo se incluiría a X_1

Así la dócima correspondiente sería:

$$H_o : \beta_2 = 0 / \beta_1 \neq 0 \quad \text{en} \quad E(Y / X_1, X_2) = \beta_1 + \beta_2 X_1 + \beta_3 X_2$$

$$H_1 : \beta_2 \neq 0 / \beta_1 \neq 0 \quad \text{en} \quad E(Y / X_1, X_2) = \beta_1 + \beta_2 X_1 + \beta_3 X_2$$

La región crítica será $F > F_\alpha(1; n - 3)$ con $F = \frac{CM Re g_{X2/X1}}{CME_{X1X2}} = \frac{SC Re g_{X2/X1}}{CME_{X1X2}}$

Puesto que la suma de cuadrados del numerador tiene 1 grado de libertad.

De manera análoga pueden plantearse y resolverse dócimas parciales suponiendo la inclusión previa en el modelo de varias variables.

Coeficientes de determinación y correlación parciales

Con un sentido análogo a la suma de cuadrados de la regresión parcial

($SC Re g_{X2/X1}$) se definen los coeficientes de determinación y correlación parcial:

$$R_{X2/X1}^2 = \frac{SC Re g_{X2/X1}}{SCE_{X1}} \quad ; \quad r_{X2/X1} = \sqrt{R_{X2/X1}^2}$$

Coeficiente de determinación ajustado.

$$R_{ajust.}^2 = 1 - \frac{SCE/g.l}{SCT/g.l}, \text{ que hace "más comparables" distintos valores de } R^2.$$

Métodos de selección de la mejor ecuación de regresión.

En el caso de un problema de regresión lineal con dos variables independientes X_1 y X_2 pueden emplearse varios algoritmos equivalentes:

a)

- Aplicar las pruebas parciales de X_2 dada X_1 y viceversa.

- Incluir las variables independientes correspondientes a las pruebas donde se obtenga el rechazo de H_0

b)

- Seleccionar la variable independiente más correlacionada con Y (R^2 , r , $SCReg$, SCE)

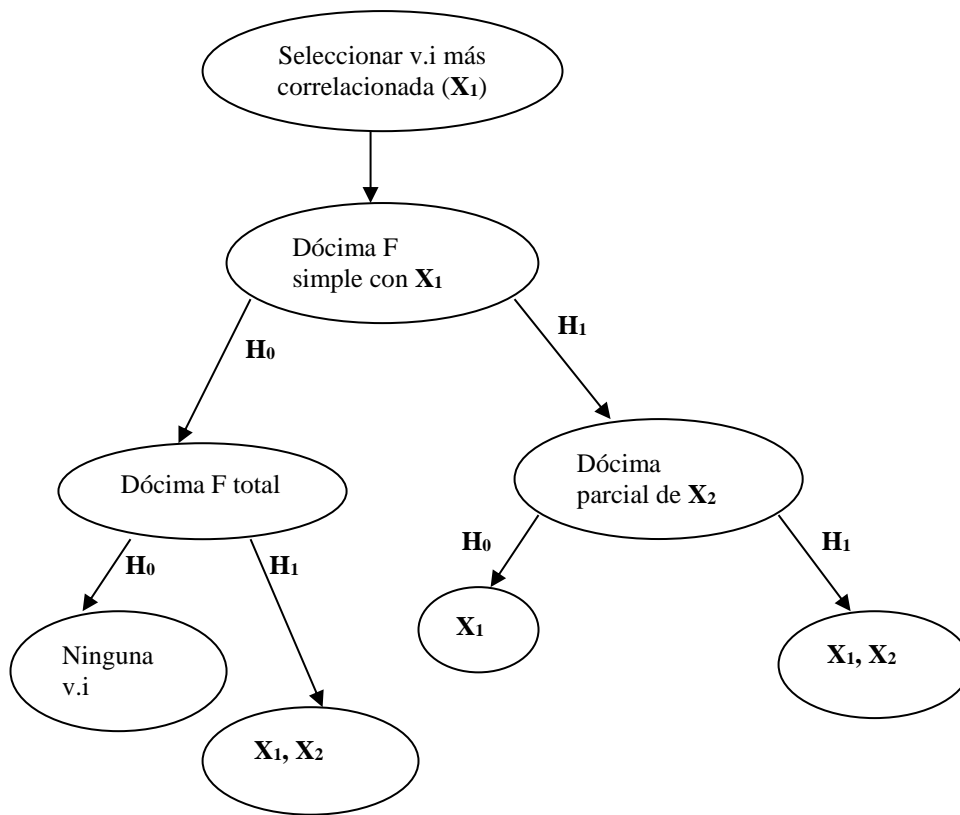
- Aplicar dócima F simple con la v.i seleccionada en 1) (Supongamos sea X_1)

