

Comparación de la eficacia de regresión logística politómica y análisis discriminante logístico en la detección del DIF no uniforme

M^a Dolores Hidalgo Montesinos y Juana Gómez Benito*
Universidad de Murcia y * Universidad de Barcelona

El objetivo de este trabajo fue evaluar la eficacia en la detección del DIF en ítems politómicos, de Análisis Discriminante Logístico y de Regresión Logística Politómica. Se ha trabajado con un test de 30 ítems con cinco categorías de respuesta, donde el DIF fue simulado en seis ítems. Las condiciones bajo estudio fueron: cantidad de DIF (0.5, 1.0 y 1.5), tamaño muestral (500 y 1000) y tipo de DIF (no-uniforme). Los resultados encontrados apuntan que el Análisis Discriminante Logístico detectó mayor porcentaje de ítems con DIF que Regresión Logística Politómica. Sin embargo, para ambos procedimientos la tasa de falsos positivos fue similar.

DIF detection using Logistic Discriminant Analysis and Polytomous Logistic Regression. This study focused on the effectiveness in polytomous item DIF detection using Logistic Discriminant Analysis and Polytomous Logistic Regression. The simulated test consisted of 30 items with five categories per item, where DIF was manipulated in six items of the test. The conditions under study were: DIF effect size (0.5, 1.0 and 1.5), sample size (500 and 1000) and DIF type (nonuniform). The results suggest that Logistic Discriminant Analysis is more accurate in DIF detection than Polytomous Logistic Regression. However, the false positives rates for both procedures were similar.

La evaluación psicológica y educativa mediante tests estandarizados requiere asegurar que las características psicométricas, tanto del instrumento mismo como de los elementos y cuestiones que lo conforman, son invariantes a través de distintas muestras pertenecientes a la misma población. En este contexto el término funcionamiento diferencial del ítem (DIF) hace referencia a aquellos ítems que presentan propiedades estadísticas distintas en función del grupo en el que ha sido administrado, después que los grupos han sido igualados en el rasgo o habilidad medida por el test (Angoff, 1993). En ítems de respuesta dicotómica el DIF existe cuando las proporciones o probabilidades de una respuesta correcta a este ítem en los grupos focal y de referencia difieren dentro de un nivel de habilidad dado. En formatos de respuesta politómica la definición de DIF es mucho más compleja, dado que éste se puede presentar, entre otras formas, constante a través de todas las categorías de respuesta o sólo en algunas de las categorías. Además, el DIF puede ser también uniforme y no-uniforme (simétrico o compensatorio y asimétrico o no-compensatorio) (Mellenbergh, 1982; De Ayala, Kim, Stapleton y Dayton, 1999). El concepto de DIF uniforme o no-uniforme se refiere a si esta diferencia en proporciones es constante o cambia a lo largo de la escala de habilidad.

Durante los últimos veinte años se han propuesto gran variedad de técnicas estadísticas para detectar el DIF en ítems de respuesta dicotómica (ver revisiones Gómez e Hidalgo, 1997a; Millsap y Everson, 1993) algunas de las cuales han sido adaptadas para de-

teccionar el DIF en ítems de respuesta politómica (ver revisiones Hidalgo y Gómez, 1999; Potenza y Dorans, 1995). En la actualidad, los estudios acerca de la efectividad de las técnicas de detección del DIF en ítems politómicos no son muy numerosos. El presente trabajo compara la eficacia en la detección de DIF no uniforme, en ítems con formato de respuesta politómica, de dos procedimientos: el análisis de regresión logística politómica y el análisis discriminante logístico. También se comparan las tasas de error tipo I de ambos procedimientos.

