CORRELACIÓN Y REGRESIÓN LINEAL MÚLTIPLE

Por Tevni Grajales G.

Con frecuencia hemos observado la relación que existe entre una variable y otra (correlación bivarible) lo cual nos permite, en algunos casos, predecir los valores de una variable a partir de los valores observados en la otra. Por ejemplo: se ha encontrado que las calificaciones que un estudiante obtiene en una prueba de ingreso a la universidad se correlacionan con las calificaciones que el alumno obtiene en su programa académico; siendo asi se podría intentar predecir la calificación final del estudiante.

Pero el mundo de la educación es muy complejo y difícilmente podemos atribuir a una sola variable los resultados en otra; la realidad nos obliga a reconocer que para predecir con mayor precisión las calificaciones finales del estudiante, es necesario observar e integrar en la predicción otras variables que también puedan estar relacionadas.

Un esfuerzo de este tipo implica la observación de más de dos variables al mismo tiempo y en el caso de una observación correlacional, requiere de una procedimiento que permita pesar el grado de impacto que cada una de las variables observadas puede tener sobre los resultados de la predicción. Por ejemplo, sabemos que el precio de la colegiatura en una institución está determinado por varias variables: costo de los servicios públicos, renta del local, gastos indirectos, tamaño y características del personal que labora en la institución, etc. También sabemos que estas variables antes mencionadas no tienen la misma importancia al momento de determinar el costo de la colegiatura, por lo que diríamos que hay que ponderar el impacto que cada una tendría sobre el costo de la colegiatura.

El procedimiento analítico que nos permite determinar cuánto de la variación en la variable observada está asociado con la variación del conjunto de variables que pretenden predecirla se denomina: Correlación Múltiple.

- El objetivo de la técnica

En el procedimiento de correlación múltiple se procura construir la mejor combinación del peso que cada variable simple aporta en la medición de la variable que se observa. Y esta mejor combinación sin duda tendrá una mayor correlación con la variable observada que la correlación que pueda tener cualquiera de las variables simples de manera independiente.

¿Cómo se puede lograr esto? Por medio de técnicas avanzadas de matemáticas para cálculo y matrices algebraicas lo cual supera los propósitos de este trabajo. Lo que pretendemos es comprender el significado de los resultados que se obtienen.

Esta técnica supone que existen más de dos variables correlacionadas y que es posible determinar la forma como se comportan las diversas correlaciones a nivel bivariable a fin de conformar una correlación combinada o total.

Como se sabe, cuando dos variables se relacionan de manera significativa esto se puede representar sobre un eje de coordenadas en la que cada variable se representa sobre cada uno de los ejes. Además esa relación puede ser representada por una línea imaginaria que cruzara de manera equidistante entre todos los puntos que representan la relación. Esta línea puede ser determinada por medio de una ecuación lineal simple que permita predecir el valor de una variable observada (Variable criterio) a partir de una variable predictora. Con este fin se determina un coeficiente b el cual explica la forma como cambia la variable criterio por cada unidad que cambia en la variable predictora. Por ejemplo: sea la ecuación de una línea $Y = A + BX$ en la que A es el valor que asume la variable Y cuando X es igual a 0 (en otras palabras, valor que tiene Y cuando la línea que representa la ecuación cruza el eje de las coordenadas) y es llamado intercepto; B por su parte representa la caída o pendiente de la recta y su valor representa la forma como los valores de Y pueden variar por cada unidad de variación en X. El valor B mencionado en el ejemplo se conoce como **coeficiente b**.

En el caso cuando se consideran más de dos variables, para cada una de las variables predictoras corresponde un **coeficiente b**, el cual a su vez se puede estandarizar conduciendo a calcular otro coeficiente denominado "**coeficiente beta**" cuyo valor está en función de dos cosas: (1)las correlaciones de las variables predictoras individuales con la variable criterio y (2) las correlaciones que existen entre las variables predictoras entre si.

Tabla No.1

caso	Valor de la variable criterio (z)	Valor compuesto derivado de las predictoras	=	($beta_{1)}$ (velocidad de lectura)	+	($Beta_{2)}$ (aptitud verbal)	+	($Beta_{3)}$ (aptitud musical)
1	1.50	1.23	=	(.27)(1.72)	+	(.51)(.98)	+	(-.22)(-1.21)
2	-.78	-.49	=	(.27)(-.97)	+	(.51)(-.15)	+	(-.21)(.33)

3	.	.		.		.		.

R es la correlación entre los valores de la segunda y tercera columna. Ejemplo: (1.50 y 1.23) (-.78 y -.48)
Los **coeficientes beta** determinados de manera que maximicen la correlación representada por R son los que aparecen en el primer paréntesis de cada una de las columnas 5,7 y 9.Ejemplo: (.27, .51, y -.22)

Se espera que una variable predictora en alta correlación con la variable criterio tendrá un mayor peso (coeficiente beta) asociado con ella y que una variable que tenga una correlación menor presenta un peso menor (coeficiente beta menor).

Pero también hay que considerar el grado de relación que pueda existir entre las variables predictoras , pues si ellas tienen alta correlacionadas entre si, se supone que se trata de lo mismo y no aportarán mucho más que lo que haría una sola de ellas al considerar su impacto conjunto (peso) en la variable criterio. Ha este fenómeno se le llama multicolinariedad.

En cambio si las mismas variables que están altamente correlacionadas con la variable criterio tienen una muy baja correlación entre si, por ser diferentes, cada una contribuirá desde perspectivas diferentes; y su aporte combinado será mayor al que sería si estuviesen altamente correlacionadas entre si.

En conclusión, para determinar los coeficientes beta, no sólo debemos tener en cuenta la correlación entre las variables predictoras y la variable criterio sino también la correlación que puede existir entre las variables predictoras entre si.

Una vez determinados los coeficientes beta (peso de cada variable predictora en la criterio) es posible correlacionar el conjunto de valores compuesto derivados de una aplicación de los coeficientes beta y los valores de la variable criterio. Esta correlación compuesta, se denomina Coeficiente de Correlación Múltiple y se representa por una **R** (r mayúscula) para distinguirla de los coeficientes **r** entre dos variables aleatorias.

Interpretación de la correlación múltiple

Como podemos suponer, el coeficiente beta puede ayudarnos a comprender la importancia de una variable en la forma como se comporta la variable criterio (principal); que a mayor valor beta mayor impacto tiene en la forma como varia el valor de la variable criterio o viceversa. Pero si deseamos conocer la contribución (importancia) relativa de las variables predictoras en la variabilidad criterio, lo podemos determinar elevando al cuadrado los respectivos coeficientes beta. Esto no dice nada respecto a la contribución absoluta de cada variable predictora , sólo presenta su importancia relativa. Por ejemplo: Siendo el coeficiente beta de la variable edad .53 mientras que el coeficiente beta de la variable nivel académico es de .22 al correlacionarlas con la variable criterio gradación del lente, se puede elevar al cuadrado ambos coeficientes obteniéndose para edad una beta cuadrado de.2809 y para nivel académico .0484 lo que significa que la edad contribuye cinco veces más a la variabilidad de la variable gradación que lo que contribuye la variable nivel académico, dentro de una investigación en particular.

De la manera como en las correlaciones simples, al elevar al cuadrado el coeficiente de correlación entre dos variables, se determina la proporción de la varianza en una de las variables atribuible o predecible a partir de otra, el valor R elevado al cuadrado R^2 representa la proporción de la varianza en la variable criterio que puede ser predicha a partir la varianza conjunta de las variables predictoras.

En fin, el concepto de regresión múltiple es una ampliación del concepto de regresión simple entre dos variables. En lugar de usarse una variable predictora para estimar los valores en una variable criterio, se hace uso de varias variables predictoras. Un ejemplo puede ser predecir el promedio de las calificaciones en la universidad a partir de variables predictoras como promedio en el colegio secundario, puntaje obtenido en una prueba estandarizada de aptitud, ingreso familiar, resultado obtenido en algunos exámenes de ingreso a la universidad, etc. Esto conduce a una ecuación cuya forma general es:

$$y'=a+b1\mathbf{X}1+b2\mathbf{X}2+b3\mathbf{X}3+...+bk\mathbf{X}k.$$

donde y´ es el valor predicho de la variable criterio y los valores de a y los coeficientes b serán determinados a partir de los datos ofrecidos por la muestra. Esta ecuación no representa una línea como en el caso de la regresión simple sino que representa diversos planos de un espacio multi-dimensional. Los valores de a y b representan la mejor solución en valores a fin de que la suma de los cuadrados de la diferencia entre la y observada y la y predicha -S $(y-y')^{2-}$ sea lo mínimo. Los componentes de la ecuación deberán ser estandarizados en puntuaciones z para poder determinar la importancia relativa de las diversas variables predictoras como aparece en la Tabla No1.