- Si se rechaza H_0 , aplicar la dócima parcial de X_2 dada X_1 . Si se rechaza de nuevo H_0 quedan incluidas ambas variables, de lo contrario solo X_1

- Si no se rechaza H_0 en la dócima F simple, aplicar la prueba F total. Si se rechaza H_0 se incluyen ambas variables, en caso contrario ninguna de las variables independientes será incluida en el modelo

Existen distintos métodos habitualmente usados para seleccionar las variables independientes a incluir en un modelo de regresión múltiple. Algunos de los más usados son:

- Método de todas las regresiones posibles
- Método de selección hacia delante
- Método de eliminación hacia atrás
- Método de selección paso a paso

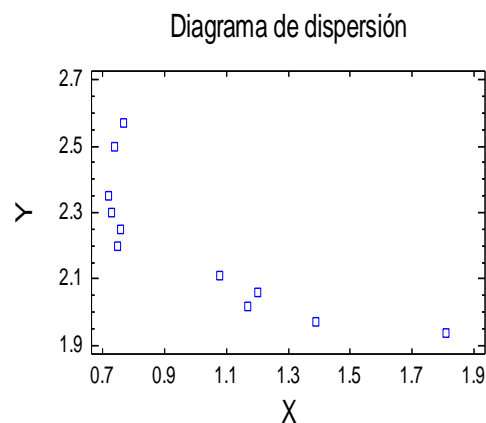


Ejemplos:

Ej1: Los siguientes datos corresponden a los precios al detalle (X- \$ por libra) y el consumo (Y-tazas diarias por persona) de café en E.U. en los últimos años.

Según la Microeconomía, la demanda de un producto de consumo depende del precio de ese producto, de los precios de otros bienes sustitutivos y complementarios y del ingreso del consumidor. Por ahora supondremos una función de demanda parcial ó ceteris paribus, suponiendo todas las demás variables independientes de influencia en la demanda de café constantes.

Año	Y	X
2001	2.57	0.77
2002	2.50	0.74
2003	2.35	0.72
2004	2.30	0.73
2005	2.25	0.76
2006	2.20	0.75
2007	2.11	1.08
2008	1.94	1.81
2009	1.97	1.39
2010	2.06	1.20
2011	2.02	1.17



Se observa la relación ya analizada entre ambas variables (mayor precio-menor consumo), en cuanto al tipo de relación pudiera asumirse una relación lineal o tal vez hiperbólica, de acuerdo a la forma del conjunto de puntos, aunque la cantidad de estos es escasa.

