Análisis de Regresión: Correlación y su Aplicación en el Tamaño de la Banda "C" en Cromosomas Humanos

Lic. Gustavo Bastarrachea García

RESUMEN

Nuestro interés principal es determinar la relación que existe entre la banda "C" y la con tracción del cromosoma. Dado que, la diferencia longitudinal del cromosoma por la banda "C", nos permite observar dos regiones bastante marcadas en el cromosoma; una de coloración más intensa, llamada banda "C" y otra de coloración más tenue identificada como lq¾h. Estas dos regiones van a determinar diferentes tipos de cromatina: La banda "C" del tipo de Heterocromatina constitutiva y la región lq¾h del tipo Eucromatina. Ambas regiones son accesibles a mediciones.

Estudiaremos el caso en que la banda "C" depende linealmente de lq¾h. Si la relación de la banda "C" con lq¾h fuera perfecta, todos los puntos que se grafican usando la banda "C" como ordenada y lq¾h como abscisa quedarían sobre una línea recta, cuya ecuación general sería.

$$H = a + b\,E \qquad (1)$$

donde a es la ordenada en el origen, es decir, el valor de H cuando lq¾h = 0; b es la pendiente de la recta, esto es, la inclinación de la recta sobre el eje de abscisas.

Si b fuera cero, H (la banda "C") no dependería de "E" (región eucromática lq¾h), sería una constante, H = a, igual para cualquier valor de "E". La recta con b = 0 es paralela al eje de abscisas cuya altura sobre dicho eje siempre es igual a a.
Si la banda "C" aumenta cuando lq¾h aumenta, b es positiva, desde luego, que al aumentar lq¾h disminuye la banda "C", entonces b es negativa.

El estudio considera las mediciones de los cromosomas 1, 9 y 16 de una muestra n=20 individuos japoneses, 15 mujeres y 5 varones.

Estas mediciones la representamos por (E_i H_i) donde E_i y H_i son los valores de E y H en el individuo i-esimo, o en la observación i-esima (i=1, 2, 3,.....n). En la práctica no obtenemos una relación perfecta entre las variables H_i, E_i . Más bien, obtenemos valores de H_i que fluctúan alrededor de una recta. La fluctuación puede deberse a muchas causas; entre otras, podemos considerar:

a) Las mediciones realizadas son imperfectas. Nuestros instrumentos son imprecisos.
b) La variable "H" puede depender de varias otras variables, a partir de la que estamos considerando. Por ejemplo del estado fisiológico de la célula.
c) El comportamiento del material biológico es variable por naturaleza.

Para describir el comportamiento de la variable "H" como función de la "E", agregamos a la ecuación de la recta un elemento aleatorio m_i , al que le llamaremos el error, que suponemos que está distribuido según una normal de media cero y de varianza constante s^2_u

$$H_i = a + b\,E_i + m_i \qquad (2)$$

La ecuación (1) es un modelo matemático que intenta describir a la H_i como función lineal de E_i . La expresión (2) es un modelo estadístico. Con la inclusión de m_i como componenete de la H_i, hemos hecho de ésta, una variable aleatoria cuya media es a + b E_i y cuya varianza es s^2_u.
En este planteamiento hay varios conceptos importantes que conviene destacar.

i) Para cada E_i la H_i correspondiente se distribuye normalmente, $H_i \sim N(a + b\,E_i\,,\ s^2_u)$. Pueden haber varias observaciones H_i para una misma E_i desde luego. Podemos visualizar cada H_i como un valor que se obtuvo conceptualmente, de calcular a + b E_i correspondientes, y al que se le

agrega un valor m_i que se obtiene aleatoriamente de la $N(0, s^2_u)$. El valor de m_i como tiene media cero, puede ser positivo o negativo. Si es positivo, H_i quedará por encima de la recta $a + b E_i$. Si es negativo, H_i quedará por debajo.

ii) La varianza de H_i dada E_i, que es varianza s^2_u, es constante, esto es, la misma para cualquier valor de E_i. A este supuesto se le conoce como la propiedad de "homosedasticidad" o "varianza homogénea" de H_i dada E_i. Esto es que el error m_i sea independiente de la E_i.

iii) Como consecuencia, para dos observaciones distintas, i, j, m_i y m_j son independientes.

iv) Las variables E_i se miden sin error. La recta (1) es la ecuación de las medias de H_i en función de la E_i. Se denomina la recta de regresión de la población de donde se tomó las muestras de observaciones.

Los parámetros de esta recta, que queremos estimar a partir de la muestra, son α y β. Este último se denomina el coeficiente de regresión de la población. El término "regresión" proviene de 1889, cuando Sir Francis Galton estudió las leyes de la herencia. Karl Pearson verificó este comportamiento, específicamente de la estatura, en 1903 en Inglaterra.

