TENDENCIAS
Revista de la Facultad de Ciencias
Económicas y Administrativas.
Universidad de Nariño
Vol. VI. Nos.1-2
Diciembre de 2005, páginas 97-108

MODELIZACIÓN ESTADÍSTICA DE VARIABLES CUALITATIVAS: UNA INTRODUCCIÓN APLICADA

Por: Julio César Riascos[1]

"*Each day is a drive thru History*"
James Douglas Morrison
(1943-1979)

RESUMEN

El artículo introduce al lector en el manejo y aplicación elemental de modelos probabilísticos, asistidos por variables dicótomas; para tal efecto determina su trascendencia en el proceso de investigación científica y su diseño a nivel básico. Se determinan, en este sentido, los modelos ANOVA de *análisis de varianza*; los modelos ANCOVA de *regresores cualitativos* y *cuantitativos*, para finalmente abordar las *ecuaciones de respuesta binaria*.

[1] **Economista**, Grado de Honor, Egresado Distinguido, Docente Hora Cátedra Universidad de Nariño. Email: julioriascos@mail.udenar.edu.co

PALABRAS CLAVE: Modelización estadística, Modelos ANOVA, ANCOVA, Modelos de respuesta cualitativa.

I. GENERALIDADES

La aplicación cada vez más frecuente del modelamiento estadístico en el campo de la investigación científica, constituye acaso la prueba irrefutable del auge cuantitativo, no sólo como herramienta de medición y predicción, sino como instrumento vital en la toma de decisiones.

Así por ejemplo, el análisis estadístico incorporado a la biología, en el caso de las ciencias de la salud, o lo que se conoce como *bioestadística*, es asistido por el examen de modelos probabilísticos en el estudio e identificación de diversas patologías.

Igualmente,

> "la misión del económetra es la de expresar las teorías económicas en términos matemáticos para verificarlas por métodos estadísticos y para medir el impacto de una variable sobre otra, así como para predecir los sucesos futuros o aconsejar la política económica que debe seguirse cuando se desea un resultado determinado"[2].

Ahora bien, la incidencia real de esta rama de la ciencia económica no se encuentra únicamente en la macroeconometría; su importancia práctica es igual en la microeconometría, aplicada dentro de la investigación de mercados y en el diseño y evaluación de proyectos empresariales.

[2] VALAVANIS, Stefan. Introducción a la Econometría. Citado por: BARBANCHO (1979: 182).

Una de las críticas convencionales a la modelización estadística descansa en el hecho de que se reduce aspectos cualitativos, a expresiones matemáticas implacables, suponiendo de ese modo una rigurosa mecánica entre relaciones que dentro de la vida cotidiana rara vez existe; pues bien, por lo menos en el trabajo del economista, no se debe perder de vista que la ciencia permanentemente se ha desbordado en el desarrollo de conceptos teóricos y análisis econométricos, y aún cuando en retrospectiva lo concebido resulta impresionante, no ha sido suficiente para afrontar los retos actuales. Con todo,

> "si el objeto social de la ciencia constituye su máximo fin, debe asegurar para tal efecto las condiciones fundamentales de su trabajo como son, en este caso, las herramientas conceptuales y estadísticas; que tampoco constituyen el límite de la economía, sino más bien, el eje que posibilita la extensión de su aporte"[3].

En ese orden de ideas, y dejando en claro que la modelización estadística, así como el conocimiento en general, constituyen solamente instrumentos de que se sirve la ciencia para intentar clarificar la realidad y que, por consiguiente, en ningún caso pueden ser un fin *per se*; el objeto central de este artículo será presentar al estudiante una introducción aplicada, muy elemental por cierto, en el manejo de modelos probabilísticos asistidos por variables "ficticias", en la cual se abordarán los modelos **ANOVA**, **ANCOVA** y de *Respuesta Cualitativa*.