Y	X	X ²	Y ²	XY
2.57	0.77	0.59	6,60	1.98
2.50	0.74	0.55	6,25	1.85
2.35	0.72	0.52	5,52	1.69
2.30	0.73	0.53	5,29	1.68
2.25	0.76	0.58	5,06	1.71
2.20	0.75	0.56	4,84	1.65
2.11	1.08	1.17	4,45	2.28
1.94	1.81	3.28	3,76	3.51
1.97	1.39	1.93	3,88	2.74
2.06	1.20	1.44	4,24	2.47
2.02	1.17	1.37	4,08	2.36
24,27	11,12	12,52	53,99	23,92

$$r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}}$$

$$r = \frac{11 \cdot 23,92 - 11,12 \cdot 24,27}{\sqrt{[11 \cdot 12,52 - (11,12)^2][11 \cdot 53,99 - (24,27)^2]}}$$

$$r = \frac{-6,7624}{\sqrt{[14,0656][4,8571]}} = -0,814$$

$$b_2 = \frac{11 \cdot 23,92 - 24,27 \cdot 11,12}{11 \cdot 12,52 - (11,12)^2} = \frac{-6,7624}{14,0656} = -0,48 \quad b_1 = \frac{24,27}{11} - (-0,48) \frac{11,12}{11} = 2,69$$

$$\hat{Y} = 2,69 - 0,48X$$

Por ejemplo, al fijar un precio de \$0,80 la libra se espera un consumo medio de alrededor de:

$$\hat{Y}_{X=0,80} = 2,69 - 0,48 \cdot 0,80 = 2,306 \text{ tazas diarias por persona.}$$

Prueba F y T.

$$H_0: \beta_2 = 0 \quad \text{en} \quad E(y/x) = \beta_1 + \beta_2 x \quad (b_1 + b_2 x \text{ no es mejor que } \bar{y}).$$

$$H_1: \beta_2 \neq 0 \quad \text{en} \quad E(y/x) = \beta_1 + \beta_2 x \quad (b_1 + b_2 x \text{ es mejor que } \bar{y}).$$

$$SCT = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 53,99 - \frac{24,27^2}{11} = 0,442$$

$$SC \text{ Reg} = b_2 \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] = -0,48 \left[23,92 - \frac{11,12 \cdot 24,27}{11} \right] = 0,293$$

$$SC \text{ Res} = SCT - SC \text{ Reg} = 0,149$$

F.Variación	S.C	g.l	C.M	F
Total	SCT	10	-	
Regresión	SCReg	1	0,293	17,69
Error	SCE	9	0,017	

$$R^2 = \frac{SC \text{ Reg}}{SCT} = \frac{0,293}{0,442} = 0,6628$$

Región crítica o de rechazo: $F > F_{\alpha}(1; n - 2) \Leftrightarrow |t| > t_{\alpha/2}(n - 2)$ con $t = \sqrt{F}$

Si fijamos $\alpha = 0,05$ se tendrá $F_{\alpha}(1; n - 2) = F_{0,05}(1,9) = 5,12$ de donde, puesto que

$F = 18,25 > 5,12$ se rechaza H_0 y por tanto, existe una influencia significativa del precio del café en su consumo a través del modelo elegido

(Para t : $t_{\alpha/2}; n - 2 = t_{0,025,9} = 2,262$ y $t = \sqrt{F} = \sqrt{18,25} = 4,27$; con el mismo resultado)

Asumiendo un nivel de confianza del 95% para la estimación se obtiene:

$$\hat{Y}_0 \mp t_{\frac{\alpha}{2}}(n-2) \sqrt{CME \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)} = 2,306 \mp t_{0,025}(9) \sqrt{0,017 \left(\frac{1}{11} + \frac{\left(0,80 - \frac{11,12}{11}\right)^2}{12,52 - \frac{11,12^2}{11}} \right)}$$

$$= 2,306 \mp 2,262 \sqrt{0,017 \left(0,0909 + \frac{0,0445}{1,2741} \right)} = 2,306 \mp 0,1032$$

Por tanto, con una confiabilidad del 95% puede afirmarse que el consumo medio de café para distintos períodos de tiempo en que este tenga un precio de \$0,80 por libra de café, estará entre 2,2028 y 2,4092 tazas diarias por persona

