Resumen en Extenso del Artículo: *Logistic Regression: a brief primer: Jill C. Stoltzfus*

Carlos

Contents

1 Resumen

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, por otra parte la regresión logística identifica iterativamente la combinación lineal más fuerte de variables con la mayor probabilidad de detectar el resultado observado.

Es importante considerar en una regresión logística la *selección de variables independientes* para esto uno debe guiarse por factores tales como teoría existente, investigaciones empíricas previas, consideraciones clínicas y análisis estadísticos univariados, reconociendo las posibles variables de confusión que deben ser consideradas.

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 fuertemente influyentes. Adicionalmente,

- existencia de un número adecuado de eventos por variable independiente para evitar un modelo sobreajustado, con un mínimo comúnmente recomendado de "reglas prácticas" que van de 10 a 20 eventos por covariable.

Respecto a las estrategias de construcción de modelos, los tres tipos generales son:

- directa/estándar,
- secuencial/jerárquica y
- por pasos/estadística,

cada uno 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, su replicabilidad dentro del mismo conjunto de datos) y su validez externa (es decir, su generalizabilidad más allá de la muestra actual).

El ajuste general del modelo de regresión logística a los datos de muestra se evalúa utilizando varias *medidas de bondad de ajuste*, donde un mejor ajuste se caracteriza por una menor diferencia entre los valores observados y los valores 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 reportarse como *razones de momios* (odds ratios, ORs) con *intervalos de confianza* (IC) del 95%.

2 Importancia de la regresión en investigación

La **regresión** es un método valioso de investigación debido a su versátil aplicación en diferentes contextos de estudio. Por ejemplo, se puede utilizar para examinar asociaciones entre un resultado y varias variables independientes (también comúnmente conocidas como covariables, predictores o variables explicativas)[1], o para determinar qué tan bien puede predecirse un resultado a partir de un conjunto de variables independientes[1, 2]. Adicionalmente, uno puede estar interesado en controlar el efecto de variables independientes específicas, particularmente aquellas que actúan como variables de confusión (es decir, cuya relación tanto con el resultado como con otra variable independiente oscurece la relación entre esa variable independiente y el resultado)[1, 3]. Esta última aplicación es especialmente útil en contextos donde no es posible asignar aleatoriamente sujetos a grupos de tratamiento, como sucede en investigaciones observacionales. Con asignación aleatoria, normalmente se puede ejercer un control adecuado sobre las variables de confusión, ya que los grupos aleatorizados tienden a tener una distribución equitativa o balanceada de dichas variables[4].

3 Regresión Logística

Existen diferentes tipos de regresión, dependiendo de los objetivos de investigación y del formato de las variables, siendo la regresión lineal una de las más utilizadas. La *regresión lineal* analiza resultados continuos (es decir, aquellos que pueden sumarse, restarse, multiplicarse o dividirse de manera significativa, como el peso) y asume que la relación entre el resultado

y las variables independientes sigue una forma funcional determinada. Sin embargo, generalmente es más deseable determinar la influencia de múltiples factores al mismo tiempo, ya que de este modo se pueden observar las contribuciones únicas de cada variable después de controlar por los efectos de las demás. En este caso, la regresión lineal multivariada es la opción adecuada.

La ecuación básica para la regresión lineal con múltiples variables independientes es:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i. \tag{1}$$

Los componentes de esta ecuación son los siguientes:

- $\hat{Y}$ es el resultado continuo estimado.
- $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ es la ecuación de regresión lineal para las variables independientes del modelo, donde:
  - β_0 es la ordenada al origen o punto en el que la línea de regresión toca el eje vertical Y. Se considera un valor constante.
  - $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ es el valor de cada variable independiente (X_i) ponderado por su respectivo coeficiente beta (β). Los coeficientes beta determinan la pendiente de la línea de regresión, cuanto mayor sea el coeficiente beta, más fuerte es la contribución de dicha variable al resultado.

Para una variable binaria, como la mortalidad, la regresión logística es el método usualmente elegido, la regresión logística puede incluir una o múltiples variables independientes, aunque examinar múltiples variables es generalmente más informativo, ya que permite revelar la contribución única de cada variable ajustando por las demás. La regresión logística tiene ecuación:

$$\text{Probabilidad del resultado}(\hat{Y}_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}. \tag{2}$$

4 Transformación logística y elección de variables independientes

Un aspecto importante de la regresión logística es que conserva muchas características de la regresión lineal en su análisis de resultados binarios. Sin embargo, existen diferencias clave entre las dos ecuaciones:

1. $\hat{Y}_i$ representa la probabilidad estimada de pertenecer a una de las dos categorías binarias del resultado (categoría i) en lugar de representar un resultado continuo estimado.
2. $e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}$ representa la ecuación de regresión lineal para las variables independientes expresadas en la escala *logit*.

La razón de esta transformación *logit* radica en los parámetros básicos del modelo de regresión logística, un resultado binario expresado como probabilidad debe estar entre 0 y 1 [1]. La escala logit resuelve este problema al transformar matemáticamente la ecuación de regresión lineal original para producir el logit (o logaritmo natural) de las razones de momios (odds) de estar en una categoría ($\hat{Y}$) frente a la otra categoría ($1 - \hat{Y}$):

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \tag{3}$$

En el contexto de estas ecuaciones, la regresión logística identifica, mediante ciclos iterativos, la combinación lineal más fuerte de variables independientes que aumente la probabilidad de detectar el resultado observado —un proceso conocido como estimación de máxima verosimilitud [2, 3]. Para asegurar que la regresión logística produzca un modelo preciso, se deben considerar factores críticos como la selección de variables independientes y la elección de la estrategia de construcción del modelo.

4.1 Variables independientes

1. **Criterios de selección.** Seleccionar cuidadosamente las variables independientes es un paso esencial. Aunque la regresión logística es bastante flexible y permite distintos tipos de variables (continuas, como la edad; ordinales, como escalas de dolor analógico visual; y categóricas, como la raza), siempre debe justificarse la selección de variables utilizando teoría bien establecida, investigaciones previas, observaciones clínicas, análisis estadístico preliminar, o una combinación razonada de estas opciones.

   Por ejemplo, se podría comenzar con un gran número de variables independientes potenciales con base en estudios previos y experiencia clínica en el departamento de urgencias, y luego analizar diferencias entre grupos mediante estadística univariada con un nivel de error tipo I más relajado (por ejemplo, $p \leq 0.25$) para determinar qué variables deben incluirse en el modelo de regresión logística. Usar un valor de p menos estricto en esta etapa protege contra la exclusión de variables potencialmente importantes. Alternativamente, uno podría optar por incluir todas las variables independientes relevantes independientemente de sus resultados univariados, ya que puede haber variables clínicamente importantes que merezcan inclusión a pesar de su desempeño estadístico. Sin embargo, siempre debe tenerse en cuenta que incluir demasiadas variables independientes en el modelo puede conducir a un resultado matemáticamente inestable, con menor capacidad de generalización más allá de la muestra actual del estudio.[1]

   Una parte clave del proceso de selección de variables es reconocer y considerar el papel de los posibles factores de confusión. Como se describió previamente, las variables de confusión son aquellas cuya relación tanto con el resultado como con otra variable independiente oculta la verdadera asociación entre esa variable independiente y el resultado.[2]

[1] Referencias 2,3

[2] Referencias 1,3

Por ejemplo, el nivel socioeconómico (SES) podría confundir la relación entre la raza y las visitas anuales a emergencias, debido a su asociación con ambas variables (es decir, ciertos grupos raciales tienden a estar sobrerrepresentados en algunas categorías de SES y los pacientes más pobres pueden usar más frecuentemente los servicios de urgencias). No obstante, como este tipo de asociaciones causales no siempre son evidentes, se recomienda evaluarlas formalmente durante el proceso de selección de variables, a fin de garantizar que se modelen adecuadamente. Los diagramas de análisis de trayectorias pueden ser particularmente útiles en este sentido.[3]

Independientemente del método para seleccionar las variables independientes, deben cumplirse ciertos supuestos básicos al aplicar regresión logística. Un supuesto es la **independencia de los errores**, lo cual significa que todos los resultados del grupo de muestra deben ser independientes entre sí (por ejemplo, que no haya respuestas duplicadas). Si los datos incluyen mediciones repetidas u otros resultados correlacionados, los errores también estarán correlacionados y el supuesto se violará.[4] Existen otros métodos para analizar datos correlacionados mediante técnicas de regresión logística, pero van más allá del alcance de este artículo; para más información, los lectores pueden consultar a Stokes et al.,[5] Newgard et al.,[6] y Allison.[7]

Un segundo supuesto es la **linealidad en el logit** para las variables continuas independientes (por ejemplo, edad), lo que significa que debe existir una relación lineal entre estas variables y sus respectivos resultados transformados en logit. Hay diversas formas de verificar este supuesto, siendo una técnica común la creación de un término de interacción entre cada variable continua independiente y su logaritmo natural. Si alguno de estos términos es estadísticamente significativo, el supuesto se considera violado.[8]

Las soluciones incluyen codificación dicotómica de la variable independiente,[9] o su transformación estadística a otra escala.[10]

Un tercer supuesto es la **ausencia de multicolinealidad**, o redundancia entre variables independientes (por ejemplo, peso e índice de masa corporal [IMC] están correlacionados, por lo que no deben incluirse en el mismo modelo). Un modelo de regresión logística con variables independientes altamente correlacionadas usualmente genera errores estándar grandes para los coeficientes beta (o pendientes) estimados. La solución común es eliminar una o más variables redundantes.[11]

El supuesto final es la **ausencia de valores atípicos altamente influyentes**, es decir, casos en los que el resultado predicho para un miembro de la muestra difiere considerablemente de su valor real...

