Uso se R en la validación de un modelo predictivo con aplicación a la enfermedad del asma

Alejandra Tapia Silva

Palabras clave: datos del asma; modelo logístico de efectos mixtos; desempeño predictivo; diagnósticos de influencia global y local; Método de Metropolis-Hastings; R

Resumen

En este trabajo se propone el uso de R para la validación de un modelo predictivo ampliamente usado en las áreas de las ciencias médicas. Esta validación se realiza a través del desarrollo de una metodología simultánea de la técnica de influencia local y global para el modelo logístico de efectos mixtos, con el objetivo de detectar observaciones influyentes que puedan afectar sus inferencias y desempeño predictivo. Para la implementación de la metodología se creó un código en R, utilizando funciones de paquetes tanto para el ajuste del modelo como para la obtención de las medidas del desempeño predictivo. Un estudio con datos reales de asma en niños y adolescentes, recolectados en un hospital público de São Paulo, Brasil, fue realizado para la aplicación. Los resultados muestran que esta metodología es útil para obtener un modelo predictivo preciso que proporcione evidencia científica cuando se toman decisiones médicas basadas en datos.

Introducción

El asma es una de las enfermedades crónicas más importante que afecta a millones de personas en todo el mundo. El asma se describe como una enfermedad heterogénea por la Iniciativa Global para el Asma (GINA: https://ginasthma.org) y se caracteriza como una inflamación crónica de las vías respiratorias. En las últimas décadas, la prevalencia del asma está aumentando en muchos países, especialmente entre los niños y adolescentes (GINA). Por lo tanto, estrategias basadas en la evidencia científica son cruciales para generar mejores medidas preventivas, así como un mayor acceso y adherencia a tratamientos que reduzcan la carga económica. La evidencia científica sobre el asma está fuertemente relacionada con el uso de modelos predictivos, que proporcionan información valiosa en las diferentes áreas de las ciencias médicas para la toma de decisiones basadas en datos. En esta dirección, uno de los modelos predictivos ampliamente utilizado para ajustar la presencia o ausencia de una enfermedad con datos agrupados corresponde al modelo logístico de efectos mixtos. Este modelo presenta desafíos estadísticos que tienen una fuerte implicación en los resultados y pueden comprometer las inferencias y predicciones y, consecuentemente, las conclusiones para la toma de decisiones basadas en datos. Así, una vez que el modelo se ha ajustado, es fundamental evaluar la calidad de éste, para comprobar la validez de su ajuste.

Metodología

En este trabajo se propone una metodología para la validación del ajuste de un modelo logístico con efectos mixtos, la cual consiste en el desarrollo simultáneo de las técnicas de influencia local y global que a menudo se utilizan por separado (ver Tapia et al. 2020 y referencias en el mismo), con el objetivo de detectar observaciones influyentes en las inferencias y desempeño predictivo. Además, de identificar observaciones que tengan un comportamiento demasiado diferente en relación a las demás; ver más en Tapia et al. (2020). Esta metodología es implementada con un código R, cuyo algoritmo es descrito a continuación.

Paso 1: Ajustar el modelo logístico de efectos mixtos utilizando la función `glmer` del paquete `{lme4}`. Calcular las medidas del desempeño predictivo: sensibilidad (Sen), especificidad (Esp) y porcentaje de clasificación correcta (PCC), utilizando las funciones `sensitivity`, `specificity` y `pcc` del paquete `{PresenceAbsence}`.

Paso 2: Fase I: Basados en la estimación de parámetros obtenida en Paso 1, implementar un código R para muestrear observaciones aleatorias desde la función de densidad condicional dada en ec. (6) de Tapia et al. (2020), a través del método de Metropolis-Hastings. Luego, basados en el método de Monte Carlo, con estas observaciones aproximar las esperanzas condicionales de las matrices dadas en ec. (7) y (8) de esta misma referencia.

Paso 3: Fase II: Con las matrices obtenidas en Paso 2, implementar un código R para calcular la medida de influencia global, junto con el punto de corte asociado, y detectar grupos y observaciones influyentes. Realizar un análisis post-eliminación, es decir, eliminar estos grupos y observaciones, y recalcular las estimaciones de los parámetros y medidas de Sen, Esp y PCC.

Paso 4: Fase III: Con las matrices obtenidas en Paso 2, implementar un código R para calcular la medida de influencia local, junto con el punto de corte asociado, y detectar observaciones influyentes. Realizar un análisis post-eliminación, es decir, eliminar estas observaciones, y recalcular las estimaciones de los parámetros y medidas de Sen, Esp y PCC. Tanto en el Paso 3 como 4 para calcular las medidas de influencia global y local se utilizaron algunas funciones para cálculo matricial del paquete `{matrixcalc}`.

Paso 5: Fase IV: Basados en los resultados de las estimaciones de los parámetros y medidas de Sen, Esp y PCC de la Fase II y Fase III, determinar los grupos y/o observaciones para un nuevo análisis post-eliminación.

Paso 6: Fase V: Basados en los resultados de la Fase IV realizar un nuevo y final análisis post-eliminación.

