UNIVERSIDAD NACIONAL DE INGENIERÍA

UNI- NORTE - SEDE REGIONAL ESTELÍ

Asignatura: Estadística II

Unidad No.1

Análisis de Regresión Multiple

Objetivos

- Introducir conceptos de Correlación y Regresión Lineal.
- Explicar la forma de cálculo.
- Realizar las pruebas de hipótesis asociadas

Docente: Ing. MS Luis María Dicovskiy Riobóo

Contenido

Análisis de Correlación

Es frecuente que estudiemos sobre una misma población los valores de dos o más variables estadísticas distintas, con el fin de ver si existe alguna relación entre ellas, es decir, si los cambios en una o varias de ellas influyen en los valores de la variable dependiente. Si ocurre esto decimos que las variables están correlacionadas o bien que hay correlación entre ellas. Este tipo de análisis funcione bien cuando las variables estudiadas son continuas, no es adecuado usar esta prueba con variables del tipo nominal.

El análisis de correlación es el conjunto de técnicas estadísticas empleado para medir la intensidad de la asociación entre dos variables. El principal objetivo del análisis de correlación consiste en determinar que tan intensa es la relación entre dos variables, estas pueden ser.

- *Variable Dependiente*.- es la variable que se predice o calcula. Cuya representación es "Y"
- *Variable Independiente*.- es la o las variables que proporcionan las bases para el calculo. Cuya representación es: “X”. Esta o estas variables suelen ocurrir antes en el tiempo que la variable dependiente.

Coeficiente de Correlación. El coeficiente de correlación más utilizado es el de Pearson, este es un índice estadístico que mide la relación lineal entre dos variables cuantitativas, es una forma de medir la intensidad de la relación lineal

entre dos variables. El valor del coeficiente de correlación puede tomar valores desde menos uno hasta uno, -1 < r < 1, indicando que mientras más cercano a uno sea el valor del coeficiente de correlación, en cualquier dirección, más fuerte será la asociación lineal entre las dos variables. El coeficiente de correlación de cálculo "r" es un estimador muestral del coeficiente poblacional Rho, ρ.

Mientras más cercano a cero sea el coeficiente de correlación, este indicará que más débil es la asociación entre ambas variables. Si es igual a cero se concluirá que no existe relación lineal alguna entre ambas variables. Hay varias maneras de equivalentes de calcular "r", a continuación se muestran tres formas.

Coeficiente Correlación Fórmula por Covarianzas y Desviaciones Típicas

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Siendo: "σ_{XY}" la covarianza de (X,Y) y "σ_X, σ_Y" las desviaciones típicas de las distribuciones de las variables independiente y dependiente respectivamente.

Coeficiente Correlación Fórmula Clásica. Poco usada para cálculo.

$$r = \sqrt{\frac{\left[\sum \left(X - \overline{X}\right)\left(Y - \overline{Y}\right)\right]^2}{\left[\sum \left(X - \overline{X}\right)^2 \sum \left(Y - \overline{Y}\right)^2\right]}}$$

Coeficiente Correlación, Fórmula por suma de cuadrados. Se usa cuando se dispone de calculadoras de mano que hacen sumas de cuadrados y no correlación.

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

Gráfico de Dispersión de puntos. Es un diagrama de dispersión de punto X Y, el cual es una representación gráfica de la relación entre dos variables, muy utilizada en las fases de comprobación de teorías e identificación de causas raíz y en el diseño de soluciones y mantenimiento de los resultados obtenidos. Tres conceptos especialmente son destacables: que el descubrimiento de las verdaderas relaciones de causa-efecto es la clave de la resolución eficaz de un problema, que las relaciones de causa-efecto casi siempre muestran variaciones, y que es más fácil ver la relación en un diagrama de dispersión que en una simple tabla de números.

Según sea la dispersión de los datos (nube de puntos) en el plano cartesiano, pueden darse alguna de las siguientes relaciones, Lineal, Logarítmica, Exponencial, Cuadrática, entre otras. Estas nubes de puntos pueden generar polígonos a partir de ecuaciones de regresión que permitan predecir el comportamiento de la variable dependiente.

Regresión

La **regresión estadística** o **regresión a la media** es la tendencia de una medición extrema a presentarse más cercana a la media en una segunda medición. La regresión se utiliza para predecir una medida basándonos en el conocimiento de otra. El término regresión fue introducido por Francis Galton en su libro *Natural inheritance* (1889), partiendo de los análisis estadísticos de Karl Pearson. Su trabajo se centró en la descripción de los rasgos físicos de los descendientes a partir de los de sus padres. Estudiando la altura de padres e hijos

llegó a la conclusión de que los padres muy altos tenían una tendencia a tener hijos que heredaban parte de esta altura, pero los datos revelaban también una tendencia a *regresar* a la media.

Los tipos de regresión más comunes entre dos variables son las del tipo polinómico como la regresión: lineal, cuadrática y cúbica. La primera regresión genera una recta, las otras diferentes tipos de parábolas. Otros tipos de regresión que se pueden usar con dos variables son la logarítmica y la exponencial, la regresión logarítmica permite transformar una curva en una línea recta. Cuando hay más de una variable independiente "x", la regresión más utilizada en la regresión múltiple. A continuación se expresan matemáticamente los diferentes modelos comentados:

REGRESIÓN	**ECUACIÓN**
Lineal	$y = b_0 + b_1 x$
Logarítmica	$y = b_0 + b_1 \text{Ln}(x)$
Exponencial	$y = b_0 e^{(b_1 x)}$
Cuadrática	$y = b_0 + b_1 x + b_2 x^2$
Cúbica	$y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$
Lineal Múltiple	$y = b_0 + b_1 x_1 + b_2 x_2 \ldots + b_n x_n$

Ecuación de Regresión Lineal

Es el tipo de regresión más utilizada y fácil de estimar, esta es una ecuación que define la relación lineal entre dos variables.

Ecuación de regresión lineal $\hat{Y} = b_0 + b_1 x$

Esta ecuación se calcula según el principio de Mínimos Cuadrados. La cual es la técnica empleada para obtener la ecuación de regresión, minimizando la suma de los cuadrados de las distancias verticales entre los valores verdaderos de "Y", los

observados y los valores estimados "$\hat{Y}$". Se debe notar que el valor observado menos el valor estimado genera un residuo que llamaremos error, este residuo o error, es la distancia que hay del valor observado a la recta de regresión. Se deduce que el error de para cada dato se encuentra de la siguiente manera:

$$\left(y_i - \hat{y}_i\right) = \varepsilon_i$$

A ε_i se le llama error aleatorio, es la diferencia entre el valor observado "y_i" menos el valor estimado "$\hat{y}_i$, esta es una distancia entre ambos valores y esta puede ser negativa o positiva y tienen la siguiente propiedad:

- $E(\varepsilon_i) = 0$
- Los desvíos, ε se distribuyen de manera Normal.

La primera propiedad indica que en promedio los errores son iguales a cero, al igual que la sumatoria de los mismos. Lo segundo que los errores se distribuyen de manera normal con promedio de 0.

La ecuación que minimizar la desviaciones de los valores de "Y" respecto a la ecuación de la recta, cuando "b_0= 0", es:

$$\hat{Y} = \left(\frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}} \right) X$$

$$\hat{Y} = b_1 X$$

Por lo tanto la Expresión del coeficiente de regresión, "b_1", queda así:

$$b_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}$$

Como podemos escribir:

$$\left(\hat{Y}-\overline{Y}\right)=b_1\left(X-\overline{X}\right)$$

Que puede replantearse como:

$$\hat{Y}=\left(Y-b_1\overline{X}\right)+b_1X$$

De tal manera que la ordenada al origen, cuando "X" vale 0, "b_0", queda definida de la siguiente manera:

$$\hat{Y}=b_0=\overline{Y}-b_1\overline{X}$$

Ejemplo de regresión correlación lineal:

Se tienen las notas de un examen parcial de diez alumnos de las asignaturas de matemáticas y español

Matemáticas	2	3	5	5	6	6	7	7	8	9
Español	2	2	5	5	6	7	5	8	7	10

Se supone que los alumnos con mejores notas en matemáticas, variable independiente "X", tienen las mejores notas en español, variable dependiente "Y". Esta pregunta se puede responder con un análisis de regresión correlación.