La regresión tiene múltiples aplicaciones. Se puede usar para verificar si, en efecto, la H_i depende de la E_i, esto es, si $\beta \neq 0$.

Para estimar los parámetros α y β de la regresión propuesta, el método generalmente utilizado consiste en encontrar los valores de α y β, estimados a partir de la muestra de n parejas (E_i, H_i); que hacen mínima $\sum m^2_i$. Esto es, se buscan valores de α y β que minimicen la expresión:

$$\sum_{i=1}^{n} (H_i - \alpha - \beta E_i)^2 \qquad (3)$$

donde H y E son las medias de las variables, longitud de la banda "C" y la regresión euromática (lq-h) correspondiente, en la muestra.

La recta de regresión estimada es:

$$H_i = a + b E \qquad (6)$$

Si llamamos E_i a las medidas de tipo eucromatina (lq¾h) de los individuos i-esimo (i =1, 2, ..., n), H_i la longitud de la banda "C" en el mismo cromosoma, ajustaremos una regresión lineal simple del tipo dado en (6). De la muestra n = 20 tenemos los siguientes resultados:

$n = 20; \quad \sum H_i E_i = 87.1825$

$\sum E_i = 57.97$

$\sum H_i = 29.72$

$\sum E^2_i = 172.1043 \quad \sum H^2_i = 45.2110$

$E_i = 2.8985 \quad H_i = 1.486$

La recta de regresión que hemos ajustado es, entonces, según (6)

$$H_i = 0.7475 + 0.2548 E \qquad (7)$$

es importante ahora analizar la calidad de ajuste de regresión que hemos hecho.

Podemos empezar calculando los valores de H que se obtiene, sustituyendo los valores observados de E_i en (7).

Los resultados de las mediciones (ver cuadro N°1) se discuten más adelante.

Calculemos ahora S^2_u, la varianza de H_i dado E_i. El cómputo se hace con la expresión:

$$S^2_u = 1/(n-2) \sum_{i=1}^{n} (H_i - H_i)^2 \qquad (8)$$

$S^2_u = 0.0435$

La varianza de β se calcula ahora de la expresión:

$$S^2(\beta) = \frac{S^2_u}{\sum_{i=1}^{n} (E_i - E)^2} \qquad (10)$$

que para el experimento que tenemos resulta,

$$S^2(\beta) = 0.1033$$

Podemos ahora construir una prueba de la hipótesis Ho : $\beta = 0$
usamos la estadística:

$$t = \frac{\beta - \beta / H_o}{S(\beta)} \qquad (11)$$

que sigue la distribución t de "student" con n-2 gra-

dos de libertad, porque ajustamos dos parámetros en la regresión: a y b, tenemos entonces:

$$t_c = \frac{0.2548 - 0}{0.1033} = 2.467$$

un valor cuya significación con $t_{\alpha,(n-2)} = t_{0.05\,(18)} = 2.101$ es decir, con una probabilidad de $P_k = 1 - \alpha = 0.95$. Rechazamos por lo tanto, H_o. Concluimos que tenemos una regresión poderosa de H sobre E.

Para estimar la varianza de H_i, podemos escribir la expresión (6) únicamente en términos de β,

$$H_i = \alpha + \beta E_i = H_i - \beta E + \beta E_i \qquad (12)$$

$$\therefore H = H_i + \beta (E_i - E)$$

Ahora, la varianza de Hi resulta:

$$\text{var}(H_i) = S^2_u \, 1/n + \frac{(E_i - E)^2}{\sum_{i=1}^{n} (E_i - E)^2} \qquad (14)$$

A partir de (14) se puede evaluar las varianzas de H_i; por ejemplo la varianza de H_i, que es 1.31 y corresponde a $E_i = 2.21$, será:

$$\text{var}(H_i) = 0.0072$$

que puede considerarse pequeña debido a la unidad de medida utilizada en el experimento, como se puede apreciar en el cuadro N°1 en sus medidas puntuales.

El coeficiente de correlación, nos permite obtener un coeficiente que condense la banda del ajuste de regresión en un número.
Se trata de medir en otras palabras, el grado de asociación lineal entre las variables H_i, E_i. El número que lo mide es el coeficiente de correlación, r, cuya expresión puede darse en términos del coeficiente de regresión mediante la relación,

$$\rho = \beta \cdot \frac{\sigma E}{\sigma H} \qquad (15)$$

En el experimento, el valor del coeficiente de correlación entre la banda "C" y la región eucromática (lq—h) resulta:

$$\rho = 0.5029$$

El coeficiente de correlación varía entre -1 y +1, si $\rho = 0$, E no está correlacionada con H (y entonces, también, $\beta = 0$). El valor de $\rho = 0.5029$ indica que existe una mediana correlación y que los puntos de coordenadas (E_i H_i) están cercanos a la recta en (6).

