El efecto colegio sobre la variabilidad del rendimiento en matemáticas

Introducción. I. El modelo Jerárquico lineal. II. Una aplicación: el caso de la variabilidad del rendimiento en matemáticas explicada por el colegio. III. Conclusiones. Referencias

Introducción

Muchas de las investigaciones en ciencias sociales tienen que ver con estructuras de datos jerárquicos, es decir, con conjuntos de información donde los datos son observados en diferentes niveles.

Generalmente se tienen variables que describen los individuos, y los individuos están agrupados en unidades más grandes de las cuales también se tienen variables que las describen.

La investigación en educación es uno de estos casos. Allí es de gran interés tratar de determinar la importancia que tienen las características propias de los colegios sobre el rendimiento que obtienen sus estudiantes. En otras palabras, los investigadores quieren saber qué porcentaje de la variabilidad del rendimiento es explicado por el "efecto colegio".

Desafortunadamente, en la mayoría de los estudios realizados hasta ahora, esta pregunta ha quedado sin respuesta debido a que las técnicas usadas, como la regresión lineal estándar, no permiten descomponer la variabilidad del rendimiento entre la variabilidad explicada por las características propias de los estudiantes y la explicada por las características de los colegios. Dichas técnicas fallan al no reconocer que en estos casos los datos son generados por estructuras jerárquicas, es decir los datos son observados en diferentes niveles: se tienen variables que describen los individuos, y los individuos están agrupados en unidades más grandes de las cuales también se tienen variables que las describen. Los estudiantes están agrupados en colegios. Se tienen variables que describen tanto a los estudiantes como a los colegios.

Este documento está organizado de la siguiente manera: en la sección I se expone brevemente la teoría relacionada con los modelos jerárquicos lineales de dos niveles y se presenta, como un submodelo de la clase general, el modelo de Análisis de Varianza de una Vía con Efectos Aleatorios que permite dar respuesta a la pregunta inicial. La sección II presenta una aplicación a la **Encuesta Saber 93**. La sección III presenta algunas conclusiones.

I. El modelo jerárquico lineal

Una vez hemos reconocido que los datos pertenecen a una estructura jerárquica, debemos pensar en las técnicas estadísticas que tengan en cuenta esta estructura. Hay dos procedimientos obvios que han sido algo desacreditados: El primero es desagregar las variables de niveles más altos al nivel individual y realizar el análisis a nivel de individuos. Por ejemplo, las características del colegio son asignadas a cada uno de los estudiantes. El problema con esta aproximación es que no podemos usar la hipótesis de independencia en las observaciones que es básica en las técnicas estadísticas clásicas. La otra alternativa es agregar las variables individuales al nivel más alto y realizar el análisis a este nivel. El problema, en este caso, es que perdemos la información dentro de grupos, la cual puede llegar a ser el 80% o 90% de la variación total antes

de iniciar el análisis. Como consecuencia, las relaciones entre las variables agregadas son con frecuencia mucho más fuertes y pueden ser muy diferentes a las relaciones entre las variables no agregadas. Por tanto, se desperdicia la información y se distorsiona la interpretación si tratamos de interpretar el nivel agregado como el nivel individual. Por tanto, los dos procedimientos no son satisfactorios.

Veamos como podemos construir un modelo que tenga en cuenta la estructura jerárquica de los datos. Supongamos que cada grupo tiene una ecuación de regresión distinta y que varios grupos son muestreados. Entonces podemos suponer que los interceptos y pendientes de los colegios seleccionados son una muestra aleatoria de una población de interceptos y pendientes. Esto define los modelos de regresión de coeficientes aleatorios:

$$Yij = \beta_{0j} + \beta_{1j} X_{1ij} + \ldots + \beta_{Qj} X_{Qij} + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \mu_{oj}$$

$$\beta_{1j} = \gamma_{10} + \mu_{1j}$$

.

.

.

$$\beta_{Qj} = \gamma_{Q0} + \mu_{Qj}$$

donde:

Y_{ij} es la respuesta del i-ésimo individuo de la unidad j, i=1,2,...,nj, j=1,2,...,J

X_{qij} es el valor de la q-ésima característica del individuo i de la unidad j, q=1,2,...,Q, i=1,2,...,n_j, j=1,2,...,J.

β_{qj} es el coeficiente aleatorio de la q-ésima característica de la unidad j, q=1,2,...,Q, j=1,2,...,J.

r_{ij} es el efecto aleatorio del i-ésimo individuo en la unidad j. Se supone que son independientes y tienen distribución normal con media cero y varianza constante σ^2.