Lo primero que se hace es construir un gráfico de dispersión de punto como el que se muestra a continuación

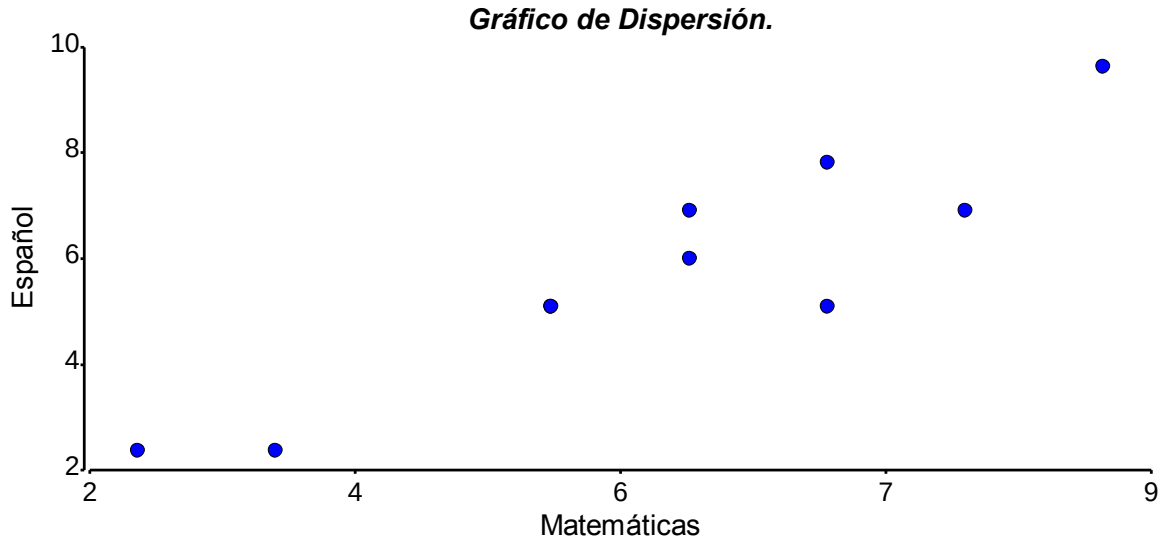

Gráfico de dispersión de puntos de las notas de las asignaturas de matemáticas y español

Datos generados con una calculadora de mano:

$$\bar{x} = 5.8, \bar{y} = 5.7, \sum x = 58, \sum x^2 = 378, \sum y = 57, \sum y^2 = 381, \sum xy = 375$$

Luego se calcula el coeficiente de correlación “r”.

$$r = \sqrt{\frac{\left(375 - (58)(57)/10\right)^2}{\left(378 - 58^2/10\right)\left(381 - 57^2/10\right)}} = 0.919$$

Este valor de “r” de 0.919 nos dice que hay una alta correlación entre las notas de matemáticas y español.

Para hacer la recta de regresión debemos calcular:

$$b_1 = \frac{375 - (58)(57)/10}{378 - 58^2/10} = 1.0673$$

$$b_0 = 5.7 - (1.0673)(5.8) = -0.4904$$

La recta de regresión queda determinada de la siguiente manera:

"$\hat{Y}$" = -0.4904 + 1.0673 X ".

A continuación se observan los valores estimados por la recta de regresión de la asignatura de español,"$\hat{Y}$", para cada valor observado "*Y*" y el desvío o error asociado a cada dato, "$Y - \hat{Y}$",estos son:

"$\hat{Y}$"	1.64	2.71	4.85	4.85	5.91	5.91	6.98	6.98	8.05	9.12
E	0.36	-.71	.15	.15	.009	1.09	-1.98	1.02	-1.05	0.88

Se puede comprobar que la suma de los desvíos es igual 0.

