Modelación estadística: La regresión lineal múltiple (parte 2)

Gabriel Cavada Ch.[1,2]
[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*
[2]*Facultad de Medicina, Universidad de los Andes.*

Statistical modeling: Multiple linear regression (second part)

El algoritmo de la estimación del modelo de regresión lineal múltiple es utilizado para comparar promedios en dos o más grupos, método que es conocido como análisis de la varianza, ANOVA. También sirve para estimar los promedios condicionados a distintos grupos ajustando por variables continuas, método conocido como análisis de covarianza o ANCOVA. Ambos métodos suponen que las unidades de muestreo son independientes, es decir cada sujeto es medido sólo una vez; si esta situación no se diera estaríamos en presencia de diseños jerárquicos o de medidas repetidas en cuyo caso la exposición de los métodos ANOVA y ANCOVA que se tratan en este artículo no son aplicables.

Antes de explicitar los modelos e ilustrarlos, es necesario definir lo que se entiende por variables "dummys" o Indicatrices; en efecto:

Si en el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_p X_p$$

es necesario incorporar variables explicativas que sean de naturaleza nominal (no numéricas), tales como sexo, raza, tratamiento u otras, debemos introducir el concepto de variable Indicatriz o variable "dummy". Estas variables son generadas a partir de una variable categórica que está medida en "k" niveles, de modo tal que esta producirá tantas variables "dummy" como niveles tenga; así, cada "dummy" indicará la pertenencia de la unidad de observación, en forma excluyente a cada nivel de la variable original. Es decir,

si se tiene X una variable categórica medida en r niveles, entonces se definen r variables "dummys" que indican en forma EXCLUYENTE la pertenencia de una unidad de observación a un determinado nivel de la variable. Así:

$$d_1 = \begin{cases} 1 & \text{si el sujeto está en el nivel } X_1 \\ 0 & \text{si el sujeto no está en el nivel } X_1 \end{cases}$$

$$d_2 = \begin{cases} 1 & \text{si el sujeto está en el nivel } X_2 \\ 0 & \text{si el sujeto no está en el nivel } X_2 \end{cases}$$

.
.
.

$$d_r = \begin{cases} 1 & \text{si el sujeto está en el nivel } X_r \\ 0 & \text{si el sujeto no está en el nivel } X_r \end{cases}$$

Ejemplo: Supongamos que se registra la variable Nivel Educacional medida en tres niveles:

$$\textit{Nivel Educacional} = \begin{cases} 1 & \text{nivel bajo} \\ 2 & \text{nivel medio} \\ 3 & \text{nivel alto} \end{cases}$$

Y se registra información de 6 sujetos, que se muestran a continuación:

Sujeto	Nivel educacional
1	1
2	1
3	2
4	2
5	3
6	3

Rincón de la Bioestadística

Al crear las respectivas variables "dummys" la tabla con la información se expande como sigue:

Sujeto	Nivel educacional	Nivel_ edu1	Nivel_ edu2	Nivel_ edu3
1	1	1	0	0
2	1	1	0	0
3	2	0	1	0
4	2	0	1	0
5	3	0	0	1
6	3	0	0	1

Debe notarse que se han creado tres variables "dummys" y cada una de ellas toma el valor 1 de acuerdo con el nivel de la variable original.

Una vez creadas estas variables, el promedio de una respuesta continua "Y", puede ser comparado a través del siguiente lineal múltiple:

$$Y = \underbrace{\beta_0}_{\text{Respuesta promedio en el grupo de referencia}} + \underbrace{\beta_2 \cdot Nivel\_edu_2}_{\text{Respuesta promedio en el grupo 3}} + \underbrace{\beta_3 \cdot Nivel\_edu_3}_{\text{Respuesta promedio en el grupo 2}}$$

Si el sujeto es de Nivel educacional 1, implica que el Nivel educacional 2 y 3 son 0 y el modelo se reduce a:

$$Y = \beta_0$$

que es la respuesta promedio en el nivel educacional 1.

Si el sujeto está en el Nivel educacional 2, implica que el Nivel educacional 2 es 1 y el nivel 3 es 0, con que el modelo se reduce a:

$$Y = \beta_0 + \beta_2$$

Lo que representa la respuesta promedio en el nivel educacional 2; además se observa que el valor de β_2 es la diferencia de la respuesta en el nivel 2.

