Fundamentos en Humanidades
Universidad Nacional de San Luis – Argentina
Año XI – Número II (22/2010) 89/99 pp.

La utilización de variables indicadoras en un Modelo de Regresión Múltiple

Indicator variables in a multiple-regression model

Fabricio Penna
Universidad Nacional de San Luis
fpenna@unsl.edu.ar

(Recibido: 16/09/09 – Aceptado: 15/02/11)

Resumen

El presente trabajo es sólo un acercamiento de modelación estadística a las Ciencias Humanas, particularmente su aplicación a la Psicología. Se hace especial referencia a la aplicación de variables indicadoras (o dummy) para una mejor aplicación de la regresión múltiple en esa área de desarrollo.

Abstract

This paper is only an approach of statistical modelation to Human Sciences, particularly its application to Psychology. It focuses on indicator (or dummy) variables for a better application of multiple-regression model in that development area.

Palabras clave

modelación estadística - variables dummy - regresión múltiple - psicología

Key words

statistical modelation - dummy variables - multiple-regression - psychology

"*The path by which we rise to knowledge must be made smooth and beaten in its lower steps, and often ascended and descended, before we can scale our way to any eminence, much less climb to the summit*"
John Herschel ([1830] 1987).

Motivación

La modelación estadística, con el tiempo y el avance tecnológico, se ha ampliado y en la actualidad se utiliza para resolver problemas en diferentes áreas científicas, particularmente en algunas disciplinas de las Ciencias Humanas, constituyéndose como base de la formulación teórica.

Esta rama de la estadística se considera como un área de estudio en la que convergen aspectos tanto teóricos como metodológicos.

Nelder y Wedderburn (1972) realizaron generalizaciones para modelos lineales donde incluyeron, entre otros, modelos de regresión lineal para variables continuas.

Por otro lado, cuando hablamos de un modelo de regresión múltiple, nos estamos enfrentando a un modelo multicausal donde la variable respuesta es continua. Que, en presencia de una variable respuesta no-continua o dicotómica, el modelo a ser aplicado debiera ser un modelo de Regresión Logística (Penna, 2006).

Muchas veces, particularmente en Ciencias Sociales, generar modelos multicausales requiere incluir todas las variables, sin embargo necesitamos trabajar con variables regresoras que, a menudo, no son variables continuas sino estamos en presencia de variables de clasificación o de atributos, variables éstas que no tienen ninguna relación numérica pura (no presentan magnitud). Si esto ocurre y queremos que los grados de libertad del error sean mayores, tendríamos que "incluir" todas las variables dentro del modelo. Frente a los atributos como variables explicativas, y como la regresión múltiple sólo acepta variables numéricas, la única manera de una posible inclusión de las mismas es a través de la creación de variables dummy (McCullagh y Nelder, 1989), también llamadas variables indicadoras o pseudo-variables. Estas variables indicadoras presentan resultados dicotómicos (uno o cero) de acuerdo a la aparición –o no– de la característica del atributo.

Lo que habitualmente se hace, a la hora de generar variables dummy, es decidir el número necesario de dichas pseudo-variables a partir de un valor k–1, donde k es el número de niveles de la variable original.

El ejemplo presentado en el presente trabajo, tomado de la realidad, posiblemente clarifique aun más la inclusión de dichas variables dummy en un modelo de regresión múltiple.

Introducción

La experiencia se llevó a cabo de la siguiente manera: se tomó una muestra de 120 estudiantes de la carrera de Licenciatura en Psicología de la Universidad Nacional de San Luis, que hubiesen aprobado los exámenes finales de las materias "Metodología de la Investigación I" (correspondiente al segundo año de la carrera) y "Metodología de la Investigación II" (del cuarto año de la carrera). Dicha muestra la constituyeron alumnos de ambos sexos donde, además, la mitad de ellos eligió la orientación cognitiva y el resto la orientación psicoanalítica.

A cada sujeto se le consideraron (en suma) las notas finales de las materias mencionadas en el párrafo anterior. La siguiente tabla nos muestra los resultados obtenidos (1):

Tabla 1: Puntajes de los sujetos, de acuerdo a la Orientación, por año y sexo

Año	Cognitiva (n=60)		Psicoanalítica (n=60)	
	Masculino	Femenino	Masculino	Femenino
1 = 2000	514	467	502	450
2 = 2001	512	470	501	459
3 = 2002	513	470	501	459
4 = 2003	509	465	495	449
5 = 2004	507	466	497	445
6 = 2005	505	461	494	444

A partir de esto consideramos, como primer "ajuste" del modelo, uno que permita examinar las diferencias entre orientaciones y entre géneros, en las regresiones del puntaje con el año. Para ello se utilizarán tres variables indicadoras: MC que valga 1 si el alumno es de sexo masculino y tiene como orientación cognitiva y 0 en caso contrario; MP para indicar los alumnos de sexo masculino con orientación psicoanalítica y FC para alumnos de sexo femenino con orientación cognitiva.