Aplicación

Un estudio con datos reales de asma en 362 niños y adolescentes, recolectados en un hospital público de São Paulo, Brasil, fue realizado para la aplicación; ver más información en Tapia et al. (2020). El objetivo de este estudio consistió en construir un modelo predictivo preciso para la probabilidad de que un paciente presente obstrucción fija de la vía aérea dado el grupo de gravedad del asma en el que fue clasificado y covariables de interés. Específicamente, la variable respuesta Y_{ij} corresponde al registro de que si los pacientes tenían ($Y_{ij} = 1$) o no ($Y_{ij} = 0$) una obstrucción fija de la vía aérea (FAO). Además, estos pacientes se clasificaron en cuatro grupos según su Gravedad del asma (Grupo 1 - Asma intermitente, Grupo 2 - Asma persistente, Grupo 3 - Asma persistente moderada, Grupo 4 - Asma grave persistente), incorporando esta variable como intercepto aleatorio, u_i. De esta forma, se asume un modelo logístico de efectos mixtos dado por: $Y_{ij}|u_i \sim Bernoulli(\pi_{ij})$, con $u_i \sim N(0, \sigma^2)$ y

$$log\Big(\frac{\pi_{ij}}{1-\pi_{ij}}\Big) = \beta_0 + \beta_1 Trat_{ij} + \beta_2 Eosi_{ij} + \beta_3 Alergia_{ij} + u_i, \quad j = 1, \ldots, n_i, i = 1, \ldots, 4, \tag{1}$$

donde *Trat* corresponde a la duración del tratamiento en años, *Eosi* a la presencia o ausencia de eosinofilia en los análisis de sangre y *Alergia* a la suma de todos los niveles de todos los factores que producen alergia siguiendo la prueba radioalergosorbente (RAST). El Cuadro 1 muestra las estimaciones, errores estándar (ES), valores-p, cambios relativos (CR) y medidas de Sen, Esp y PCC para el modelo inicial (1) ajustado. Este modelo es denominado inicial debido a que es ajustado con todas las observaciones del conjunto de datos.
Para validar el ajuste de este modelo, se aplica la metodología propuesta, donde los resultados obtenidos en la Fase V son mostrados en Modelo Fase V del Cuadro 1. Se puede observar que al eliminar los casos del Grupo 4 para $y_{ij} = 0$ y Grupo 3 para $y_{ij} = 1$, las covariables *Eosi* y *Alergia* dejan de ser significativas al 5 % y 10 %, respectivamente, y las medidas de Sen, Esp y PCC aumentan sustancialmente. Sin embargo, si se considera el Modelo Fase V reducido, es decir, sin esas covariables, la estimación de la varianza asociada a la Gravedad del asma aumenta y las medidas Sen, Esp y PCC disminuyen. Por lo tanto, estas covariables deben permanecer en Modelo Fase V, siendo éste el modelo predictivo final. En conclusión, con el uso de R se implementa una metodología que nos permite obtener un modelo preciso con mayor capacidad predictiva y algunos cambios inferenciales, proporcionando evidencia científica mejorada para los datos de la enfermedad del asma en la toma de decisiones médicas. Cabe destacar que esta metodología permite identificar situaciones que no podrían detectarse si utilizaramos estas técnicas por separado. Además, de identificar observaciones que tengan un comportamiento demasiado diferente en relación a las demás con las que se pueda realizar un escrutinio adicional.

Cuadro 1: Resultados de las estimaciones, errores estándar (ES), valores-p, cambios relativos (CR) y medidas de Sen, Esp y PCC para los modelos ajustados.

Casos eliminados	Efecto	Estimación (ES)	valor-p	CR	Sen	Esp	PCC
Modelo inicial							
Ninguno	Intercepto	-4.0430 (0.7343)	<0.0001	-	0.6969	0.7598	0.7541
	Tratamiento	0.1871 (0.0579)	0.0012	-			
	Eosinofilia	0.1101 (0.0481)	0.0220	-			
	Alergia	-0.7006 (0.4026)	0.0817	-			
	Gravedad del asma	0.5417	-	-			
Modelo Fase V							
Group 4 - $y_{ij} = 0$	Intercept	-4.7920 (3.2765)	0.1435	18.5261	0.8750	0.8860	0.8851
Group 3 - $y_{ij} = 1$	Tratamiento	0.2987 (0.1181)	0.0114	59.6048			
	Eosinofilia	0.0916 (0.0863)	0.2883	16.7512			
	Alergia	-0.4026 (0.7687)	0.6004	42.5413			
	Gravedad del asma	5.5658	-	927.5726			
Modelo Fase V reducido							
Group 4 - $y_{ij} = 0$	Intercepto	-4.2963 (3.2379)	0.1845	6.2670	0.8333	0.8382	0.8378
Group 3 - $y_{ij} = 1$	Tratamiento	0.2869 (0.1159)	0.0133	53.3056			
	Gravedad del asma	5.7493	-	5121.0510			

Tapia, A., Giampaoli, V., Leiva, V., and Lio, Y. (2020). "Data-Influence Analytics in Predictive Models Applied to Asthma Disease." *Mathematics*, 8, 1587. https://doi.org/10.3390/math8091587.

GINA. The Global Strategy for Asthma Management and Prevention; GINA Report: Fontana, WI, USA, (2020). Available online: https://ginasthma.org/gina-reports (accessed on 11 September 2020).

Alejandra Tapia Silva
Universidad Católica del Maule, Chile
alejandraandreatapiasilva@gmail.com