II. DISEÑO DE VARIABLES

No todos los eventos estadísticos son obligatoriamente medibles, o cuantificables, y el hecho de que ello sea así, no implica que la incidencia de dichos elementos deba ignorarse. Es más, en ocasiones estos factores basados en aspectos cualitativos, sobrepasan con amplitud la significancia de aquellos fenómenos pertenecientes a una determinada escala de razón[4].

---

[3] RIASCOS, Julio César. **Principales determinantes económicos del desempleo en San Juan de Pasto**. Tesis de grado para optar el título de Economista. Facultad de Ciencias Económicas y Administrativas. Programa de Economía. Universidad de Nariño. Pasto, 2004. Pág. 137.

[4] Tradicionalmente el uso de *escalas de razón* se asoció al desarrollo de la *investigación cuantitativa* y, de manera similar, el uso de *escalas nominales* se vinculó a la producción de

La construcción de variables "dummy" es llevada a cabo mediante el uso del sistema binario; la presencia de un atributo o cualidad implica que cada observación tomará valores equivalentes a 1 y, en caso de ausencia, cada evento adoptará valores iguales a 0.

Para efectos prácticos considérese el modelo lineal expresado en la ecuación (1).

$$Yt = \beta 1 + \beta 2X2t + \mu t \qquad (1), \text{ donde:}$$

Yt = Salarios Nominales (Variable Endógena)
X2t = Educación (Variable Exógena)
μt = Término de Error
$\beta 1$= Parámetro Autónomo[5]
$\beta 2$= Parámetro de Impacto[6]

Ahora supóngase que se cuenta con la siguiente información: (Cuadro A)

El lector advertirá que la educación se descompone en 5 categorías diferentes: Primaria, Secundaria, Técnica, Pregrado y Postgrado. Intuitivamente, el diseño de regresores ficticios supondrá construir una variable para cada cualidad; así entonces de X2t se generarían D2, D3, D4, D5 y D6, donde D2 representaría la educación Primaria, siendo 1 cuando las observaciones presentan la existencia de ese evento específico y 0 para todo evento que le sea distinto. D3, D4, D5 y D6 comprenderán la educación Secundaria, Técnica, Pregrado y Postgrado, respectivamente, teniendo en cuenta la presencia y ausencia de cada categoría mediante el sistema binario. De ese modo, las variables ficticias habrán originado los datos del cuadro B

---

investigación cualitativa; no obstante, en la actualidad buena parte de los avances en materia de *investigación científica,* tienen a bien asistir una combinación de áreas *cuasicuantitativas* o *cuasicualitativas* en los recientes diseños de ***investigación experimental***.

[5] El valor del intercepto estará asociado con el efecto promedio que sobre los salarios nominales ejercen el conjunto de variables omitidas del modelo.

[6] Reflejará la incidencia de los niveles educativos sobre los salarios nominales.

CUADRO A

INFORMACIÓN GENERAL EMPLEADOS ALFAOMEGA S.A.

(Ejemplo hipotético)

SALARIOS (Miles de pesos)	NIVELES EDUCATIVOS	GÉNERO	PROCEDENCIA
200	Primaria	M	Putumayo
205	Primaria	M	Nariño
220	Secundaria	M	Cauca
228	Secundaria	F	Nariño
252	Técnica	F	Nariño
264	Técnica	M	Putumayo
272	Técnica	M	Putumayo
315	Pregrado	F	Cauca
324	Pregrado	F	Putumayo
340	Pregrado	M	Putumayo
618	Postgrado	F	Nariño
720	Postgrado	M	Nariño
800	Postgrado	F	Cauca

Fuente: **RIASCOS**, Julio. Economía: Retos y Posibilidades. (Investigación en curso), 2003, 2004, 2005.

CUADRO B

VARIABLES DUMMY DE LOS NIVELES DE EDUCACIÓN

D2	D3	D4	D5	D6
1	0	0	0	0
1	0	0	0	0
0	1	0	0	0
0	1	0	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	0	1
0	0	0	0	1
0	0	0	0	1

Se entiende que, aunque la educación es una variable no medible, como sucede con elementos de estirpe cualitativa, es posible aproximarse, sino a su cuantificación, sí por lo menos a la incidencia de su participación.