La estimación correspondiente para un período dado estaría dada por

$$\hat{Y}_0 \mp t_{\frac{\alpha}{2}}(n-2) \sqrt{CME \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)} = 2,306 \mp t_{0,025}(9) \sqrt{0,016 \left(1 + \frac{1}{11} + \frac{\left(0,80 - \frac{11,12}{11}\right)^2}{12,52 - \frac{11,12^2}{11}} \right)}$$

$$= 2,306 \mp 2,262 \sqrt{0,016 \left(1 + 0,0909 + \frac{0,0445}{1,2741} \right)} = 2,306 \mp 0,1366$$

Con una confiabilidad del 95% puede afirmarse que el consumo de café para un período en que este tenga un precio de \$0,80 por libra de café, estará entre 2,1694 y 2,4426 tazas diarias por persona

En el ejemplo anterior propondremos probar con el modelo inverso de X es decir:

$$E(Y/X) = \beta_1 + \beta_2 \frac{1}{X} \Leftrightarrow Y = \beta_1 + \beta_2 \frac{1}{X} + \varepsilon$$

Aquí podemos ver que obtenemos resultados superiores al anterior:

Coefficientes

	Mínimos Cuadrados	Estándar	Estadístico	
<i>Parámetro</i>	<i>Estimado</i>	<i>Error</i>	<i>T</i>	<i>Valor-P</i>
Intercepto	1.57774	0.131788	11.9718	0.0000
Pendiente	0.578952	0.1171	4.94408	0.0008

Análisis de Varianza

Fuente	Suma de Cuadrados	Gl	Cuadrado Medio	Razón-F	Valor-P
Modelo	0.323094	1	0.323094	24.44	0.0008
Residuo	0.11896	9	0.0132178		
Total (Corr.)	0.442055	10			

Coefficiente de Correlación = 0.854923

R-cuadrada = 73.0893 por ciento

R-cuadrado (ajustado para g.l.) = 70.0992 por ciento

Error estándar del est. = 0.114969

Error absoluto medio = 0.0768492

Estadístico Durbin-Watson = 0.640994 (P=0.0009)

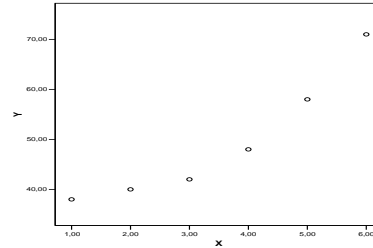
Autocorrelación de residuos en retraso 1 = 0.425029

Ej2: En una fábrica se han registrado los datos relativos al *costo de producción (Y)* y el *consumo de materias primas (X)* durante varios períodos, con los siguientes resultados:

X	Y
1	38
2	40
3	42
4	48
5	58
6	71

- Dibuje el diagrama de dispersión
- Calcule e interprete el coeficiente de correlación lineal
- Calcule e interprete el coeficiente de correlación lineal entre Y y X². Analice los resultados y su relación con el diagrama de dispersión.
- Halle la ecuación que mejor exprese la relación entre ambas variables.
- Estime el costo de producción para un consumo de materias primas de 3,5

X	Y	X ²	Y ²	XY	X ² Y	X ⁴
1	38	1	1444.00	38	38.00	1.00
2	40	4	1600.00	80	160.00	16.00
3	42	9	1764.00	126	378.00	81.00
4	48	16	2304.00	192	768.00	256.00
5	58	25	3364.00	290	1450.00	625.00
6	71	36	5041.00	426	2556.00	1296.00
21	297	91	15517.00	1152.00	5350.00	2275.00



$$b) r = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} = \frac{6.1152 - 21.297}{\sqrt{[6.91 - 21^2][6.15517 - 297^2]}} = \frac{675}{\sqrt{[546 - 441][93102 - 88209]}} = \frac{675}{\sqrt{105.4893}} = \frac{675}{\sqrt{513765}} = \frac{675}{716,77} = 0,9417$$

$$c) r_{X^2Y} = \frac{n \sum X^2Y - \sum X^2 \sum Y}{\sqrt{[n \sum X^4 - (\sum X^2)^2][n \sum Y^2 - (\sum Y)^2]}} = \frac{6.5350 - 91.297}{\sqrt{[6.2275 - 91^2][6.15517 - 297^2]}} = \frac{5073}{\sqrt{[13650 - 8281][93102 - 88209]}} = \frac{5073}{\sqrt{5369.4893}} = \frac{5073}{\sqrt{26270517}} = \frac{5073}{5125,48} = 0,9898$$