μ_{qj} es el efecto aleatorio de la q-ésima característica en la unidad j. Se supone que son independientes entre sí y de r_{ij} y tienen distribución conjunta normal (Q+1)-variada con vector de medias ceros y matriz de covarianzas T= (τ_{lm}) donde:

τ_{lm} = varianza (β_{mj}) si l=m

τ_{lm} = covarianza(β_{mj}, β_{lj}) si l diferente de m

Las τ_{lm} son llamadas componentes de varianza y covarianza del modelo.

En esta clase de modelos no hay posibilidad de incorporar variables de niveles más altos y por esta causa es necesario definir los modelos multinivel o modelos jerárquicos, en los cuales el modelo de cada nivel es de nuevo un modelo lineal. En el caso de dos niveles el modelo se puede escribir como

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{1ij} + \ldots + \beta_{Qj} X_{Qij} + r_{ij},\ i=1,2,\ldots,nj,\ j=1,2,\ldots,J \qquad (1)$$

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1} W_{1j} + \gamma_{q2} W_{2j} + \ldots + \gamma_{q,sq} W_{sq,j} + \mu_{qj}, \qquad (2)$$

para q=0,1,2,...,Q,

y donde:

Y_{ij} es la respuesta del i-ésimo individuo de la unidad j, i=1,2,...,nj, j=1,2,...,J

X_{qij} es el valor de la q-ésima característica del individuo i de la unidad j, q=1,2,...,Q, i=1,2,...,nj, j=1,2,...,J.

β_{qj} es el coeficiente aleatorio de la q-ésima característica de la unidad j, q=1,2,...,Q, j=1,2,...,J.

γ_{qs} es el coeficiente fijo que captura la influencia de la s-ésima variable predictora de la ecuación q de segundo nivel.

r_{ij} es el efecto aleatorio del i-ésimo individuo en la unidad j. Se supone que son independientes y tienen distribución normal con media cero y varianza constante σ^2.

μ_{qj} es el efecto aleatorio de la q-ésima característica en la unidad j. Se supone que son independientes entre sí y de r_{ij} y tienen distribución conjunta normal (Q+1)-variada con vector de medias cero y matriz de covarianzas T= (τ_{lm}), donde:

τ_{lm} = varianza(μ_{qj}), si l=m

τ_{lm} = covarianza(μ_{mj} , μ_{lj}), si l es diferente de m, l,m=0,1,...,Q.

Así llegamos a una clase de modelos que tiene en cuenta la estructura jerárquica de los datos y que hace posible la incorporación de variables predictoras en todos los niveles. Este modelo es llamado **modelo jerárquico lineal de dos niveles**. La ecuación (1) es llamada modelo de primer nivel o de nivel micro, y las Q+1 ecuaciones en (2) son llamadas modelo de nivel 2 o modelo de nivel macro.

A. Estimación y contraste de hipótesis

En un modelo jerárquico se pueden estimar y contrastar hipótesis sobre tres tipos de parámetros: los efectos fijos (las γ_{qs}), los coeficientes aleatorios del primer nivel (los β_{qj}) y las componentes de varianza y covarianza (los γ_{lm}). Una descripción de los procedimientos de estimación empleados se encuentra en Bryck y Raudenbush (1992).

B. El modelo análisis de varianza de una vía con efectos aleatorios

Un caso particular de el modelo completo jerárquico lineal de dos niveles es:

$$Y_{ij} = \beta_{0j} + r_{ij} \quad (3)$$

$$\beta_{0j} = \gamma_{00} + \mu_{oj} \quad (4)$$

Este es el más simple de los modelos jerárquicos y es llamado modelo de **Análisis de Varianza de una Vía con Efectos Aleatorios.**

Este modelo predice la respuesta de cada unidad del nivel 1 con el parámetro del nivel 2, el intercepto β_{0j}. En este caso β_{0j} es la respuesta media de la j-ésima unidad. Es decir $\beta_{0j} = u_{yj}$.

En el modelo de nivel 2, γ_{00} representa la gran media de la respuesta en la población, y μ_{oj} es el efecto aleatorio asociado a la j-esima unidad.

El modelo combinado es

$$Y_{ij} = \gamma_{00} + \mu_{oj} + r_{ij}$$

el cual es claramente un modelo de Análisis de Varianza de una Vía con una gran media γ_{00}, con un efecto de grupo (nivel 2), μ_{oj}, y un efecto de individuo (nivel 1), r_{ij}.

Observe que la varianza de la respuesta es

$$\text{var}(Y_{ij}) = \text{var}(r_{ij} + \mu_{oj}) = \tau_{00} + \sigma^2$$

Este es el resultado es el más importante que proporciona el modelo, pues informa como se descompone la variabilidad total de la respuesta en términos de las variabilidades de cada uno de los niveles. El parámetro σ^2 representa la variabilidad dentro de grupos y τ_{00} captura la variabilidad entre grupos.