Esto era de esperarse dado el valor de t en la prueba de hipótesis $H_O : \beta = 0$. En vista de que si $\beta = 0$, también $\rho = 0$, una prueba de ambos igual a cero puede hacerse con el coeficiente de correlación.

En efecto, la expresión para S^2_u

$$S^2_u = 1/n\text{-}2 \;(1 - \rho^2) \sum_{i=1}^{n} (H_i - H)^2 \qquad (18)$$

De aquí puede inferirse que $\rho^2 = \sum_{i=1}^{n} (H_i - H)^2$ es la proporción de $\sum_{i=1}^{n} (H_i - H)^2$ que se explica por el ajuste de regresión. La parte que puede explicarse por la regresión de H_i sobre E_i, la mide el cuadrado del coeficiente de correlación, que por ello se denomina el coeficiente de determinación.

En el estudio con $\rho = 0.5029$, esto es el 25.29% de la suma de cuadrados de H_i se explica por el ajuste de regresión. Por consecuencia la varianza restante, S^2_u, es grande.

Con el resultado (18) se puede construir una prueba de $H_o : \rho = 0$ (ó $\beta = 0$, es lo mismo). Se calcula una t que es:

$$t_c = p \sqrt{\frac{n - 2}{1 - p^2}} \qquad (19)$$

$t_c = 2.468$

Al comparar este resultado con la t de las tablas con n - 2 grados de libertad y el nivel de significación 0.05 el valor de t resulta 2.101 con 18 grados de libertad, es decir, con una probabilidad de 0.95,

rechazamos la hipótesis de que $\rho = 0$ y concluimos que el grado de asociación entre las variables es significativo.

El simple valor de ρ , aunque sea alto, no necesariamente implica que la asociación entre H_i y E_i es rigurosa, por que depende del número de observaciones en que se base las mediciones realizadas, (o más precisamente depende del grado de libertad para estimar ρ) y del nivel de significación que se use para trabajar.

ANÁLISIS DE LOS RESULTADOS:

El cuadro N°1 presenta el análisis de regresión y correlación obtenidos entre la longitud de la banda "C" y el tamaño de la regresión eucromática (lq—h), aproximándonos al análisis de Balicek P., Zizka J. y Skalska H., Hum Genet (1977).

En consecuencia la metodología presentada, tendrá los resultados que persigue el Instituto de Biomatemática en apoyo a los investigadores de la Facultad de Ciencias Biológicas, quien proporcionó los datos.

El diagrama de regresión entre la inclinación (β) del tamaño de la banda "C" sobre (lq—h) y su longitud promedio presentada en el Cuadro N°1, donde los parámetros obtenidos son: $a = 0.089$ y $b = 0.1595$ y el coeficiente de correlación $r = 0.4661$ para $n = 9$ y $\alpha \leq 0.05$, no es grande como se hubiera deseado, pero confirma el análisis y las consideraciones observadas al inicio del estudio.

Cuadro No. 1

REGRESIÓN Y CORRELACIÓN OBTENIDA ENTRE LA LONGITUD DE BANDA "C" Y EL TAMAÑO DE LA REGIÓN EUCROMÁTICA (lq— h)

BANDA "C" (2)	Tamaño de Muestra (3)	Promedio (X) $\pm S_H$	Ángulo de Inclinación (b) (1)	Intercepto (a)	b/ x	Coefic. de Correlac. ρ
Largo (h^+_1)	20	1.59 ± 0.32	0.051	1.283	0.032	0.15
Medio (h_1)	20	1.52 ± 0.13	0.041	1.323	0.027	0.22
Pequeño (h^-_1)	20	1.47 ± 0.23	0.048	1.290	0.032	0.25
Largo (h^+_1)	20	1.56 ± 0.07	0.263	0.73	0.169	0.40
Medio (h_1)	20	1.50 ± 0.05	0.272	0.70	0.181	0.17
Pequeño (h^-_1)	20	1.40 ± 0.06	0.262	0.72	0.187	0.28
Largo (h^+_1)	20	0.88 ± 0.05	0.041	0.806	0.047	0.14
Medio (h_1)	20	0.87 ± 0.02	0.040	0.808	0.045	0.17
Pequeño (h^-_1)	20	0.86 ± 0.03	0.045	0.794	0.052	0.12

1) Valores al nivel de significación 0.01.
2) Corresponde a la ubicación de la banda "C" dentro del conjunto de medidas.
3) FUENTE: Muestra tomada por Luis Alberto Rodríguez en el Laboratorio de Genética Humana, U.N.M.S.M. - Facultad de Ciencias Biológicas.

REFERENCIAS

-BERNAR OSTLE Estadística Aplicada – Centro Regional de ayuda técnica México, 1965.

-JOHNE. FREUND Estadística Matemática

-R. E. WALPOLE conAplicaciones México, 1980.

-FREDERICK E.CROXTON

-DUCLLEY J. COWDE Fondo de la Cultura Económica Buenos Aires, 1965.