Erróneamente podría formularse el siguiente modelo:

$$Yt = \beta 1 + \beta 2D2i + \beta 3D3i + \beta 4D4i + \beta 5D5i + \beta 6D6i + \mu t \qquad (2).$$

Al aplicar **M. C. O.** para estimar los parámetros β **estimado** $= (X'X)^{-1} (X'Y)$, el cálculo de la matriz inversa A exp.-1 = (1/ $|A|$)* (Adj. A) encontrará que el determinante $|A|$ será equivalente a 0, con lo que dicha matriz se hace singular. Si A exp.-1 tiene n*n filas y columnas, en este caso el rango es menor que "n", implicando que existe una relación lineal perfecta entre los regresores que componen el modelo; por lo tanto, existiría ***multicolinealidad exacta***[7], ocasionada a su vez por lo que se conoce como la ***trampa de las variables ficticias***.

III. MODELOS ANOVA O MODELOS DE ANÁLISIS DE VARIANZA

El apartado anterior ha dejado una lección valiosa en cuanto al manejo y precaución de variables dicótomas. Nótese el cuadro B, para comprender la existencia de *multicolinealidad perfecta* al estimar la ecuación (2). En la primera fila, la presencia de D2i implica la ausencia de D3i, D4i, D5i y D6i. De igual forma, en cualquier observación se tiene que la existencia de cualquier evento explica de manera exacta la carencia de las demás. El desacierto estriba en abarcar todas las posibilidades en que incurre un mismo fenómeno sin determinar una categoría que sirva de base. En otras palabras, el investigador debe "sacrificar" una posibilidad en favor de las demás, lo que supone *eliminar una variable que haga las veces de referencia.*

[7] Los parámetros a estimar β son indeterminados y por lo tanto sus errores típicos serán infinitos.

De acuerdo con los propósitos del análisis, la modelización estadística definirá qué elementos serán prioritarios y qué factor se omitirá para establecerse como base; no obstante, recuérdese que $\beta 1$, o el parámetro autónomo, ponderará, entre las variables excluidas, la incidencia de aquella que se ha dejado de lado.

Si, verbigracia, el investigador tomara como referencia la *Educación Primaria*, formularía el modelo de la ecuación (3)

$$Yt = \alpha 0 + \alpha\ 3D3i + \alpha\ 4D4i + \alpha\ 5D5i + \alpha\ 6D6i + \mu t \qquad (3), \text{ donde}$$

Yt = Salario Nominal (Variable Endógena)
D3i = Educación Secundaria (Variable Exógena)
D4i = Educación Técnica (Variable Exógena)
D5i = Educación de Pregrado (Variable Exógena)
D6i = Educación de Postgrado (Variable Exógena)
μt = Término de Error.
$\alpha 0$ = Parámetro Autónomo[8]
$\alpha\ 3$, $\alpha\ 4$, $\alpha\ 5$ y $\alpha\ 6$ = Parámetros de Impacto; cada uno reflejará correspondientemente la incidencia de la *Educación Secundaria*, *Técnica, de Pregrado y de Postgrado*.

> *"El mensaje es: si una variable cualitativa tiene m categorías, solo hay que agregar (m-1) variables dicótomas"* (GUJARATI, 2004: 289)

La expresión (3) es un modelo estadístico, que explica el comportamiento de los salarios nominales en función de la Educación Secundaria, Técnica, de Pregrado y de Postgrado; modelo que puede ser estimado a través de M. C. O.

[8] La teoría supone que en $\alpha 0$ se cuantificarán los efectos de los términos excluidos y entre ellos, el que le asiste a la *educación primaria* que ha servido de referente.

CUADRO C
LS // Dependent Variable is SALARIO
Date: 10/15/05 Time: 05:39
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	202.5000	32.79720	6.174308	0.0003
D3	21.50000	46.38224	0.463539	0.6553
D4	60.16667	42.34100	1.421002	0.1931
D5	123.8333	42.34100	2.924667	0.0192
D6	510.1667	42.34100	12.04900	0.0000

R-squared	0.966140	Mean dependent var	366.0000
Adjusted R-squared	0.949211	S.D. dependent var	0.000000
S.E. of regression	46.38224	Akaike info criterion	7.957556
Sum squared resid	17210.50	Schwartz criterion	8.174845
Log likelihood	-65.17032	F-statistic	57.06743
Durbin-Watson stat	1.734905	Prob(F-statistic)	0.000006

Ahora bien, es posible que el investigador omita una categoría que sea precisamente la más importante; para evitar entonces este tipo de riesgos, es del todo viable incluir esa variable que se podría haber excluido, de tal modo que nuevamente se considerarían el conjunto total de eventos, dejando de lado esta vez el intercepto, cuya justificación aparentemente desaparecería al incluir todas las posibilidades[9].

En tal sentido podría formularse, de manera alternativa, el siguiente modelo:

$$Yt = \alpha\, 2D2i + \alpha\, 3D3i\ \alpha\, 4D4i + \alpha\, 5D5i + \alpha\, 6D6i + \mu t \qquad (4), \text{ donde}$$

α 2 = Educación Primaria

[9] Aunque lo anterior es válido, existe una amplia discusión sobre variables que explican un fenómeno y que sin embargo no se tienen en cuenta, es lo que se denomina *"externalidades"*; aquí por lo tanto podría también existir una contradicción cuando se elimina el término que pondera tales eventos.

D2i = Parámetro de Impacto, que pondera la incidencia de la Educación Primaria al explicar los Salarios.

CUADRO D
LS // Dependent Variable is SALARIO
Date: 10/15/05 Time: 06:22
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	T-Statistic	Prob.
D2	202.5000	32.79720	6.174308	0.0003
D3	224.0000	32.79720	6.829852	0.0001
D4	262.6667	26.77880	9.808754	0.0000
D5	326.3333	26.77880	12.18626	0.0000
D6	712.6667	26.77880	26.61309	0.0000

R-squared	0.966140	Mean dependent var	366.0000
Adjusted R-squared	0.949211	S.D. dependent var	0.000000
S.E. of regression	46.38224	Akaike info criterion	7.957556
Sum squared resid	17210.50	Schwartz criterion	8.174845
Log likelihood	-65.17032	F-statistic	57.06743
Durbin-Watson stat	1.734905	Prob(F-statistic)	0.000006

Los modelos que explican una variable cuantitativa en función de variables cualitativas, como la ecuación (3) y (4), se conocen como **Modelos ANOVA** y, en términos generales, aplican las mismas pruebas estadísticas de un modelo de estirpe cuantitativa.

IV. MODELOS ANCOVA Y DE RESPUESTA CUALITATIVA

Los modelos ANCOVA combinan al mismo tiempo variables exógenas cualitativas y cuantitativas, en la explicación de un fenómeno medible o de escala de razón. Supóngase que se tiene además la siguiente información:

CUADRO E
GASTOS FAMILIARES Y DE CONSUMO INDIVIDUAL
EMPLEADOS ALFAOMEGA S. A
(Miles de pesos)

GASTO FAMILIAR	CONSUMO PERSONAL
125	50
130	65
140	80
142	86
162	90
172	90
200	70
215	80
225	80
400	100
325	130
360	140
380	160

Ahora el investigador perfectamente podría añadir al modelo (3) los datos del cuadro E, teniendo como resultado la ecuación (5):

$Yt = \alpha 0 + \alpha\ 3D3i + \alpha\ 4D4i + \alpha\ 5D5i + \alpha\ 6D6i + \beta 3X3t + \beta 4X4t + \mu t$ (5), donde

X3t = Gasto Familiar (Variable Exógena)
X4t = Consumo Personal

$\beta 3$ y $\beta 4$ = Parámetros de Impacto que miden la incidencia del Gasto Familiar y del Consumo Personal sobre el salario nominal respectivamente.