Regresión Logística Politómica

El análisis de regresión logística politómica (RLP) es una extensión del análisis de regresión logística dicotómico (RLD) para variables de respuesta con más de dos categorías. Swaminathan y Rogers (1990) aplicaron el análisis de regresión logística para detectar el DIF no-uniforme en ítems dicotómicos, esta técnica nos permite modelar la respuesta a un ítem en función de variables como la pertenencia a grupo (G), la habilidad de los sujetos (X) y la interacción habilidad con grupo (XG), y comprobar su relevancia o no para pronosticar la probabilidad de acertar el ítem. Rogers y Swaminathan (1993) mostraron la superioridad de RLD para detectar el DIF no-uniforme frente al estadístico de Mantel-Haenszel. Recientemente se ha propuesto el análisis de RLP como técnica para evaluar el DIF uniforme y no-uniforme en ítems politómicos (French y Miller, 1996; Hidalgo, 1995; Gómez e Hidalgo, 1997b). En estas situaciones la respuesta al ítem se recodifica en $m-1$ variables dicotómicas, siendo m el número de categorías del ítem, siguiendo uno de los esquemas siguientes (Agresti, 1984, 1990; French y Miller, 1996): categorías adyacentes, categorías continuas y acumulativas. Cada una de estas $m-1$ variables dicotómicas se modelan bajo un modelo de RLD, debiendo ajustarse $m-1$ modelos.

Correspondencia: M^a Dolores Hidalgo Montesinos
Facultad de Psicología
Universidad de Murcia
30080 Murcia (Spain)
E-mail: mdhidalg@fcu.um.es

Si asumimos que la variable de respuesta está medida en una escala ordinal, es posible ajustar un modelo de regresión logística politómica asumiendo pendientes paralelas (bajo el esquema de categorías acumulativas). En este modelo, sólo el parámetro de intercepción es diferente para las *m-1* funciones, y se asume que los efectos de las variables predictoras son constantes a través de las comparaciones.

En el estudio del DIF, si el efecto del grupo (G) resulta estadísticamente significativo y el efecto de la interacción habilidad x grupo (XG) no, entonces el ítem presentaría DIF uniforme. Por el contrario si la interacción XG resulta estadísticamente significativa el ítem presentaría DIF no-uniforme. Estas hipótesis pueden someterse a comprobación mediante el estadístico de Wald (Hosmer y Lemeshow, 1989).

Análisis Discriminante Logístico

Miller y Spray (1993) proponen el análisis discriminante logístico (ADL) como una alternativa en la evaluación del DIF, más sencilla y de menor costo computacional que el análisis de RLP. Básicamente este procedimiento es un análisis de regresión logística para variables dicotómicas donde se toma como variable dependiente el grupo de pertenencia (focal y referencia) y como variables predictoras, la respuesta al ítem (U), la habilidad y la interacción habilidad y respuesta al ítem (XU). El modelo de ADL para el DIF no-uniforme se formula en los siguientes términos:

$$P(G|X,U) = \frac{\exp(G-1)(\beta_0 + \beta_1 X + \beta_2 U + \beta_3 XU)}{1 + \exp(\beta_0 + \beta_1 X + \beta_2 U + \beta_3 XU)}$$

El ADL permite evaluar, al igual que RLP, la presencia tanto de DIF uniforme como no-uniforme. Así, si el efecto de la variable respuesta al ítem (U) resulta estadísticamente significativo y la interacción no, el ítem mostraría DIF uniforme. Si el efecto de la interacción habilidad x respuesta al ítem (XU) resulta significativo el ítem presentaría DIF no-uniforme.

Método

Los procedimientos de regresión logística politómica y análisis discriminante logístico se comparan sometiendo a análisis distintos tipos de tests en los que se han simulado diferentes condiciones de DIF. En todos los casos se ha trabajado con un test de 30 ítems con cinco categorías de respuesta, bajo el modelo de respuesta graduada de Samejima (1969). Se simularon tres condiciones de cantidad de DIF no uniforme (0.5, 1.0 y 1.5), que se generaron usando como medida el área entre las funciones de respuesta al ítem de los dos grupos (focal y de referencia) (Cohen, Kim y Baker, 1993). El DIF fue simulado siempre en el mismo sentido, favoreciendo al grupo de referencia, y constante en todas las categorías del ítem. En cada uno de los tests bajo estudio el número de ítems en los que se manipuló DIF fue de 6, es decir, un 20% de los ítems funcionaban diferencialmente. Además, a efectos de estudiar la tasa de error tipo I de ambos procedimientos, se sometió a análisis un test sin DIF.

Se han utilizado dos tamaños muestrales de 500 y 1000 sujetos tanto para el grupo de focal como para el grupo de referencia. La habilidad de los sujetos se generó aleatoriamente según una distribución normal tipificada en el intervalo de -3 a +3.