El coeficiente de correlación lineal entre X² y Y es mayor que entre X e Y. Ello indica que la relación entre las variables está mejor representada a través de un modelo del tipo $Y = \beta_1 + \beta_2 X^2 + \varepsilon$ que a través del modelo $Y = \beta_1 + \beta_2 X + \varepsilon$, lo cual concuerda con lo que se observa en el diagrama de dispersión

Quedarían para el modelo elegido como:

$$b_2 = \frac{n \sum X^2Y - \sum X^2 \sum Y}{n \sum X^4 - (\sum X^2)^2} = \frac{5073}{5369} = 0,94 \quad b_1 = \bar{Y} - b_2 \frac{\sum X^2}{n} = \frac{297 - 0,94.91}{6} = \frac{211,46}{6} = 35,24$$

$$\hat{Y} = 35,24 + 0,94.X^2$$

$$e) \hat{Y}_{X=3,5} = 35,24 + 0,94.3,5^2 = 35,24 + 11,52 = 46,76$$

Regresión Simple - Y vs. X

Lineal: $Y = a + b \cdot X$

Coeficientes					Análisis de Varianza					
	MC	Estándar	Estadístico		Fuente	SC	Gl	CM	Razón-F	Valor-P
Parámetro	Estimado	Error	T	Valor-P	Modelo	723.214	1	723.214	31.35	0.0050
Intercepto	27.0	4.4716	6.0381	0.0038	Residuo	92.2857	4	23.0714		
Pendiente	6.42857	1.1482	5.59882	0.0050	Total (Corr.)	815.5	5			

Coefficiente de Correlación = 0.941719 R-cuadrada = 88.6835 por ciento R-cuadrado (ajustado para g.l.) = 85.8544 por ciento
 Error estándar del est. = 4.80327 Error absoluto medio = 3.38095 Estadístico Durbin-Watson = 1.03317 (P=0.0113)
 Autocorrelación de residuos en retraso 1 = 0.210526

Regresión Simple - Y vs. X

Cuadrado de X: $Y = a + b \cdot X^2$

Coeficientes					Análisis de Varianza					
	MC	Estándar	Estadístico		Fuente	SC	Gl	CM	Razón-F	Valor-P
Parámetro	Estimado	Error	T	Valor-P	Modelo	798.886	1	798.886	192.35	0.0002
Intercepto	35.1695	1.32661	26.5107	0.0000	Residuo	16.6135	4	4.15338		
Pendiente	0.944869	0.0681287	13.8689	0.0002	Total (Corr.)	815.5	5			

Coefficiente de Correlación = 0.989761 R-cuadrada = 97.9628 por ciento R-cuadrado (ajustado para g.l.) = 97.4535 por ciento
 Error estándar del est. = 2.03798 Error absoluto medio = 1.58397 Estadístico Durbin-Watson = 1.05503 (P=0.0137)
 Autocorrelación de residuos en retraso 1 = 0.266304

Ej3: Un ejecutivo desea determinar si existe relación entre el promedio de Ventas de sus tiendas con el número de competidores cercanos y el ingreso promedio de las personas del distrito donde se sitúan sus establecimientos. Para ello se tomó una muestra en diez tiendas diferentes y los resultados son los siguientes:

(Y)- Promedio de Ventas Diarias (Miles de dólares) $\sum Y = 20,5$ $\sum Y^2 = 49,13$

(X₁)- Número de Competidores en el distrito. $\sum X_1 = 98$ $\sum X_1^2 = 1604$ $\sum YX_1 = 133,6$

(X₂)- Ingreso per cápita Anual (Miles). $\sum X_2 = 339,8$ $\sum X_2^2 = 12734,16$ $\sum YX_2 = 788,15$

- Realice todos los pasos necesarios por el método paso a paso para establecer la mejor relación lineal. En caso de incluir las dos variables tenga en cuenta que: SCReg $X_1X_2 = 7,087$, $b_0 = 1$, $b_1 = 0,04$, $b_2 = -0,04$.
- Estime el promedio de Ventas de una tienda cuyo número de competidores en el distrito sea 15 y el ingreso per cápita anual del mismo sea 20 000 \$.