CUADRO F
LS // Dependent Variable is SALARIO
Date: 10/15/05 Time: 06:17
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	T-Statistic	Prob.
C	58.77773	75.77253	0.775713	0.4674
D3	-39.94652	53.60103	-0.745257	0.4843
D4	-4.043173	48.94460	-0.082607	0.9369
D5	46.25580	57.25687	0.807865	0.4500
D6	293.6460	110.8878	2.648137	0.0381
X3	0.053229	0.313076	0.170021	0.8706
X4	2.381487	1.442229	1.651254	0.1498

R-squared	0.980287	Mean dependent var	366.0000
Adjusted R-squared	0.960573	S.D. dependent var	0.000000
S.E. of regression	40.86593	Akaike info criterion	7.724326
Sum squared resid	10020.14	Schwartz criterion	8.028530
Log likelihood	-61.65432	F-statistic	49.72682
Durbin-Watson stat	1.581633	Prob(F-statistic)	0.000074

Por último, los *modelos de respuesta cualitativa* se caracterizan fundamentalmente porque la variable endógena es una regresada binaria; es decir, el fenómeno que se está explicando es de naturaleza nominal.

En ese orden de ideas podría plantearse, a manera de ejemplo, el siguiente modelo uniecuacional:

$$Yt = \theta 1 + \theta\ 2\ D7i\ + \theta\ 3\ X5t\ + \mu t \qquad (6), \text{ donde:}$$

Yt = Educación de Postgrado (Variable Endógena Binaria)
D7i = Procedencia (Variable Exógena Binaria)
X5t = Salarios (Variable Exógena Cuantitativa)
μt = Término de Error
θ 1= Parámetro Autónomo
θ 2= Parámetro de Impacto que medirá la incidencia de la región de procedencia sobre la Educación de Postgrado

θ 3= Parámetro de Impacto, reflejará el impacto que ejercen los salarios sobre la educación de postgrado.

El lector advertirá que la variable procedencia está constituida por 3 categorías, y que por lo tanto deberán construirse 2 variables dicótomas; no obstante y aún cuando la estimación se haga mediante M.C.O., es posible sus propiedades estadísticas no sean las deseables[10].

Sin embargo, los desarrollos inferenciales han posibilitado formas alternativas de estimación al de M.C.O., mediante *métodos de cálculo binario* en ***modelos logit*** para funciones de distribución acumulativa *(logística),* y en ***modelos probit*** para funciones de distribución normal acumulativa, conceptos cuya exposición desbordaría con amplitud los alcances de este análisis introductorio, pero que pueden ser trabajados por el estudiante en los textos de Gujarati y Gourieroux, con la obvia asistencia de un paquete estadístico relativamente reciente.

REFERENCIAS BIBLIOGRÁFICAS

BARBANCHO, Alfonso (1979). **Fundamentos y Posibilidades de la Econometría**. Editorial Ariel. Barcelona.

CARRASCAL, Ursicinio (2004). **Análisis Econométrico con Eviews**. Alfaomega. Madrid.

GOURIEROUX, Christian (2000). **Econometrics of Qualitative Dependent Variables**. University Press. Nueva York.

GUJARATI , N. Damodar (2004). **Econometría**. Mc. Graw-Hill. México.

MADALLA, G. S. (2001). **Introducción a la Econometría**. Prentice-Hall. México.

PINDYCK, Roberts (2001). **Econometría: Modelos y Pronósticos**. Mc. Graw-Hill. México.

[10] Los errores residuales no siguen una distribución normal, toda vez que su distribución es la Bernoulli; las probabilidades de Heterocedasticidad se incrementan, las estimaciones son proclives a rebasar los valores entre 0 y 1 y, entre otros inconvenientes, los valores de R cuadrado no tendrán a priori mayor poder explicativo