Las matrices de datos (sujetos x ítems) se simularon bajo el modelo de respuesta graduada de Samejima (1969). A fin de encontrar resultados estables se simularon 50 réplicas por condición.

El DIF se evaluó ajustando el modelo de regresión logística politómica (RLP) asumiendo idénticas pendientes a través de las categorías y ajustando el modelo discriminante logístico (ADL). En ambos casos el modelo analizado incluía todos los términos. Cuando se utilizó RLP el modelo ajustado incluía X, G y la interacción X x G; por otro lado, cuando se utilizó ADL, el modelo incluía X, U y la interacción X x U. Se utilizó el estadístico de Wald para la identificación del DIF. El nivel de confianza se estableció al 95%. Un resumen de las condiciones utilizadas en este estudio se presenta en la tabla 1.

Resultados

Con el objetivo de evaluar la eficacia de los dos procedimientos de detección del DIF, se ha tenido en cuenta el porcentaje de ítems con DIF correctamente identificados (IC) y el porcentaje de ítems que sin presentar DIF han sido detectados como tales, es decir, el porcentaje de falsos positivos (FP). La tabla 2 presenta el porcentaje de IC en cada procedimiento (RLP y ADL) en función del tamaño muestral y la cantidad de DIF.

El número de ítems con DIF correctamente identificados, independientemente del tamaño muestral y cantidad de DIF, fue mayor cuando se utilizó ADL que cuando se utilizó RLP. Tanto ADL como RLP detectaron mayor número de ítems con DIF cuando el tamaño muestral fue mayor (N=1000), y cuando la cantidad de DIF fue alta (área entre las curvas de respuesta al ítem de 1.00 o mayor). En las condiciones más favorables, mayor tamaño muestral de los grupos focal y de referencia y mayor cantidad de DIF, la tasa de IC fue del 87% cuando se empleó ADL frente a un 42.33% cuando se utilizó RLP. En las condiciones de menor tamaño muestral y menor cantidad de DIF la tasa de IC fue de 10.67% cuando se utilizó ADL frente al 5.33% cuando se utilizó RLP (ver tabla 2).

La tabla 3 presenta el porcentaje de FP en cada procedimiento y en función del tamaño muestral y la cantidad de DIF.

Tabla 1
Resumen de las condiciones experimentales

| Tamaño muestral | Tamaño del test | Cantidad de DIF no uniforme | Número de ítems con DIF en el Test | Procedimientos de detección del DIF | Réplicas |
|-----------------|-----------------|-----------------------------|------------------------------------|-------------------------------------|----------|
| 500 | 30 | 0.5 | 6 (20%) | RLP | 50 |
| 1000 | (m=5) | 1 | | ADL | |
| | | 1.5 | | | |

Tabla 2
Porcentaje de IC en las distintas condiciones manipuladas

| Tamaño Muestral | Cantidad de DIF | RLP | ADL |
|-----------------|-----------------|-------|-------|
| 500 | 0.5 | 5.33 | 10.67 |
| | 1.0 | 13.00 | 41.67 |
| | 1.5 | 32.67 | 74.00 |
| 1000 | 0.5 | 8.00 | 17.33 |
| | 1.0 | 16.67 | 57.00 |
| | 1.5 | 42.33 | 87.00 |

El porcentaje de falsos positivos en la situación de no DIF, es decir, la tasa de error Tipo I se mantuvo cerca de los niveles nominales para ambos procedimientos RLP y ADL, siendo ligeramente inferior cuando se aplicó ADL. Cuando el tamaño muestral fue de 500 sujetos la tasa de error Tipo I fue del 4.87% para ADL y del 5.40% cuando se utilizó RLP. Cuando el tamaño muestral fue de 1000 sujetos, la tasa de error Tipo I fue de 4.00% para ADL y de 4.33% para RLP.