Como sabemos, el modelo estadístico se puede pensar como un "constructor mental" que nos permite estudiar y entender cualquier fenómeno subyacente en una relación causa-efecto. Este concepto es clave al momento de definir y entender ciertos procesos inferenciales, considerando la medición en una variable explicativa sobre la unidad de estudio.

Ahora, como todo modelo presenta una parte sistemática (X'β) y una aleatoria (ε), lo podemos definir de la siguiente manera (Dobson, 1983):

$$Y = X'\beta + \varepsilon$$

Siendo β el vector de parámetros, X' la matriz (transpuesta) de variables aleatorias y ε, el error aleatorio no observable que, generalmente, se supone distribuido normalmente.

Donde los supuestos (2), como requisitos y limitaciones teóricas para la aplicación de un modelo de regresión múltiple, son: linealidad, normalidad y equidistribución de los residuos, variables independientes, colinealidad y outliers.

Para nuestro caso, según lo propuesto, el modelo sugerido está dado por:

$$\text{Puntaje} = \text{Constante} + \beta_1(\text{Año}) + \beta_2(\text{MC}) + \beta_3(\text{MP}) + \beta_4(\text{FC}) + \beta_5(\text{MC*Año}) + \beta_6(\text{MP*Año}) + \beta_7(\text{FC*Año}) + \text{error}$$

Donde "Puntaje" es la variable dependiente (o variable respuesta); entre paréntesis se encuentran las variables independientes, las variables dummy y las interacciones; los βi (i=1, ..., 7) son los coeficientes del modelo que minimizan los residuos y "error" es el error aleatorio no observable.

El paquete estadístico utilizado en el presente trabajo es el *Infostat* (Infostat 2002, 2003) y la técnica aplicada para la eliminación de coeficientes es la *backward*.

Los resultados obtenidos, después de aplicar la técnica de eliminación propuesta en el párrafo anterior, son los siguientes:

Salida 1

Análisis de regresión lineal

Variable	N	R^2	R^2Aj	ECMP
Puntaje	24	0,99	0,98	14,56

Coeficientes de regresión y estadísticos asociados

Eliminación backward. Máximo p-valor para retener regresoras: 0,15

Coef.	Est.	EE	LI(95%)	LS(95%)	T	p-valor	CpMallows
Const.	457,28	1,77	453,57	460,98	258,64	<0,0001	
Año	-1,79	0,36	-2,55	-1,04	-4,96	0,0001	27,41
MC	59,00	1,75	55,34	62,66	33,79	<0,0001	1088,42
MP	47,33	1,75	43,68	50,99	27,10	<0,0001	701,97
FC	15,50	1,75	11,84	19,16	8,88	<0,0001	78,89

De acuerdo a los resultados que podemos ver en la Salida 1, el modelo estimado (que llamaremos modelo 1) será:

$$\text{Puntaje} = 457{,}28 - 1{,}79(\text{Año}) + 59{,}00(\text{MC}) + 47{,}33(\text{MP}) + 15{,}50(\text{FC})$$

Del modelo 1 se desprende lo siguiente:

1.1. Cuando queremos evaluar el comportamiento de un sujeto de sexo masculino que eligió la orientación cognitiva (MC=1, cero para el resto) el modelo estimado –ecuación 1.1– será:

$$\text{Puntaje} = 457{,}28 - 1{,}79(\text{Año}) + 59{,}00 \Rightarrow \text{Puntaje} = 516{,}28 - 1{,}79(\text{Año})$$

1.2. Cuando queremos evaluar el comportamiento de un sujeto de sexo masculino con orientación psicoanalítica (MP=1, cero para el resto) el modelo estimado –ecuación 1.2– será:

$$\text{Puntaje} = 457{,}28 - 1{,}79(\text{Año}) + 47{,}33 \Rightarrow \text{Puntaje} = 504{,}61 - 1{,}79(\text{Año})$$