El gráfico de regresión es el siguiente:

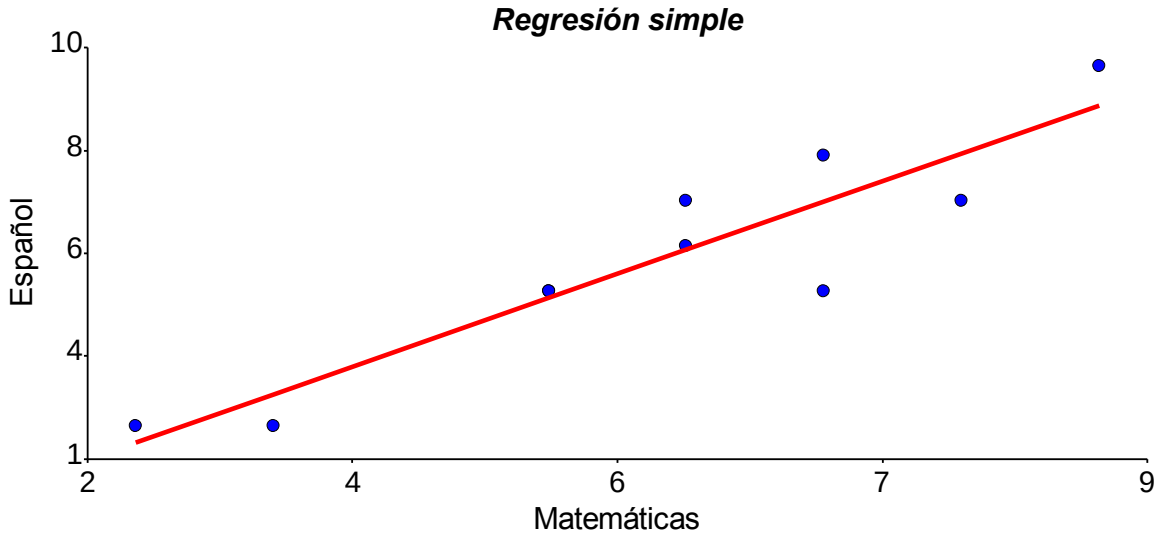

Gráfico de Regresión de la asignatura Matemática y Español. Se observa la recta de regresión y los datos observados en forma de línea discontinua.

Verificación del modelo de regresión.

Para verificar si el modelo de regresión lineal es correcto para ser utilizado con los datos que se tiene, se puede hacer con el programa INFOSTAT un gráfico Q-Q plot de residuos para observar si estos tienen un comportamiento normal. Este

gráfico se utiliza para evaluar el grado de ajuste de un conjunto de observaciones a una distribución teórica.