Con idéntico razonamiento se encuentra que la respuesta promedio en el nivel educacional 3 es:

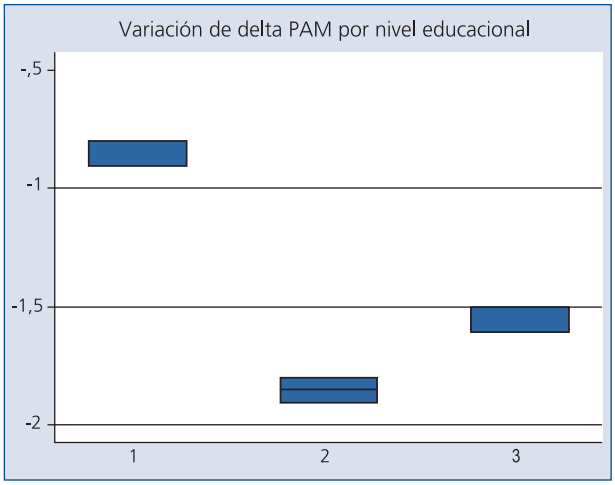
$$Y = \beta_0 + \beta_3$$

Ejemplo: Un médico sospecha que la efectividad de un tratamiento hipotensor, debido a su complejidad, depende del nivel educacional del paciente y de su edad. La principal respuesta es el cambio de PAM (mm Hg) al cabo de un mes de tratamiento.

El promedio y desviación estándar del cambio en la PAM por nivel educacional se muestran a continuación:

Nivel educacional	Promedio	DE	N
1	-0,86	0,05	10
2	-1,85	0,05	10
3	-1,56	0,05	10
Total	-1,42	0,43	30

El gráfico de esta situación se ve a continuación:



Al proponer el modelo:

$$deltaPAM = \beta_0 + \beta_2 \cdot Nivel\_edu_2 + \beta_3 \cdot Nivel\_edu_3$$

la estimación del mismo se muestra en la siguiente tabla:

deltaPAM	Coeficiente	Valor p	Intervalo de Confianza 95%	
Nivel_edu2	-0,99	0,0000	-1,04	-0,94
Nivel_edu3	-0,70	0,0000	-0,75	-0,65
constante	-0,86	0,0000	-0,89	-0,83

$$deltaPAM = -0{,}86 - 0{,}99 \cdot Nivel\_edu_2 - 0{,}7 \cdot Nivel\_edu_3$$

Promedio de deltaPAM en nivel educacional 1= -0,86

Promedio de deltaPAM en nivel educacional 3 = -0,86-0,7 = -1,56

Promedio de deltaPAM en nivel educacional 2 = -0,86-0,99 = -1,85

También sería de interés saber si la edad del paciente está asociada al cambio en la PAM; para ello se propone el modelo:

$$Y = \beta_0 + \beta_2 \cdot Nivel\_edu_2 + \beta_3 \cdot Nivel\_edu_3 + \beta_4 \cdot edad$$

Esta expresión contiene 3 ecuaciones lineales simples a saber:

- Si el nivel educacional es el de referencia las variables "dummys" $Nivel\_edu_2$ y $Nivel\_edu_3$ toman el valor cero y la ecuación se reduce a:

$$Y = \beta_0 + \beta_4 \cdot edad$$

- Y si el nivel educacional es el nivel educacional 2, $Nivel\_edu_2 = 1$ y $Nivel\_edu_3 = 0$, la ecuación se reduce a:

$$Y = \beta_0 + \beta_2 + \beta_4 \cdot edad$$

Con lo que el intercepto de esta recta es $\beta_0 + \beta_2$.

- Por último, si el nivel educacional es el nivel educacional 3, $Nivel\_edu_2 = 0$ y $Nivel\_edu_3 = 1$, la ecuación es:

$$Y = \beta_0 + \beta_3 + \beta_4 \cdot edad$$

Con lo que el intercepto de esta recta es $\beta_0 + \beta_3$.

Como puede observarse se trata de tres rectas que tienen igual pendiente pero distintos interceptos. La estimación de este modelo se muestra en la siguiente tabla:

deltaPAM	Coeficiente	Valor p	Intervalo de Confianza 95%	
Nivel_edu2	-1,00	0,0000	-1,03	-0,97
Nivel_edu3	-0,70	0,0000	-0,73	-0,67
edad	0,01	0,0000	0,01	0,02
constante	-1,71	0,0000	-1,99	-1,43

Considerando las interpretaciones anteriores, el gráfico de las ecuaciones para los distintos niveles educacionales que muestran el cambio de deltaPAM por edad es el siguiente: