Regresión Logística en la Investigación Médica

Carlos

Abril 2025

1. Resumen

Las técnicas de regresión son versátiles en su aplicación a la investigación médica, ya que pueden medir asociaciones, predecir resultados y controlar los efectos de confusión. La regresión logística es una técnica eficiente y poderosa para analizar el efecto de un grupo de variables independientes sobre un resultado binario al cuantificar la contribución única de cada variable independiente.

Utilizando componentes de la regresión lineal reflejados en la escala logit, la regresión logística identifica iterativamente la combinación lineal más fuerte de variables con la mayor probabilidad de detectar el resultado observado. Consideraciones importantes incluyen seleccionar variables independientes relevantes, cumplir con los supuestos y elegir una estrategia adecuada de construcción del modelo.

Los supuestos básicos que deben cumplirse para la regresión logística incluyen independencia de errores, linealidad en el logit para variables continuas, ausencia de multicolinealidad y falta de valores atípicos influyentes. Se recomienda tener un número adecuado de eventos por variable para evitar el sobreajuste.

Las estrategias de construcción de modelos incluyen enfoques directo, secuencial/jerárquico y escalonado, cada uno con énfasis diferente. Antes de llegar a conclusiones definitivas, se debe cuantificar formalmente la validez interna y externa del modelo. El ajuste del modelo se evalúa mediante medidas como el logaritmo de verosimilitud y pruebas de bondad de ajuste. Los resultados se presentan habitualmente como razones de momios (OR) con intervalos de confianza del 95 %.

2. Tipos de Regresión

Existen distintos tipos de regresión dependiendo de los objetivos de investigación y el formato de las variables. La regresión lineal se usa comúnmente para resultados continuos y asume una relación lineal entre la variable dependiente y las independientes:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

donde:

- β_0 es la ordenada al origen.

- $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ representa el valor ponderado de las variables independientes.

La regresión logística, sin embargo, es preferida cuando el resultado es binario. La fórmula es:

$$P(\hat{Y}_i) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i}}$$

Esta transforma la salida de la regresión lineal a una probabilidad entre 0 y 1 mediante la escala logit:

$$\ln\left(\frac{\hat{Y}}{1 - \hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

3. Supuestos de la Regresión Logística

- **Independencia de errores:** Las observaciones deben ser independientes entre sí.
- **Linealidad en el logit:** Las variables continuas deben tener una relación lineal con el logit del resultado.
- **Ausencia de multicolinealidad:** No debe haber redundancia entre variables independientes.
- **Ausencia de valores atípicos influyentes:** Estos pueden distorsionar los coeficientes y reducir la validez del modelo.

4. Número de Variables a Incluir

Se recomienda seguir la regla práctica de tener entre 10 y 20 eventos por variable independiente. Incluir demasiadas variables con un tamaño de muestra pequeño puede causar sobreajuste y generar errores estándar elevados.

5. Estrategias de Construcción del Modelo

- **Directa:** Todas las variables se introducen simultáneamente.
- **Secuencial/Jerárquica:** Se introducen variables por orden de prioridad.
- **Escalonada (Stepwise):** Basada en criterios estadísticos, como selección hacia adelante o eliminación hacia atrás.

6. Validación del Modelo

6.1. Validación Interna

- Método *holdout*: dividir el conjunto en entrenamiento y prueba.
- *k-fold cross-validation*: subdividir la muestra en k subconjuntos.
- *Bootstraping*: remuestreo con reemplazo.

6.2. Validación Externa

Evaluar el modelo en una muestra diferente al conjunto original, para estimar su aplicabilidad clínica y robustez.

7. Interpretación del Modelo

7.1. Ajuste General del Modelo

Se evalúa mediante estadísticos como Chi-cuadrado, desviación residual, y prueba de bondad de ajuste de Hosmer-Lemeshow. Un buen ajuste implica poca diferencia entre observados y predichos.

7.2. Discriminación del Modelo

Se evalúa con:

- Tablas de clasificación.
- Área bajo la curva ROC (AUROC), donde 0.5 indica azar y 1.0 discriminación perfecta.

7.3. Resultados de las Variables Independientes

Los coeficientes se interpretan mediante razones de momios (odds ratios, OR):

$$OR = e^{\beta_i}$$

- Un OR de 1.5 implica que una variable aumenta en un 50 % las probabilidades del evento.
- Los OR ajustados consideran el efecto de las demás variables.
- Se reportan con intervalos de confianza del 95 %.

8. Conclusión

La regresión logística es una herramienta estadística robusta para analizar resultados binarios en investigación médica. Su correcto uso requiere comprensión teórica, cumplimiento de supuestos y validación adecuada.

Resumen

Las técnicas de regresión son versátiles en su aplicación a la investigación médica porque pueden medir asociaciones, predecir resultados y controlar efectos de variables de confusión. Como una de estas técnicas, la regresión logística es una forma eficiente y poderosa de analizar el efecto de un grupo de variables independientes sobre un resultado binario, cuantificando la contribución única de cada variable independiente. Utilizando componentes de la regresión lineal reflejados en la escala logit, la regresión logística identifica iterativamente la combinación más fuerte de variables con la mayor probabilidad de detectar el resultado observado. Consideraciones importantes al realizar una regresión logística incluyen la selección de variables independientes, asegurando que se cumplan los supuestos relevantes y eligiendo una estrategia de construcción de modelo apropiada. Para la selección de variables independientes, uno debe guiarse por factores como teoría aceptada, investigaciones empíricas previas, consideraciones clínicas y análisis estadísticos univariantes, con reconocimiento de variables de confusión potenciales que deben tenerse en cuenta. Los supuestos básicos que deben cumplirse para la regresión logística incluyen independencia de errores, linealidad en el logit para variables continuas, ausencia de multicolinealidad y falta de valores atípicos altamente influyentes. Además, debe haber un número adecuado de eventos por variable independiente para evitar un modelo sobreajustado, con reglas generalesrecomendadas que van de 10 a 20 eventos por covariable. En cuanto a estrategias de construcción de modelos, hay tres tipos generales: directa/estándar, secuencial/jerárquica y paso a paso/estadística, cada una con un énfasis y propósito diferente. Antes de llegar a conclusiones definitivas a partir de los resultados de cualquiera de estos métodos, se debe cuantificar formalmente la validez interna del modelo (es decir, replicabilidad dentro del mismo conjunto de datos) y la validez externa (es decir, generalizabilidad más allá de la muestra actual). El ajuste general del modelo de regresión logística a los datos de la muestra se evalúa utilizando diversas medidas de bondad de ajuste, donde un mejor ajuste se caracteriza por una menor diferencia entre los valores observados y predichos por el modelo. También se recomienda el uso de estadísticas de diagnóstico para evaluar aún más la adecuación del modelo. Finalmente, los resultados para las variables independientes suelen informarse como razones de momios (odds ratios, OR) con intervalos de confianza del 95 % (IC).

Referencias

1. Darlington RB. Regression and Linear Models. Columbus, OH: McGraw-Hill Publishing Company, 1990.

2. Tabachnick BG, Fidell LS. Using Multivariate Statistics. 5th ed. Boston, MA: Pearson

Education, Inc., 2007.

3. Hosmer DW, Lemeshow SL. Applied Logistic Regression. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2000.

4. Campbell DT, Stanley JC. Experimental and Quasi-experimental Designs for Research. Boston, MA: Houghton Mifflin Co., 1963.

5. Stokes ME, Davis CS, Koch GG. Categorical data analysis using the SAS system (2nd ed). Cary, NC: SAS Institute, Inc., 2000.

6. Newgard CD, Hedges JR, Arthur M, Mullins RJ. Advanced statistics: the propensity score–a method for estimating treatment effect in observational research. Acad Emerg Med. 2004; 11:953–61.

7. Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research–part 2: multiple imputation. Acad Emerg Med. 2007; 14:669–78.

8. Allison PD. Logistic Regression Using the SAS System: Theory and Application. Cary, NC: SAS Institute, Inc., 1999.

9. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49:1373–9.

10. Agresti A. An Introduction to Categorical Data Analysis. Hoboken, NJ: Wiley, 2007.

11. Feinstein AR. Multivariable Analysis: An Introduction. New Haven, CT: Yale University Press, 1996.

12. Altman DG, Royston P. What Do We Mean by Validating a Prognostic Model? Stats Med. 2000; 19:453–73.

13. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Quebec, Canada, August 20–25, 1995. 1995:1137–43.

14. Efron B, Tibshirani R. An Introduction to the Bootstrap. New York: Chapman & Hall, 1993.

15. Miller ME, Hiu SL, Tierney WM. Validation techniques for logistic regression models. Stat Med. 1991; 10:1213–26.

16. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. Stat Med. 1997; 16:965–80.

17. Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. Stat Med. 2002; 21:3789–801.

18. Zou KH, OMalley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 2007; 115:654–7.