1.3. Cuando queremos evaluar el comportamiento de un sujeto de sexo femenino con orientación cognitiva (FC=1, cero para el resto) el modelo estimado –ecuación 1.3– será:

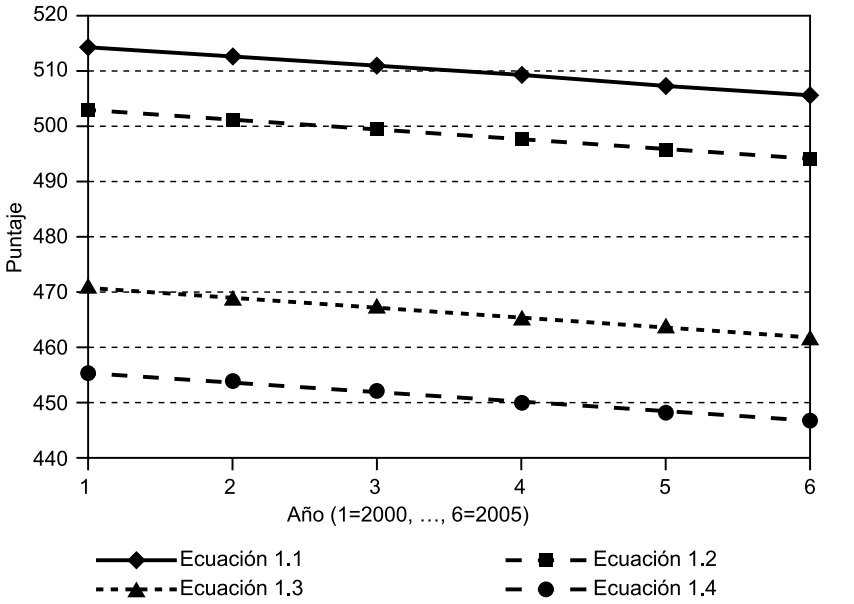
$$\text{Puntaje} = 457{,}28 - 1{,}79(\text{Año}) + 15{,}50 \Rightarrow \text{Puntaje} = 472{,}78 - 1{,}79(\text{Año})$$

1.4. Cuando queremos evaluar el comportamiento de un sujeto de sexo femenino con orientación psicoanalítica (MC=MP=FC=0) el modelo estimado –ecuación 1.4– será:

$$\text{Puntaje} = 457{,}28 - 1{,}79(\text{Año})$$

Por último, para aquellos lectores "amantes" de las representaciones gráficas, presentamos el Gráfico 1 que nos muestra el comportamiento, de manera conjunta, de las cuatro ecuaciones generadas en el punto anterior:

Gráfico 1: Ecuaciones de regresión generadas a partir del modelo 1



Algunas conclusiones sobre el modelo presentado (modelo 1)

Es de notar que en el modelo 1, no aparecen las interacciones. Esto nos está hablando que las mismas fueron excluidas en la selección por ser no significativas, lo que nos permite ver que las pendientes son iguales y que no existe dependencia entre año, sexo y orientación.

Ahora bien, si analizamos las ecuaciones propuestas anteriormente obtenemos que todas las rectas, como ya dijimos, tienen la misma pendiente negativa, decreciendo (casi) 2 puntos por año.

Notamos, además, que tienen mejor puntaje los varones que las mujeres (para ambas orientaciones); si miramos el sexo, en ambos casos, los puntajes obtenidos en la orientación cognitiva son mejores que los obtenidos en psicoanálisis.

El segundo modelo que vamos a presentar es con dos variables dummy y las interacciones correspondientes, siendo de las variables indicadoras relacionadas con Sexo y otra con la Orientación. Decidimos generar un segundo modelo en función de poder "decidir" cual de los dos es más representativo del comportamiento de la información presentada.

Para este caso, según lo propuesto, se ajusta otro modelo con dos variables indicadoras: Sexo (masculino = 1 y femenino = 0) y Orientación (cognitiva = 1 y psicoanalítica = 0) y las interacciones correspondientes. Donde el modelo sugerido está dado por:

$$\text{Puntaje} = \text{Constante} + \beta_1(\text{Año}) + \beta_2(\text{Sexo}) + \beta_3(\text{Orientación}) + \beta_4(\text{Sexo*Orientación}) + \beta_5(\text{Sexo*Año}) + \beta_6(\text{Orientación*Año}) + \beta_7(\text{Orientación*Año*Sexo}) + \text{error}$$

Mediante la utilización de *Infostat*, y aplicando la eliminación *backward*, fueron obtenidos los siguientes resultados:

Salida 2

Análisis de regresión lineal

Variable	N	R^2	R^2Aj	ECMP
Puntaje	24	0,99	0,98	14,21

Coeficientes de regresión y estadísticos asociados

Eliminación backward. Máximo p-valor para retener regresoras: 0,15

Coef.	Est.	EE	LI(95%)	LS(95%)	T	p-valor	CpMallows
Const.	458,23	1,71	454,66	461,81	267,33	<0,0001	
Año	-1,79	0,37	-2,57	-1,01	-4,79	0,0001	24,93
Sexo	45,42	1,28	42,75	48,08	35,55	<0,0001	1206,56
Orientación	13,58	1,28	10,92	16,25	10,63	<0,0001	110,70

De acuerdo a los resultados que podemos ver en la Salida 2, el modelo estimado (que llamaremos modelo 2) será:

$$\text{Puntaje} = 458{,}23 - 1{,}79(\text{Año}) + 45{,}42(\text{Sexo}) + 13{,}58(\text{Orientación})$$

Del modelo 2, se desprende lo siguiente:

2.1. Cuando se quiere evaluar el comportamiento de un sujeto de sexo masculino que eligió orientación cognitiva (Sexo=1, Orientación=1) el modelo estimado –ecuación 2.1– es:

$$\text{Puntaje} = 458{,}23 - 1{,}79(\text{Año}) + 45{,}42 + 13{,}58 \Rightarrow \text{Puntaje} = 517{,}23 - 1{,}79(\text{Año})$$

2.2. Cuando se quiere evaluar el comportamiento de un sujeto de sexo masculino que eligió orientación psicoanalítica (Sexo=1, Orientación=0) el modelo estimado –ecuación 2.2– es

$$\text{Puntaje} = 458{,}23 - 1{,}79(\text{Año}) + 45{,}42 \Rightarrow \text{Puntaje} = 503{,}69 - 1{,}79(\text{Año})$$

2.3. Cuando se quiere evaluar el comportamiento de un sujeto de sexo femenino que eligió orientación cognitiva (Sexo=0, Orientación=1) el modelo estimado –ecuación 2.3– es:

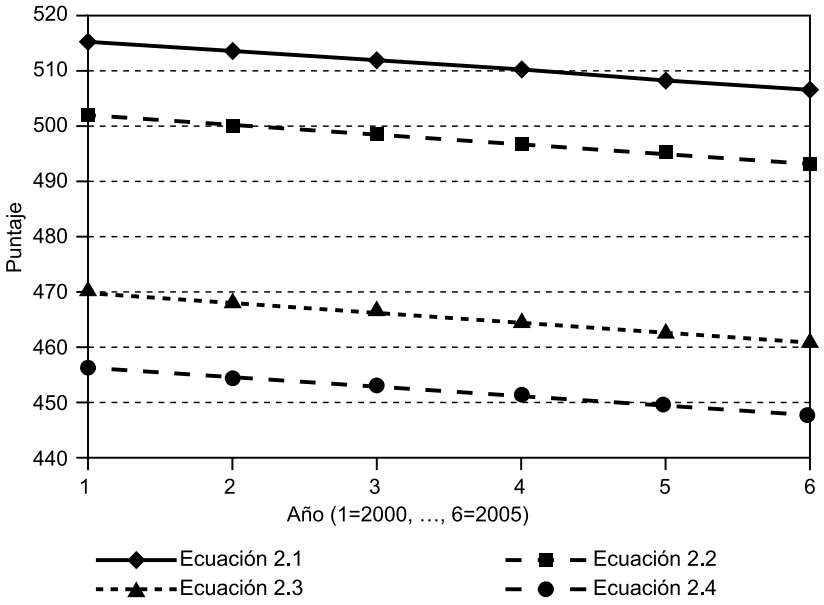
$$\text{Puntaje} = 458{,}23 - 1{,}79(\text{Año}) + 13{,}58 \Rightarrow \text{Puntaje} = 471{,}81 - 1{,}79(\text{Año})$$

2.4. Cuando se quiere evaluar el comportamiento de un sujeto de sexo femenino con orientación psicoanalítica (Sexo=0, Orientación=0) el modelo estimado –ecuación 2.4– es

$$\text{Puntaje} = 458{,}23 - 1{,}79(\text{Año})$$

Del mismo modo que para el modelo 1, presentamos el Gráfico 2 que nos muestra el comportamiento, de manera conjunta, de las cuatro ecuaciones generadas en el punto anterior:

Gráfico 2: Ecuaciones de regresión generadas a partir del modelo 2



Algunas conclusiones sobre el modelo presentado (modelo 2)

Como las interacciones no son significativas (motivo por el cual fueron excluidas en la selección), nos permite probar que las pendientes son iguales y que no existe dependencia entre Año, Sexo y Orientación.