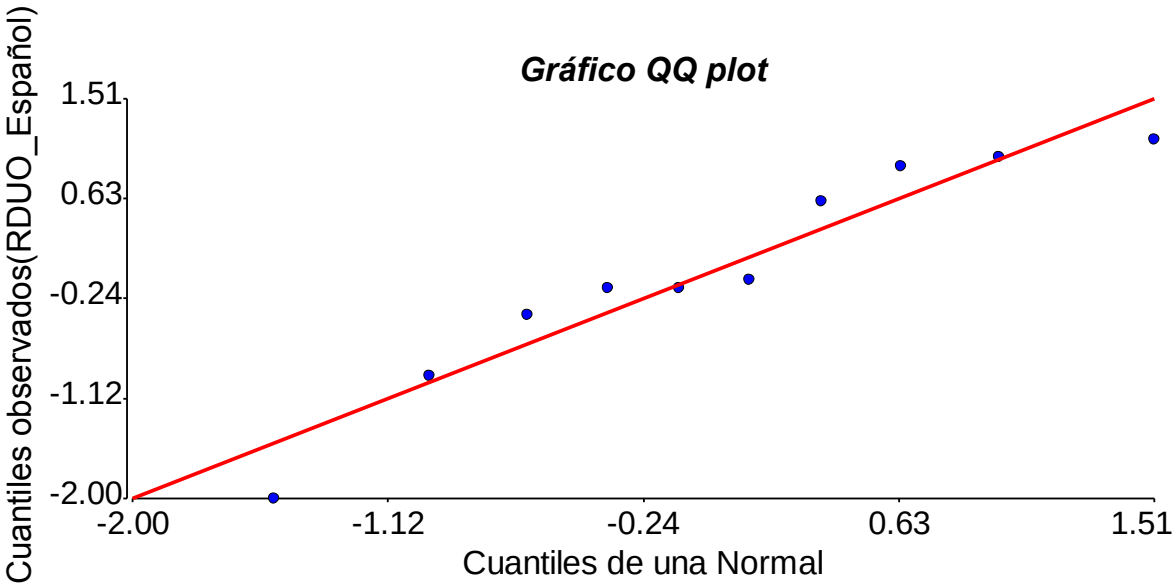

Gráfico QQ plot de los residuos de la regresión para verificar normalidad.

Sobre el rótulo del eje X se muestran los parámetros de la distribución teórica estimados a partir de la muestra. La normalidad se cumple si los puntos generados por residuos se distribuyen al azar cerca de la recta de regresión normal. También en el gráfico Q-Q plot se presenta el coeficiente de correlación lineal "*r"* de la correlación entre los cuantiles observados versus los cuantiles de la distribución teórica seleccionada, este valor debe ser de al menos de "0.95" para aceptar la normalidad. Como el "r" observado fue de 0.96, se acepta la normalidad.

Ejercicios:

a) Tomar el peso y la altura de 10 personas, hacer el gráfico de dispersión, calcular el coeficiente de correlación y la recta de regresión de estos datos.
b) Hay una hipótesis de investigación que sugiere que el gasto en comida por familia, expresado en C$ por mes, está influido directamente por el ingreso familiar mensual en C$. Haga estudio de regresión y correlación de las dos variables. Trabaje con calculadora.

Tabla de datos

Ingreso observado por familia, en cientos C$	**Gasto observados en alimentación, en cientos C$**
30	21
34	26
17	5
26	19
29	18
18	7
32	23
32	25

¿Calcular el coeficiente de regresión Lineal, "r"?

¿Construya la recta de regresión, determinar los parámetros b_0 y b_1.?

¿Determine los gastos estimados ("y" estimada) por la recta de regresión, para los ingresos observados?

¿Se quiere saber si la correlación obtenida con la muestra, es diferente de 0 en la población. Realice una prueba de hipótesis para el coeficiente de correlación "r". El valor "t" de tabla es 2.3?

¿Comente brevemente sobre los coeficientes obtenidos. Responda la hipótesis de la investigación?

Análisis de Regresión Múltiple

A menudo en una investigación el objetivo es explicar el comportamiento de una variable en términos de más de una variable, por ejemplo sea la variable **"Y"**, cuyo comportamiento explicaremos en términos de las variables $X_1, X_2,\ldots,X_k$; ahora estudiaremos la situación donde el comportamiento de la variable **"Y"** (llamada dependiente o respuesta) se explicará mediante una relación lineal en función de

las variables $X_1, X_2,\ldots,X_k$ (llamadas independientes o también explicativas). La variable respuesta y las variables explicativas deben ser cuantitativas.

Modelo

Sea **"Y"** una variable respuesta y variables in $X_1, X_2,\ldots,X_k$ dependientes; deseamos describir la relación que hay entre la variable respuesta y las variables explicativas, si entre ellas hay una relación lineal se espera que:

$$\widehat{Y_i} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}$$

Donde $\widehat{Y}$ es la variable respuesta cuantitativa para el i-ésimo objeto, este es un valor estimado.

β_k Son los parámetros poblacionales (valores constantes fijas) llamados coeficientes. Siendo "n" el número de objetos u observaciones donde i = 0, 1, 2,...., n.

Se espera que la variable dependiente varíe linealmente con las variables independientes.

Además cada valor observado "$\mathbf{y_i}$" se puede descomponer de la siguiente manera

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_{ik}$$ Para i=1, 2,..., n

Donde ε_{ik} es el desvío o error de cada observación, este valor hace único a cada dato.

Restricciones al modelo de Regresión Múltiple.

El modelo de una regresión múltiple sufre de restricciones cuando sus valores se quieren generalizar a una población, estas son:

- Las x_i son variables fijas, no aleatorias y el modelo solo se aplica a los conjuntos de x_i estudiados y no para algún conjunto mayor de valores de x_i.