Respuesta: Método paso a paso.

- Coeficiente de correlación lineal para cada variable.

$$r_{x_1Y} = \frac{\sum X_1Y - \frac{\sum X_1 \sum Y}{n}}{\sqrt{\left[\sum X_1^2 - \frac{(\sum X_1)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}} = \frac{133,6 - \frac{98 \cdot 20,5}{10}}{\sqrt{\left[1604 - \frac{98^2}{10}\right] \left[49,13 - \frac{20,5^2}{10}\right]}} = \frac{-67,3}{\sqrt{[643,6][7,105]}} = 0,9952$$

$$r_{x_2Y} = \frac{\sum X_2Y - \frac{\sum X_2 \sum Y}{n}}{\sqrt{\left[\sum X_2^2 - \frac{(\sum X_2)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}} = \frac{788,15 - \frac{339,8 \cdot 20,5}{10}}{\sqrt{\left[12734,16 - \frac{339,8^2}{10}\right] \left[49,13 - \frac{20,5^2}{10}\right]}}$$

$$= \frac{91,56}{\sqrt{[1187,756][7,105]}} = 0,9967$$

variable más correlacionada X₂.

- Décima pendiente para X_2

$$H_0: \beta_3 = 0 \quad \text{en} \quad E(y/x) = \beta_1 + \beta_3 x_2.$$

$$H_1: \beta_3 \neq 0 \quad \text{en} \quad E(y/x) = \beta_1 + \beta_3 x_2.$$

$$SCT = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 7,105$$

$$SC \text{ Reg} = b_2 \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] = \frac{\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right]^2}{\sum X_2^2 - \frac{(\sum X_1)^2}{n}} = \frac{(91,56)^2}{1187,756} = 7,058$$

$$SC \text{ Res} = SCT - SC \text{ Reg} = 0,047$$

F.Variación	S.C	g.l	C.M	F
Total	7,105	9	-	
Regresión X_2	7,058	1	7,058	1176,33
Error	0,047	8	0,006	

Región crítica o de rechazo: $F > F_\alpha(1; n - 2)$

Si fijamos $\alpha = 0,05$ se tendrá $F_\alpha(1; n - 2) = F_{0,05}(1,8) = 5,32$ de donde, puesto que

$F = 1176,25 > 5,32$ se rechaza H_0 y por tanto, existe una influencia significativa del Ingreso per cápita en las ventas.

- Décima parcial de X_1/X_2

$$H_0: \beta_2 = 0 / \beta_3 \neq 0 \quad \text{en} \quad Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

$$H_1: \beta_2 \neq 0 / \beta_3 \neq 0 \quad \text{en} \quad Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

$$SC \text{ Reg}_{X_1/X_2} = SC \text{ Reg}_{X_1 X_2} - SC \text{ Reg}_{X_2} = 7,087 - 7,058 = 0,029$$

$$SCE_{X_1 X_2} = SCT - SC \text{ Reg}_{X_1 X_2} = 7,105 - 7,087 = 0,018$$

F.Variación	S.C	g.l	C.M	F
Reg X_1 / X_2	0,029	1	0,029	11,15
Error $X_1 X_2$	0,018	7	0,0026	

Región crítica o de rechazo: $F > F_\alpha(1; n - 3)$

$F = 11,15 > F_{0,05}(1, 7) = 5,59$. Se rechaza H_0 . Se incluye también la variable X_1 .

