Aplicación de Correlación en la Investigación
Correlation Application in Research

Guillen, A., M.H. Badii & M.S. Acuña Zepeda
UANL, San Nicolás, N.L., México, aguillen77@yahoo.com

Abstract: Se describe y explica la aplicación de correlación paramétrica en la investigación. Se demuestra con ejemplos la forma de establecer la significancia de los parámetros de correlación y regresión. Se explica y se calcula los intervalos de confianza para estos parámetros.
Palabras clave: Estadística paramétrica, intervalos de confianza, correlación y regresión, parámetros

Resumen: Parametric correlation technique is described and explained. Examples are provided to show the establishment of statistical significance for correlation and regression parameters. Confidence intervals for these parameters are calculated and explained.
Keywords: Confidence intervals, correlation and regression parameters, parametric statistics

Introducción

Existen dos predicciones en todos los campos de la ciencia. Uno que se trata de buscar diferencias entre grupos de datos, y la segunda es hacer predicción lo cual versa sobre la búsqueda de patrones repetitivos en escala espacio-temporal. La regresión y la correlación son dos técnicas estrechamente relacionadas. En forma más especifica el análisis de correlación y regresión comprende el estudio de los datos muestreales para saber cómo se relacionan entre sí dos o más variables en una población. El análisis de correlación produce un parámetro que resume el grado de asociación o correlación entre dos variables; y el análisis de regresión da lugar a una ecuación matemática que explica la dependencia de la variable independiente o predictora sobre la variable dependiente o respuesta y por tanto, sirve para predicción.
Se usa la correlación (Pearson, 1920) para estimar cuáles variables son potencialmente importantes, el interés radica básicamente en el grado de la relación, mientras que la regresión da lugar a una ecuación que describe, explica y predice dicha relación en términos matemáticos (Badii et al. 2007, 2009).
En la práctica se ha notado que cuando en un variable se presenta mayor intensidad, otra variable se afecta en alguna proporción. Ejemplos de estos podrán ser la altura y el peso de las personas, edad y vigor de rebrote en árboles, etc.; en el primer ejemplo se puede ver que esta relación es positiva, es decir, al aumentar una variable tiende aumentar la otra, en cambio, en el segundo ejemplo esta relación es negativa, pues al aumentar la edad, el vigor tiende a disminuir. Por lo tanto, los datos necesarios para análisis de regresión y correlación provienen de observaciones de variables relacionadas. Cuando solamente dos variables están involucradas en el análisis de regresión y correlación, se trata de la técnica regresión o correlación simple y cuando están implicadas tres o más variables, se trata de una regresión o correlación múltiple.

Correlación lineal simple y su medición

El coeficiente de correlación es el parámetro que mide el nivel de la relación entre las dos variables involucradas. Es la asociación de estas, más no de la dependencia, por lo tanto, no

olvidar que los cambios de los valores de una variable no son la causa del cambio de la otra. A continuación se describe e explica la noción de correlación simple y también el cálculo del coeficiente de la correlación (*r*), la medida usual del grado de correlación basándose en una muestra de *n* pares de valores de la variable bajo el estudio.

$$r = \sum(XY) - [(\sum X \sum Y)/n] / \{[\sum(X^2) - (\sum X)^2/n] [\sum(Y^2) - (\sum Y)^2/n]\}^{1/2}$$

Donde, r = coeficiente de correlación, y los demás anotaciones son relacionados con las variables X y Y, de hecho el numerados es la covarianza y el denominador es el raíz cuadrado de suma de cuadrados tanto de X $(\sum(X^2) - (\sum X)^2/n)$ como de Y $(\sum(Y^2) - (\sum Y)^2/n)$.
El nivel de precisión en la predicción depende de R^2 o coeficiente de determinación lo cual es el cuadrado de coeficiente de correlación y explique el porciento de la variabilidad en la variable repuesta o dependiente que se explica por la regresión. La ecuación de la regresión simple es: $\hat{Y} = \alpha \pm \beta X$ para el caso de la población, y $\hat{Y} = a \pm b X$ para la muestra. Las variables X y Y son las independientes y dependientes, respectivamente y los parámetros α y β (población) o a y b (la muestra) son la intersección y la pendiente de la regresión, respectivamente. La técnica de la regresión, tiene importantes condiciones o premisas para asegurar la validez del modelo, entre las que destacan la independencia de las observaciones sobre la variable independiente, la normalidad e independencia de los residuales y la homogeneidad de las varianzas, disponiéndose de alternativas no paramétricas para la correlación cuando estas supuestos no se cumplen.
Existen dos tipos de asociación o correlación. 1) Correlación positiva: cuando *r* está entre 0 y +1. Los valores de X y Y tienden a moverse en la misma dirección. 2) Correlación negativa: cuando r está entre -1 y 0. Entonces los valores de X y Y y tienden a moverse en dirección opuesta; cuando uno aumenta el otro tiende a disminuir y viceversa.

Medición de la significancia estadística

Una vez que hemos calculado la ecuación de la línea recta de regresión, el siguiente paso es analizar si la regresión en efecto es significativa y la podemos utilizar para predicción de los valores de $\hat{Y}$ en función del cambio unitario en los valores de X. Para ello debemos contrastar si la coeficiente de regresión entre ambas variables es distinta de cero o si el modelo de regresión es significativo en el sentido de demostrar si el análisis de nuestra variable endógena (Y) es válido a través de la influencia de la variable explicativa (X).
Aceptación o rechazo de las hipótesis del modelo en estudio, ya sea para coeficiente de regresión (Ho: b = 0 vs. Ha: $b \neq 0$), coeficiente de correlación (Ho: r = 0 vs. Ha: $r \neq 0$) y la intersección "a" es decir, altura de la línea a partir de la intersección con la ordenada (Ho: a = 0 vs. Ha: $a \neq 0$) determina la significancia de dichos parámetros.
Supongamos por un lado que el coeficiente de correlación lineal r, y el coeficiente de regresión (b) tienen valores muy altos, y por tanto solo por sus magnitudes parece indicar la existencia de una correlación y dependencia alta entre los valores de la muestra. Pero la magnitud alta de estos coeficientes muestreales entre ambas variables no necesariamente refleja la misma situación en la población. Para poder contrastar esta suposición, una vez que hemos estimado la recta de regresión y hemos obtenido las estimaciones de los parámetros del modelo, debemos comprobar si estas estimaciones del modelo son significativas de tal forma que la variable (X) es relevante para explicar la variable de respuesta (Y). Entonces debemos

contrastar si la "*r*" de correlación y la "*b*" de la recta de regresión poblacional son significativamente distintos de cero, de ahí tendríamos que, en efecto, existe una relación (*r*) y una dependencia (*b*) significativa entre ambas variables poblacionales.
Las hipótesis que se ponen a prueba indican que no existen diferencias en las medias de las poblaciones en los diferentes niveles del factor (Ho), es decir, que la relación y la dependencia no son significativas y que, por lo tanto, la variable independiente no tiene un efecto sobre la variable de respuesta. A continuación se demuestran las pares de hipótesis para cada parámetro de correlación y regresión. En el caso 1) el grado de correlación estimado por el coeficiente de correlación (*r*), 2) la tasa de la inclinación del pendiente estimado por el coeficiente de regresión (*b*), y 3) la altura de la línea o la intersección de la línea de regresión con el ordenado (*a*), las hipótesis son similares en el significado, es decir en cada uno de estos 3 casos, la hipótesis nula indica la ausencia de correlación, la independencia de variable "Y" sobre la variable "X", y la salida de la línea de regresión a partir del origen, es decir, (X = 0 & Y = 0). Estas declaraciones (nulidad de relación causal, o la hipótesis nula) manifestado por $r = 0$, $b = 0$, y $a = 0$, se contrastan con 1) existencia de correlación $r \neq 0$, $b \neq 0$ y $a \neq 0$, respectivamente.
Para poder comprobar las hipótesis planteadas se utilizan pruebas de comparación de estimadores, como la prueba de t-student (parámetro de comparación de las medias de las variables en estudio). En este caso, lo que se desea investigar es si los promedios de las muestras sometidas a diferentes métodos o tratamientos (distintos niveles de algún factor de variación), manifiestan diferencias significativas, es decir, si los intervalos de confianza de los valores paramétricos estimados no se traslapan.
Las formulas estadísticas de t-student para comprobar la significancia de las hipótesis planteadas para los coeficientes de regresión y correlación son los siguientes:

1. t-student para el coeficiente de regresión (*b)*:
$t_b = (b_C - 0) / EE_b$
$EE_b = [VE / SC_X]^{1/2}$
Donde, t_b = t calculado para "*b*", *b*c = valor calculado de "*b*", EEb = error estándar de "*b*", VE = varianza de error, SC_X = suma de cuadrados de X.

2. t-student para el coeficiente de correlación (*r*):
$t_r = (r_C - 0) / EE_r$
$EE_r = [(1 - r^2) / (n - 2)]^{1/2}$
Donde, tr = t calculado para "*r*", *r*c = valor calculado de "*r*", EEr = error estándar de "*r*".

3. t-student para la pendiente la intersección con la ordenada (*a*):
$t_a = (a_C - 0) / EE_a$
$EEa = [VE [(1/n) + (m^2_X / SC_X)]^{1/2}$
Donde, EEa = error estándar de "*a*", *a*c = valor calculado de "*a*", ta = t calculado para "*a*", VE = varianza de error, m_X = media de la variable X, SC_X = suma de cuadrados de X.

Ejemplo. Supongamos una población de quejas de seleccionados en forma aleatoria. Analizar la relación entre *X* y *Y*, donde, *Y* es el aumento en nivel de confianza en función de número de quejas (*X*) y determinar si existe una asociación significativa entre estas dos variables (Tabla 1).

Tabla 1. Datos de la población de quejas.	
Número de quejas (*X*)	Incremento en nivel de confianza (*Y*)
3	1.4
4	1.5
5	2.2
6	2.4
8	3.1
9	3.2
10	3.2
11	3.9
12	4.1
14	4.7
15	4.5
16	5.2
17	5.0

Estimar la asociación "*r*":

$r = 0.9866 \approx 99\ \%$

Existe una asociación de 99% entre la *X* y la *Y*.

Significancia de "*r*":

$t_r = (r_C - 0) / EE_r$

$EE_r = [(1 - r^2) / (n - 2)]^{1/2}$

$EE_r = [(1 - r^2) / (n - 2)]^{1/2} = [(1 - 0.987^2) / (13 - 2)]^{1/2} = 0.0486$

$t_r = (r_C - 0) / EE_r = (0.987 - 0) / 0.0486 = 19.9272$

A nivel de $\alpha = 0.05$ y grados de libertad igual a 11 (13 – 2), el valor calculado (19.9272) es mayor que tabulado (2.201) y por tanto, se rechaza Ho, es decir existe una asociación significativa y positiva entre la *X* y la *Y*.

En estimaciones de intervalo, debemos cuantificar el IC o intervalo de confianza lo cual es un intervalo dentro el cual se estima que se ubique el parámetro poblacional con una determinada probabilidad. Formalmente, un intervalo esta limitado por dos valores calculados a partir de datos de una muestra, y el valor desconocido ubicado dentro del intervalo es un parámetro poblacional. El IC se representa por "1 – α" y se denomina nivel de confianza. En estas circunstancias, α es el llamado error tipo I o nivel de significación, esto es la probabilidad de rechazar erróneamente una hipótesis nula cierta (Ostle, 1963, 1977, 1994).

El nivel de confianza y la amplitud del intervalo varían conjuntamente, de forma que un intervalo más amplio tendrá más probabilidad de acierto (mayor nivel de confianza), y viceversa. Los límites de intervalo de confianza son los límites inferior (LI) y superior (LS), y se determinan sumando y restando a la media de la muestra un cierto número Z de la tabla normal (dependiendo del nivel de error tipo I o α) multiplicado con el errores estándar de la media $\sigma_{\bar{X}}$ (Sokal & Rohlf, 1969, 2006).

Se declara que con 95% de confianza la media poblacional se encuentra entre los límites inferior y superior del rango de IC. Esto significa que solo hay un 5% (unilateral) o 2.5%

(bilateral) de probabilidad que el parámetro poblacional se ubique afuera del rango establecido por los límites inferior y superior de IC. Esta área traducida en los valores normalizados de la tabla normal o "Z" corresponde a 1.96. Por ejemplo el 95% de IC para la media o μ se calcula en base a IC = $\mu \pm \sigma_{\bar{X}}$ Zα/2, donde, $\sigma_{\bar{X}}$ = error estándar de la media, μ = la media muestreal y Zα/2 = valor de la tabla normal con una α igual a 5% (Montgomery et al., 2006). Para el ejemplo arriba, se va a determinar el intervalo para el coeficiente de la regresión, el coeficiente de la correlación y la intersección.