Ejemplo 1

En una clase de métodos de investigación constituida por alumnos del pregrado en la UM se procuró conocer de qué manera los resultados obtenidos en pruebas de redacción, de búsqueda bibiliográfica, asi como el género, la edad y la carrera que cursa el alumno pueden contribuir a predecir la calificación que pueda obtener el alumno en una prueba parcial a final de la primera parte del curso. Con el fin de dar respuesta a la pregunta, se usó una regresión lineal múltiple con los siguientes resultados:

R múltiple .63377
R^2 .40167
R^2 ajustada .33801

Columna 1	Columna 2	Columna 3	Columna 4	Columna 5	
Variable	**Coeficiente b**	**Coeficiente beta**	**Valor T**	**Significatividad de T**	
Examen de redacción	.540472	.541046	4.615	.0000	
Examen de bibliografía	-.768937	-.101261	-.859	.3949	
Edad	-.624305	-.273969	-1.720	.0920	
Género	3.310275	.221553	1.824	.0746	
Carrera	-.481764	-.270745	-1.763	.0844	
Constante (a)	51.166291		3.876	.0003	

Podemos notar que los valores de la columna 2 se refieren a los valores absolutos de b en el caso de pretender predecir los resultados de en el examen a partir de estas variables. Pero si deseamos determinar cuál variable tiene mayor impacto en los posibles valores de la prueba parcial, podemos notar que en la columna 3 aparece un coeficiente beta de .541046 para el examen de redacción lo que nos dice que ésta calificación no sólo está relacionada sino que aporta mucho más que las otras a la variabilidad de la prueba parcial. Para hacer una comparación eleve al cuadrado los respectivos coeficiente beta (v.g. .2916; .0102; .0750; .0490; .0733). Una observación adicional nos permite notar que el valor T para cada una de las variables aparece en la columna 4 y está relacionado con un nivel de significatividad que aparece en la columna 5. En el caso de que el nivel de significatividad de la columna 5 sea <.05 se puede considerar que la variable puede formar parte de la ecuación predictora pero en la medida como no sea significativa, su poder de predecir nulo. Esto tiene que ver con los criterios a seguir al momento de determinar las variables que puedan constituir una ecuación predictora.

Los paquetes estadísticos ofrecen diversos procesos para determinar la mejor ecuación. Entre los más conocidos está stepwise, backward, forward, remove. En el anexo A se presentan los resultados de una regresión lineal siguiendo el procedimiento enter y luego la misma según el procedimiento stepwise. Este último muestra que únicamente la variable Examen de redacción es apropiada para una ecuación predictora.

Ejemplo No. 2

A continuación se presentan los resultados de una correlación lineal múltiple entre las variables predictoras: categoría de trabajo, edad y nivel educativo con el fin de determinar los valores de la variable criterio: salario del empleado.

R múltiple .82822
R^2 .68596
R^2 ajustada .68395

Columna 1	Columna 2	Columna 3	Columna 4	Columna 5	
Variable	**Coeficiente b**	**Coeficiente beta**	**Valor T**	**Significatividad de T**	
Nivel educativo	891.147917	.376387	12.136	.0000	
Edad	5.065970	.008743	.324	.7462	

Categoría de trabajo	2796.581373	.575457	19.262	.0000	
Constante	-4201.895169		-3.499	.0005	

Al comparar los resultados de este segundo ejemplo notamos que estas variables ejercen un mayor impacto en la variabilidad de la variable criterio, (compare las R múltiples y las R^2) las R ajustadas nos permiten observar la conducta a nivel poblacional.

En el segundo ejemplo notamos que, en esta muestra, el nivel educativo (beta .376387) y la categoría de trabajo (beta .575457) son los mejores predictores del salario. Es evidente que ambas variables pueden ser incluídas en una ecuación predictora dado el nivel de significatividad de las respectivas T.

Actividad:

1. Escriba la ecuación de regresión que corresponde a los ejemplos anteriores.

2. Según la información que aparece en el Anexo A stepwise, determine el valor correspondiente a la calificación de un estudiante que obtuvo 60 puntos en la prueba de redacción según el ejemplo No.1.

Altius
tgrajales.net