El porcentaje de falsos positivos en las situaciones de tests con DIF, no muestran una tendencia clara de superioridad de un procedimiento sobre otro. Cuando el tamaño muestral fue N=500, la tasa de FP fue ligeramente mayor para RLP que para ADL. Por el contrario, cuando el tamaño muestral fue mayor, la tasa de FP fue ligeramente menor para RLP que para ADL.

| Tamaño Muestral | Cantidad de DIF | RLP | ADL |
|-----------------|-----------------|------|------|
| 500 | 0 | 5.40 | 4.87 |
| | 0.5 | 5.58 | 4.83 |
| | 1.0 | 5.67 | 5.42 |
| | 1.5 | 6.67 | 6.17 |
| 1000 | 0 | 4.33 | 4.00 |
| | 0.5 | 5.25 | 5.58 |
| | 1.0 | 6.17 | 7.50 |
| | 1.5 | 8.08 | 9.17 |

En cuanto a las condiciones de DIF, en las situaciones de menor tamaño muestral y menor cantidad de DIF e independientemente del procedimiento de detección del DIF utilizado, el porcentaje de FP se mantuvo cerca de los niveles nominales. El número de falsos positivos se incrementó cuando aumentó el tamaño muestral de los grupos focal y de referencia, y la cantidad de DIF, situándose por encima del 5% en ambos procedimientos. Así, cuando N=1000 y la cantidad de DIF fue 1,50, el porcentaje de FP se situó en el 8.08% para RLP y en el 9.17% para ADL (ver tabla 3).

Conclusiones

En resumen, los resultados encontrados en este estudio apuntan que:

- ADL identificó correctamente mayor porcentaje de ítems con DIF que RLP
- Tanto ADL como RLP resultaron más efectivas en la detección del DIF cuando el tamaño muestral y la cantidad de DIF fueron mayores (N=1000 y cantidad de DIF=1.5)
 - La tasa de error Tipo I se mantuvo cerca de los niveles nominales para ambos procedimientos
 - En las condiciones de DIF, el número de FP fue similar en ambos procedimientos.
 - En las condiciones de mayor tamaño muestral y mayor cantidad de DIF el porcentaje de FP fue superior al 5%.

Referencias

- Agresti, A. (1984). *Analysis of Ordinal Categorical Data*. New York, NY: Wiley.
- Agresti, A. (1990). *Categorical Data Analysis*. New York, NY: Wiley.
- Angoff, W.H. (1993). Perspectives on differential item functioning methodology. En P.W. Holland y H. Wainer (Eds.), *Differential Item Functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, A.S., Kim, S.H. y Baker, E. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement*, 17, 335-350.
- De Ayala, R.J., Kim, S.H., Stapleton, L.M. y Dayton, C. (1999). *A Reconceptualization of Differential Item Functioning*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada.
- French, A.W. y Miller, T.R. (1996). Logistic Regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement*, 33, 315-332.
- Gómez, J. e Hidalgo, M.D. (1997a). Evaluación del funcionamiento diferencial en ítems dicotómicos: Una revisión metodológica. *Anuario de Psicología*, 74, 3-32.
- Gómez, J. e Hidalgo, M.D. (1997b). *A comparison of two procedures of ability purification on the detection of differential item functioning using multinomial logistic regression*. Poster presented at the 10th European Meeting of the Psychometric Society, Santiago de Compostela, Spain.
- Hidalgo, M.D. (1995). *Evaluación del funcionamiento diferencial del ítem en ítems dicotómicos y politómicos: Un estudio comparativo*. Tesis Doctoral no publicada. Murcia: Universidad de Murcia.
- Hidalgo, M.D. y Gómez, J. (1999). Técnicas de detección de funcionamiento diferencial en ítems politómicos. *Metodología de las Ciencias del Comportamiento*, 1, 39-60.
- Hosmer, D.W. y Lemeshow, S. (1989). *Applied Logistic Regression*. New York, NY: Wiley.
- Mellenbergh, G.J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics*, 7, 105-108.
- Miller, T.R. y Spray, J.A. (1993). Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement*, 30, 107-122.
- Millsap, R.E. y Everson, H.T. (1993). Methodology Review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, 17, 297-334.
- Potenza, M.T. y Dorans, N.J. (1995). DIF assessment for polytomously scored items: A framework for classification and evaluation. *Applied Psychological Measurement*, 19, 23-37.
- Rogers, H.J. y Swaminathan, H. (1993). A comparison of Logistic Regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, 17, 105-116.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph Supplement*, 17.
- Swaminathan, H. y Rogers, H.J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, 27, 361-370.