Un parámetro muy útil asociado a este modelo y derivado de la igualdad anterior es el **coeficiente de correlación intraclase,** definido como:

$$\rho = \tau_{00}/(\sigma^2 + \tau_{00})$$

ρ mide la proporción de la varianza de la respuesta que es explicada por el nivel 2.

La confiabilidad de la media muestral del rendimiento en el colegio j, como estimador del verdadero rendimiento medio β_{oj} ,se define como:

$$\lambda_j = \text{conf(media muestral}_j) = \text{Var}(\beta_{oj})/\text{Var(Media muestral}_j) = \tau_{00}/(\tau_{00} + \sigma^2/nj)$$

Un estimador se obtiene reemplazando τ_{00} y σ^2 por sus estimadores. λ_j toma valores entre 0 y uno y mientras más cerca esté de uno mayor confiabilidad tendrá la media muestral. En general, la confiabilidad varía de colegio en colegio debido a que el tamaño muestral n_j varía. Una medida global de la confiabilidad es el promedio de las confiabilidades λ_j, $\lambda = \Sigma \lambda_j /J$.

II. Una Aplicación: El caso de la variabilidad del rendimiento en matemáticas explicada por el colegio.

En un estudio sobre los factores que inciden en el rendimiento en matemáticas de los estudiantes de tercero de primaria, basados en la Encuesta Saber 93 calendario A, se quería investigar que proporción de la variabilidad de los rendimientos era explicada por los colegios. De la muestra, este trabajo solamente usa los colegios con 10 o más alumnos, con el fin de obtener estimaciones de mayor confiabilidad. Se consideraron un total de 566 colegios con un promedio de aproximadamente 18 estudiantes por colegio y para un total de 10150 estudiantes analizados.

Para dar respuesta a la pregunta inicial se corrió el modelo de Análisis de Varianza de una Vía de efectos aleatorios descrito en las ecuaciones (3) y (4). Se empleó el programa HLM (Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs) que arrojó los siguientes resultados:

Sigma cuadrado = 6.88063

Tau del intercepto = 3.62278

Coeficiente Aleatorio del nivel 1		estimación de la confiabilidad
Intercepto	β_{0j}	0.869

Valor función de verosimilitud en la iteración 6 : -4.131696E+004

Efectos fijos		Coeficiente	Error Std	Cociente t	valor p
Intercepto	β_{0j}				
intercepto	γ_{00}	7.866840	0.074715	105.291	0.000

Efecto aleatorio		Error Std	Comp. de Var.	Gl	Chi-cuadr	valorp
Intercepto	μ_{oj}	1.65575	2.74151	565	4350.54201	0.00
nivel 1,	r_{ij}	2.61207	6.82293			

Estimación del Coeficiente de correlación intraclase:

$\rho^* = 2.74151/(2.74151+6.82293) = 0.286636$

De los resultados anteriores concluimos:

i) La confiabilidad global de las medias muestrales de los rendimientos como estimadores de los verdaderos rendimientos promedios de cada colegio (β_{0j}) es aproximadamente 0.87. Esto indica que las medias muestrales tienden ser bastante confiables.

ii) La estimación del rendimiento promedio general en matemáticas es de aproximadamente 7.9, con un error estándar de .074715. De aquí, un intervalo del 95% confianza para el rendimiento promedio general es $7.9 \pm 1.96(.074715) = (7.72, 8.01)$. Los valores del rendimiento se encuentran entre 0 y 13.

iii) El coeficiente de correlación intraclase indica que aproximadamente el 29% de la variabilidad del rendimiento de los estudiantes en matemáticas es explicada por el nivel colegio.

iv) En el contraste de Ho: $\tau_{00} = 0$, se encuentra fuerte evidencia de que el logro promedio de los colegios es aleatorio.

III. Conclusiones

En este trabajo se analizaron los datos de la Encuesta Saber 93 para los colegios del calendario A. Se quería investigar la proporción de la variabilidad del rendimiento en matemáticas que es explicada por los colegios: Se encontró que aproximadamente el 29% de dicha variabilidad es explicada por las características de los colegios. El modelo jerárquico de dos niveles de Análisis de Varianza de una Vía con Efectos Aleatorios, nos permite responder esta pregunta adecuadamente puesto que su estructura tiene en cuenta dos niveles de información: la información correspondiente a los estudiantes y la información correspondiente a los colegios.

Referencias

Bryck, A.S. and Raudenbush, S.W. (1992). *Hierarchical linear Models,* London, Sage Publications.

HLM. *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L Programs* (1996), Chicago, Scientific Software International Inc.

Mason, W. M.; Wong, G. M. and Entwistle, B. (1983) "Contextual Analysis Through the Multilevel Linear Model". En: S. Leinhardt (Ed.), *Sociological Methodology* (pp. 72-103), San Francisco, Jossey-Bass.