El modelo quedaría:

$$\hat{Y} = 1 + 0,04 X_1 - 0,04 X_2$$

$$b) \hat{Y} = 1 + 0,04 \cdot 15 - 0,04 \cdot 20 = 0,8$$

Se estima que el promedio de ventas será de 800 dólares como promedio.

Ej4: Una Empresa de transporte necesita realizar un estudio acerca de los Ingresos anuales que obtiene por concepto de renta de autos. Supone que estos puedan tener relación con la gestión que realizan sus empleados a partir de los años de experiencia y el número de modelos de carros que poseen en los diferentes puntos de renta. Para ello se tomó una muestra de diez diferentes empleados y los resultados son los siguientes:

(Y)- Ingresos (Miles de dólares) $\sum Y = 20,5$ $\sum Y^2 = 49,13$

(X₁)- Número de modelos de carros. $\sum X_1 = 53$ $\sum X_1^2 = 395$ $\sum YX_1 = 83$.

(X₂)- Años de experiencia. $\sum X_2 = 85$ $\sum X_2^2 = 778$ $\sum YX_2 = 179$

- a) Realice todos los pasos necesarios para establecer la mejor relación lineal. En caso de incluir las dos variables tenga en cuenta que: SCReg X₁X₂=311, b₀= 0,03 , b₁= -0,15, b₂= 0.33.

Ej5: Dados los siguientes valores observados de las variables Y, X₁ y X₂

Y	X ₁	X ₂
10	1	3
11	1,2	3,5
9,4	2	5
6	2,6	6
3	3	6,8
7,3	3,4	7
8	3,7	7,8
6,9	4	8
10	4,2	8,6
10,3	4,8	9

Determine cuáles variables independientes serán incluidas en un modelo de regresión lineal, sabiendo que SCReg_{X₁ X₂} = 18,98

Datos:

$$\sum Y = 81,9 \quad \sum X_1 = 29,9 \quad \sum X_1^2 = 104,13 \quad \sum X_2 Y = 519,3 \quad \sum X_2^2 = 457,29$$

$$\sum X_1 Y = 240,06 \quad ; \quad \sum Y^2 = 725,35$$

Solución (Análisis inicial a través de SCReg)

$$SCT = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 725,35 - \frac{81,9^2}{10} = 725,35 - 670,761 = 54,589$$

Para el modelo con X₁ solamente

$$b_2 = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2} = \frac{10 \cdot 240,06 - 29,9 \cdot 81,9}{10 \cdot 104,13 - 29,9^2} = \frac{-48,21}{147,29} = -0,33$$

$$b_1 = \bar{Y} - b_2 \bar{X} = 9,17$$

$$SCReg = b_2 \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] = 1,58$$

$$SCRes = SCT - SCReg = 54,589 - 1,58 = 53,009$$

Para el modelo con X₂ solamente

$$b_2 = -0,27 \quad b_1 = 9,96$$

$$SCReg = b_2 \left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n} \right] = 2,9$$

$$SCRes = SCT - SCReg = 54,589 - 2,9 = 51,689$$

X₂ es la variable independiente más correlacionada con Y pues posee la mayor suma de cuadrados de la regresión

Dócima simple con X_2

$$H_0: \beta_2 = 0 \text{ en } Y = \beta_1 + \beta_2 X_2 + U$$

$$H_1: \beta_2 \neq 0 \text{ en } Y = \beta_1 + \beta_2 X_2 + U$$

Región crítica o de rechazo: $F > F_{\alpha}(1; n - 2)$

F.Variación	S.C	g.l	C.M	F
Total	54,589	9	-	
Regresión	2,9	1	2,9	0,45
Error	51,689	8	6,46	

$$F = 0,45 < F_{0,05}(1,8) = 5,32$$

La variable X_2 por sí sola no ejerce una influencia significativa en Y.