- Hay una sub población de “y” con distribución normal, para cada conjunto de x_i.
- Las variancias de estas subpoblaciones de “y” de cada X_i son homocedásticas, lo que quiere decir que estiman una misma varianza poblacional.
- Los valores de “y” son independientes entre sí.

Coeficientes de Correlación Parcial y Múltiple

Coeficientes de correlación parcial

La correlación entre dos variables cuando una o más variables permanecen fijas a un nivel constante, se denomina correlación parcial, este coeficiente suele mejorar su valor respecto al coeficiente de correlación simple. También se utiliza para encontrar el coeficiente de correlación múltiple de manera general.

En el caso de tres viables, la correlación parcial entre “Y” y “X_1”con un “X_2” fijo se denota “$r_{yx1.x2}$”, y se calcula a partir de las correlaciones simples de la siguiente manera:

$$r_{yx_1.x_2} = \sqrt{\frac{(r_{yx_1} - r_{yx_2} r_{x_1x_2})^2}{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}}$$

Análogamente “$r_{yx2.x2}$” se calcula de igual forma

$$r_{yx_2.x_1} = \sqrt{\frac{(r_{yx_2} - r_{yx_1} r_{x_1x_2})^2}{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}}$$

Generalizando, siempre existe una ecuación general que permite calcular un coeficiente parcial de cualquier orden “k” conocemos tres coeficientes parciales de un orden inferior donde.

Coeficiente de correlación parcial de manera general

$$r_{yx_1.x_2x_3\ldots x_k} = \sqrt{\frac{(r_{yx_1.x_3\ldots x_k} - r_{yx_2.x_3\ldots x_k}\, r_{x_1x_2.x_3\ldots.x_k})^2}{(1-r^2_{yx_2.x_3\ldots x_k})(1-r^2_{x_1x_2.x_3\ldots x_k})}}$$

Coeficiente de correlación múltiple $r_{y.x_1x_2..x_k}$

El coeficiente de correlación múltiple mide la asociación entre varias variables independientes y una dependiente. En el caso de regresión lineal simple coincide con el coeficiente de correlación de simple.

El coeficiente de correlación múltiple se puede definir de manera general como la raíz cuadrada de la suma de los cuadrados explicados por la regresión sobre la suma de los cuadrados totales.

$$r_{y.x_1x_2..x_k} = \sqrt{\frac{\sum(\hat{Y}_i - \bar{\bar{Y}})^2}{\sum(Y_i - \bar{\bar{Y}})^2}}$$

Este coeficiente tiene una desventaja, su valor se incrementa cuando se introducen nuevas variables independientes en el modelo, por tanto resulta engañoso para el análisis.

De manera general es posible encontrar una ecuación general de coeficiente de correlación múltiple que incluye “k” variables independientes, esta se puede construir a partir de los coeficientes de correlación parciales:

$$1-r^2_{y.x_1x_2x_3....x_k}=(1-r^2_{yx_1})(1-r^2_{yx_2.x_1})(1-r^2_{yx_3.x_1x_2})....(1-r^2_{yx_k.x_1....x_{k-1}})$$

Ejemplo de cómo calcular “r” el coeficiente de correlación múltiple con tres variables “Y”, “X_1” y “X_2” a partir de correlaciones simples.

De manera operacional un ejemplo de tres variables se resuelve como

$$r_{y.x_1x_2}=\sqrt{\frac{r^2_{yx_1}+r^2_{yx_2}-2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r^2_{x_1x_2}}}$$

Se debe notar que en este ejemplo para hacer $r_{y.x1.x2}$ es necesario calcular previamente tres correlaciones simples de dos variables.

Como calcular los coeficientes b_1 y b_2 de una regresión múltiple con dos variables independientes x_1 y x_2.

Construcción del modelo:

Se parte de la ecuación de regresión múltiple

$$b_0+b_1x_1+b_2x_2=\hat{y}$$

Y se construye un sistema de ecuaciones normales

$$b_0n+b_1\sum x_1+b_2\sum x_2=\sum\hat{y}$$

$$b_0\sum x_1+b_1\sum x_1^2+b_2\sum x_2x_1=\sum\hat{y}x_1$$

$$b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2 = \sum \hat{y} x_2$$

Si se plante la ecuación en términos de desviaciones respecto a la media $\sum (x_1 - \bar{x}) = 0$, como la suma de las desviaciones es 0, entonces $\sum x_1 = \sum x_2 = \sum y = 0$. Esto implica que se anula el primer término de las tres ecuaciones y la primera ecuación, quedando el modelo de forma operativa de la siguiente manera:

Modelo de resolución

$$b_1 \sum (x_1 - x_1)^2 + b_2 \sum (x_1 - x_1)(x_2 - x_2) = \sum (x_1 - x_1)(y - y)$$

$$b_1 \sum (x_1 - x_1)(x_2 - x_2) + b_2 \sum (x_2 - x_2)^2 + = \sum (x_2 - x_2)(y - y)$$

Luego se debe despejar b_1 y b_2, se puede usar el método de Gauss Jordán o reducción, usado en álgebra lineal para resolver sistemas de ecuaciones lineares.

Para poder resolver una regresión múltiple se puede usar una calculadora de mano que tenga incorporada la función de regresión y permita calcular directamente suma de cuadrados y suma de productos de los valores de "x y". Para esto se deben utilizar las siguientes igualdades conocidas:

$$\sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n} \qquad \sum (x - \bar{x})(\sum y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$

Como todas las sumatorias se pueden calcular, este sistema de ecuaciones se resuelve haciendo cero a b_1 ó b_2 y luego despejando b_0

Ejercicio: Hay una hipótesis que sugiere que el consumo de un producto dado, expresado en unidades compradas por persona en un año está influido por: el

ingreso por persona que trabaja y el tamaño de habitantes de la ciudad. Hacer estudio de regresión u correlación para responder a la suposición.

Datos

Millones de habitantes por ciudad x_1	Ingreso per capita, en cientos C$ por habitante	Consumo del producto, unidades año
0.6	30	11
1.4	34	16
1.3	17	9
0.3	26	9
6.9	29	8
0.3	18	7
4.2	32	11
0.6	32	8

El coeficiente de regresión múltiple $r_{y.x1x2}$ es igual a

$$r_{y.x_1x_2} = \sqrt{\frac{0.00246 + 0.33 - 2(-0.049)(0.574)(0.274)}{1 - 0.075}} = 0.613$$

La regresión se plantea como un sistema de ecuaciones normales, con los siguientes valores obtenidos a partir de las sumatorias antes definidas.

$$b_1 38.38 + b_2 29.25 = -2.35$$

$$b_1 29.5 + b_2 293.5 = 74.25$$

Luego se despeja b_1 y b_2, en este ejemplo los valores son respectivamente -0.26 y 0.28. Luego se despeja b_0 sabiendo que

$$b_0 = \overline{y} - b_1 \overline{x_1} - b_2 \overline{x_2}$$

$$b_0 = 9.875 - (-0.26)1.95 - (0.28)27.25 = 2.752$$

Pruebas de Hipótesis de Correlación y Regresión

Prueba de Hipótesis del Coeficiente de correlación simple ó múltiple

Prueba de hipótesis del coeficiente de correlación poblacional Rho, (letra griega) se estima con “r” y responde a la siguiente hipótesis:

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

El estadístico de Contraste es una prueba “t” donde el:

$$"t_{calculado}" = r\sqrt{\frac{n-2}{1-r^2}}$$

Esta prueba se hace con n-2 grados de libertad.

Ejemplo con los datos del problema de regresión y correlación con las asignaturas de “matemáticas y español” donde:

$$"t_{calculado}" = 0.915\sqrt{\frac{10-2}{1-.0.915^2}} = 6.59$$

El valor 6.59 es mayor que el valor “t” de tabla de 2.3, por lo tanto se acepta como era de esperar la hipótesis alternativa, Rho es diferente de 0.

En la regresión múltiple, se deben quitar de la ecuación aquellos parámetros no significativos, junto con su variable asociada.

Análisis de Variancia, ANDEVA, para la Regresión Simple ó Múltiple

El ANDEVA en este caso responde a la pregunta de hipótesis siguiente:

$$H_0 : \beta_1 = \beta_2 = \beta_3 ... = \beta_k = 0$$

$$H_1 : no\_todos\_los\_\beta = 0$$

Esta prueba se puede usar en casos de regresión simple o de regresión múltiple.

Tabla de Análisis de Variancia, Andeva

Fuente Variación	**Suma de Cuadrados SC**	**Grados de Libertad GL**	**Cuadrado Medio CM**	**"F" Calculada**
Total	$\sum_{i=1}^{n}(y-\overline{y})^2$	n-1		
Regresión	$r^2_{y.x_1x_2..x_k}\sum(Y-\bar{Y})^2$	k	$\frac{SC_R}{GL_{Rl}}$	$\frac{CM_R}{CM_{El}}$
Desviación, error	$(1-r^2_{y.x_1x_2..x_k})\sum(Y-\bar{Y})^2$	n-k-1	$\frac{SC_E}{GL_{El}}$	

Donde "k" es el número de variables independientes y el "n" número de individuos a los cuales se les toma los datos.

Se debe considerar que:

$$\sum_{i=1}^{n}(y-\overline{y})^2=\sum_{i=1}^{n}y_i^2-\frac{(\sum_{i=1}^{n}y)^2}{n}$$

Se hizo con INFOSTAT el análisis de variancia del ejercicio anterior y se obtuvo el siguiente cuadro.

Cuadro de Análisis de la Varianza

F.V.	***SC***	***GL***	***CM***	***F***	***p-valor***
Total	56.88	7			
Regresión	21.43	2	10.71	1.51	0.3066
Error	35.45	5	7.09		

Como el p-valor es mayor a 0.05 aceptamos la H_0, los coeficientes β tienen un valor de 0, por lo tanto la regresión estimada no sirve para predecir el consumo.

Prueba de hipótesis para los coeficientes Betas

De manera particular es posible hacer una prueba de hipótesis “t” para cada coeficiente beta, donde.

$$H_0 : \beta_i = 0$$

$$H_1 : \beta_i \neq 0$$

Donde: $t_{calculado} = \frac{b_i}{S_{b_i}}$

Con n-k-1 grados de libertad

Intervalos de Confianza de los coeficientes Betas

También se pueden construir intervalos de confianza para los diferentes coeficientes de regresión Betas, estos se harían de la siguiente manera:

$$\beta_i \pm t_{(1-\alpha/2),(n-k-1)} S_{b_i}$$

Donde $S_{b_i} = \sqrt{\frac{CM_{error}}{\sum x^2 - (\sum x)^2 / n}}$

Ejercicio

Se hizo un estudio correlación múltiple con 4 variables independientes, que se cree sirven para caracterizar el valor de venta de un producto industrial. Las variables independientes son “vida útil del producto”, “Resistencia del producto”, “apreciación visual de la calidad” y “precio de costo del producto”. La variable dependiente era “valor de venta”, fijado por los compradores. Se hizo la regresión y el análisis de variancia de la regresión

Análisis de Variancia de la regresión

Modelo	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	“F”
Regresión	18.5	4		
Residual	12.0	20		
Total	30.5	24		

Nota: el valor F de tabla es 2.87

¿Plantee las 2 hipótesis correspondientes del ANDEVA para una regresión múltiple, con 4 variables independientes?

¿Complete la Tabla de ANDEVA. El valor “F” de tabla es 2.71?

¿Interprete el valor “F” del ANDEVA y responda a la prueba de hipótesis?

Referencias

Sifuentes, V.2002. Curso Análisis Multivariante aplicado a la industria pesquera. IMARPE.

Daniel, W. 2006. Bioestadística. Base para el análisis de las ciencias de la salud 4ta Edic. Edit Limusa Wiley. 924 p

Little T y Hills, J. 1990. Métodos estadísticos para la investigación en la agricultura. Edit Trillas. 270 pp.

Ross,S. 2002. Probabilidad y estadística para ingenieros. Ed Mc Graw Hill. 585 pp.