1. Coeficiente de Regresión (b): IC = $b \pm t_{\alpha(n-2)}$, IC = 0.27 ± 2.201 (0.141), IC varía de 0.24 a 0.30. Con un 95% de confianza (α = 0.05) el parámetro (b) se ubica entre los extremos de 0.24 y 0.30.

2. Coeficiente de asociación (r): IC = r ± $t_{\alpha(n-2)}$, IC = 0.9876 ± 2.201 (0.0495), IC varía de 0.88 a 1.10, Con un 95% de confianza (α = 0.05) el parámetro (r) flota entre los límites de 0.88 y 1.10.

3. Intersección con la ordenada(a): IC = a ± $t_{\alpha(n-2)}$, IC = 0.718 ± 2.201 (0.1555), IC varía de 0.37 a 1.05, Con un 95% de confianza (α = 0.05) el parámetro (a) flota entre los límites de 0.37 y 1.05.
A un nivel de X = 0 (cero quejas) es decir la ausencia de quejas por los ciudadanos, hay un nivel de confianza ciudadana igual a 0.718 y el nivel de confianza ciudadana crece con una tasa promedia de 0.27 por cada queja interpuesta por la ciudadana, y este crecimiento varia dentro el rango de 0.24 a 0.30 (IC para el pendiente o el coeficiente de la regresión). La ecuación predictiva es: $\hat{Y}$ = a + bX =0 .718 + 0.27X, donde, $\hat{Y}$ = el nivel predictivo de la confianza ciudadana, y X = número de quejas interpuestas por los ciudadanos. Con esta ecuación se puede predecir que por ejemplo el interponer 5 quejas tendrá el siguiente efecto en el nivel de confianza de ciudadana lo cual se estima por: $\hat{Y}$ = 0.718 + 0.27X = 0.718 + 0.27 (5) = 0.718 + 1.55 = = 2.268. Si observamos la Tabla 1, nos damos cuenta que el valor observado del nivel de confianza para el caso de X = 5 quejas es igual a 2.2. Por tanto, $\hat{Y}$ – Y = el valor predictivo (2.268) – el valor observado (2.2) = 0.68. Este valor de 0.68 se denomina el margen de error o error de estimación.

Conclusión

El uso de correlación y regresión en la investigación científica es esencial para poder predecir patrones. Un patrón es una tendencia establecido de tipo ascendente, constante o descendente lo cual el investigador persigue a través de la investigación. Hay que recordar que la ciencia es la búsqueda de la realidad espacio-temporal por medio del uso del método científico basado en los sentidos y inferencias lógicas. La ciencia se trata de apoyarnos a descubrir la realidad espacio-temporal acerca de la naturaleza y su funcionamiento. La búsqueda de los patrones es uno de los objetivos de la ciencia y la herramienta estadística de correlación y regresión nos permite a alcanzar a detectar los patrones escondidas en la naturaleza.

Referencias

Badii, M.H., J. Castillo, J. Rositas & G. Alarcón. 2007. Uso de un método de pronóstico en investigación. Pp. 137-155. In: M.H. Badii & J. Castillo (eds.). Técnicas Cuantitativas en la Investigación. UANL, Monterrey.

Baddi, M.H., J. Castillo, J. Landeros & K. Cortez. 2009. Papel de la estadística en la investigación científica. Pp. 1-43. In: M.H. Badii & J.Castillo (eds). Desarrollo Sustentable: Métodos, Aplicaciones y Perspectivas. UANL. Monterrey.

Montgomery, D.C., E. A. Peck & G.G. Vining. 2006. Introduction to Linear Regression Analysis. Wiley Interscience, USA.

Ostle, B. 1963. Statistics in research 2a ed. Iowa State Press, Ames, Iowa.

Ostle, B., 1994. Estadística Aplicada, Primera edición, Editorial Limusa, México, 447-452 p.

Ostle, B.1977. Estadística Aplicada Técnicas de la Estadística Moderna, Cuando y Donde Aplicarlas. Limusa, México, D.F.

Pearson, K. 1920. Notes on the history of correlation. Biometrica, 25-45.

Sokal, R..R. & F.J. Rohlf. 2006. Biometry. The Principles and Practice of Statistics in Biological Research. Freeman & Company. San Francisco.

Sokal R.R. & F.J. Rohlf. 2006. Introducción a la bioestadística. Reverté, S.A.