Será necesario realizar la dócima F total incluyendo ambas variables independientes

$$H_0: \beta_2 = \beta_3 = 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

$$H_1: \beta_2 \neq 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

Región crítica o de rechazo: $F > F_{\alpha}(2; n - 3)$

F.Variación	S.C	g.l	C.M	F
Total	54,589	9	-	
Regresión	18,98	2	9,49	1,87
Error	35,609	7	5,087	

$$F = 1,87 < F_{0,05}(2,7) = 4,74$$

Las variables X_1 y X_2 conjuntamente no explican una cantidad significativa de la variación de Y
Ninguna de las dos variables independientes serían incluidas en el modelo de regresión.

Ej6: El Ministerio de la Industria Básica está evaluando distintos proyectos de inversión para los cuáles intenta determinar el *costo de inversión* a partir de los indicadores X_1 : *costo de montaje* y X_2 : *costo de construcción*, expresados en miles de pesos. Al recopilar los datos de estas variables en 30 proyectos anteriores de tipo similar a los que se consideran se obtuvieron los siguientes resultados al ajustar los distintos modelos posibles:

Modelo	b_1	b_2	b_3	SCT	SCE	SCReg
$Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$	9,88	0,35	0,27		7,32	
$Y = \beta_1 + \beta_2 X_1 + U$	10,2	0,24		57,73		33,2
$Y = \beta_1 + \beta_3 X_2 + U$	9,23		0,19			30,6

a) Determine cuál de las variables X_1 ó X_2 es necesaria para estimar el costo de inversión. Use $\alpha = 0,01$.

b) Estime el costo de inversión de un proyecto con un costo de montaje igual a 30 mil pesos y un costo de construcción de 20 mil

R/ a) La variable independiente más correlacionada con Y es X_1 (mayor SCReg)

Dócima simple con X_1

$$H_0: \beta_2 = 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + U$$

$$H_1: \beta_2 \neq 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + U$$

$$F = 37,73 > F_{0,01}(1,28) = 7,56$$

Se incluye la variable X_1

F.Variación	S.C	g.l	C.M	F
Total	57,73	29	-	
Regresión	33,20	1	33,20	37,73
Error	24,53	28	0,88	

Dócima parcial de X_2/X_1

$$H_0: \beta_3 = 0 / \beta_2 \neq 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

$$H_1: \beta_3 \neq 0 / \beta_2 \neq 0 \text{ en } Y = \beta_1 + \beta_2 X_1 + \beta_3 X_2 + U$$

$$SC\ Re\ g_{X_2/X_1} = SC\ Re\ g_{X_1 X_2} - SC\ Re\ g_{X_1} = (57,73 - 7,32) - 33,20 = 50,41 - 33,20 = 17,21$$

F.Variación	S.C	g.l	C.M	F
Total	57,73	29	-	
Regresión X_1	33,20	1	33,20	37,73
Reg X_2 / X_1	17,21	1	17,21	63,74
Reg X_1X_2	50,41	2		
Error X_1X_2	7,32	27	0,27	

Región crítica o de rechazo: $F > F_{\alpha}(1; n - 3)$

$F = 63,74 > F_{0,01}(1,27) = 7,77$. Se rechaza H_0 . Se incluye también la variable X_2 – costo de construcción. Ambas variables son necesarias para estimar el costo de inversión

$$\begin{aligned} \text{b) } \hat{Y} &= 9,88 + 0,35X_1 + 0,27X_2 = 9,88 + 0,35 \cdot 30 + 0,27 \cdot 20 = \\ &= 9,88 + 10,5 + 5,4 = 20,38 + 5,4 = 25,78 \end{aligned}$$

Se estima un costo promedio de inversión de 25 mil 780 pesos para un proyecto con un costo de montaje igual a 30 mil pesos y un costo de construcción de 20 mil

Estudio Independiente:

Ejercicios L/T. Miller pág. 315; 331; 343; 350.

Ejercicios L/T. Walpole pág. 398; 436 (Ejercicios de Repaso)

Ejercicios Guía Tema IV.