Como ocurriera en el modelo1, las ecuaciones propuestas anteriormente nos muestran que las rectas tienen la misma pendiente negativa y decrecen (casi) 2 puntos por año.

Notamos, además, que tienen mejor puntaje los varones que las mujeres (para ambas disciplinas); si miramos el género, en ambos casos, los puntajes obtenidos en alumnos que eligieron la orientación cognitiva son mejores que los obtenidos por aquellos que eligieron psicoanalítica.

Interpretación conjunta: conclusión general

Como podemos ver en las interpretaciones de los modelos 1 y 2, las características son similares para los dos enfoques. En ambos casos, las interacciones resultaron no significativas, las pendientes son las mismas

(negativas y decrecen casi 2 puntos por año) y los mejores puntajes, para ambas disciplinas, los tienen los varones y si consideramos los géneros por separado, los mejores puntajes fueron obtenidos en la orientación cognitiva.

Las que varían (pero no significativamente) son las ordenadas al origen.

Podríamos, además, comparar los resultados obtenidos en ambos modelos por el R^2Aj (3) pero, como podemos ver tanto en la Salida 1 como en la Salida 2, el valor es igual para ambos (0,98). Otra mirada para determinar el "mejor" modelo, son los puntajes obtenido por ambas ecuaciones en el ECPM (4). Si miramos los resultados a partir del análisis de regresión, vemos que, para el modelo 1, el ECPM es igual a 14,56; y para el modelo 2 es de 14,21.

A la hora de tomar una decisión y a pesar que los resultados son bastante parecidos (por no decir que el comportamiento es estadísticamente igual), nos quedamos con el enfoque del modelo 2 ya que el valor del ECMP es menor que el del modelo 1, esto nos lleva a asegurar que en el segundo modelo hay más datos influyentes que en primero.

San Luis (Argentina), 2 de julio de 2010.

Notas

1) Se recomienda a los lectores del presente trabajo, no se sientan mal ni se "devanen" los sesos preguntándose cómo fue tomada la muestra, si la misma fue aleatoria, si es representativa, etc., pues todas las respuestas convergen en un "no" rotundo. Sucede que el "nudo gordiano" de esta investigación no es la muestra per se (que sí podría ser tema de otro trabajo), sino que es la aplicación de la metodología empleada.

2) Detalle en el cual no vamos a detenernos en la presente investigación pero, a la hora de realizar un trabajo, hay que tenerlos presentes para la correcta aplicación de un modelo de regresión múltiple.

3) Coeficiente de Determinación ajustado o corregido. Este coeficiente se llama así pues corrige o "penaliza" al modelo –disminuyendo el valor del mencionado coeficiente– a medida que le agregamos variables (Infostat, 2002).

4) Error cuadrático medio de predicción, el cual representa la función objetivo que debe ser minimizada en el proceso de reestimación del modelo a los fines predictivos (Infostat, 2002).

Referencias bibliográficas

Dobson, A. J. (1983). *Introduction to Statistical Modelling* (pp 45 - 56). London: Chapman & Hall Ltd.

Grupo InfoStat (2002). *InfoStat, versión 1.1. Manual del Usuario*. Córdoba: Brujas.

Grupo Infostat (2003). *Infostat/Profesional versión 1.5*. Grupo Infostat, Facultad de Ciencias Agrarias, Universidad Nacional de Córdoba. Córdoba: Autor.

McCullagh, P. y Nelder, J. A. (1989). *Generalized Linear Models* (pp. 48 - 97) (segunda edition). London: Chapman y Hall Ltd.

Nelder, J. A. y Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, Serie A, 135, pp. 370-384.

Penna, F. (2006). Modelo de Regresión Logística aplicada a niños con maloclusión dental. *Fundamentos en Humanidades*. Año VII; Nº I-II (13-14), pp. 201-211.

Herschel, J. ([1830] 1987). A preliminary Discourse on the Study of Natural Philosophy. Chicago: University of Chicago Press.