[3]Referencia 1
[4]Referencia 2
[5]Referencia 5
[6]Referencias 6,7
[7]Referencia 8
[8]Referencias 2,3
[9]Referencia 3
[10]Referencias 2,3
[11]Referencia 2

...resultado. Si hay demasiados valores atípicos, la precisión general del modelo puede verse comprometida. La detección de valores atípicos se realiza examinando los residuales (es decir, la diferencia entre los valores predichos y los resultados reales) junto con estadísticas diagnósticas y gráficas.[12] Luego, se puede comparar el ajuste general del modelo y los coeficientes beta estimados con y sin los casos atípicos. Dependiendo de la magnitud del cambio, uno podría conservar los valores atípicos cuyo efecto no sea dramático[13] o eliminar aquellos con una influencia particularmente fuerte sobre el modelo.[14]

Además de comprobar que se cumplan los supuestos anteriores, se puede considerar incluir términos de interacción que combinen dos o más variables independientes. Por ejemplo, es posible que la interacción entre la edad y la raza de los pacientes sea más importante para explicar un resultado que cualquiera de estas variables por separado[15] (por ejemplo, la relación entre la edad y la mortalidad relacionada con trauma varía entre asiáticos, blancos e hispanos). Sin embargo, los términos de interacción pueden complicar innecesariamente el modelo de regresión logística sin aportar mucho beneficio.[16] Por ello, se debe pensar cuidadosamente antes de incluirlos, obteniendo orientación de diagnósticos estadísticos (por ejemplo, observando cuánto cambian los coeficientes beta estimados, o pendientes, de una variable independiente al añadir otra al modelo), y evaluando si las interacciones tienen sentido clínico.[17]

2. Número de variables a incluir. Como parte del proceso de selección de qué variables independientes incluir, también se debe decidir cuántas. El reto es seleccionar el menor número posible de variables independientes que expliquen mejor el resultado sin descuidar las limitaciones del tamaño de muestra.[18] Por ejemplo, si se seleccionan 50 personas para el estudio y se incluyen 50 variables independientes en el análisis de regresión logística, el resultado es un modelo sobreajustado (y por tanto inestable). En términos generales, un modelo sobreajustado tiene coeficientes beta estimados para las variables independientes mucho mayores de lo que deberían ser, además de errores estándar más altos de lo esperado.[19] Este tipo de situación genera inestabilidad en el modelo porque la regresión logística requiere más resultados que variables independientes para poder iterar soluciones diferentes en busca del mejor ajuste a través del método de máxima verosimilitud.[20]

Entonces, ¿cuál es el número correcto de resultados para evitar un modelo sobreajustado? Aunque no existe un estándar universalmente aceptado, hay algunas "reglas generales" derivadas en parte de estudios de simulación. Una de estas reglas sugiere que por

[12]Referencias 2,3
[13]Referencia 3
[14]Referencias 2,3
[15]Referencia 3
[16]Referencias 2,3
[17]Referencia 3
[18]Referencias 2,3
[19]Referencia 3
[20]Referencias 2,3

cada variable independiente, debe haber al menos 10 resultados por cada categoría binaria (por ejemplo, vivo/muerto), siendo el resultado menos frecuente el que determina el número máximo de variables independientes.[21] Por ejemplo, en un estudio de mortalidad por sepsis, si se asume que 30 pacientes murieron y 50 sobrevivieron, el modelo podría acomodar, como máximo, tres variables independientes (ya que 30 es el resultado menos frecuente). Algunos estadísticos recomiendan una "regla general" aún más estricta de 20 resultados por variable independiente, dado que una relación más alta tiende a mejorar la validez del modelo.[22] Sin embargo, el tema no está completamente resuelto y algunos argumentan que menos de 10 resultados por variable pueden ser apropiados en ciertos contextos de investigación.[23]

Estrategias de Construcción del Modelo

Además de la cuidadosa selección de las variables independientes, se debe elegir el tipo adecuado de modelo de regresión logística para el estudio. De hecho, seleccionar una estrategia de construcción del modelo está estrechamente relacionado con la elección de variables independientes, por lo que estos dos componentes deben considerarse simultáneamente al planear un análisis de regresión logística.

Existen tres enfoques generales para la construcción del modelo que se aplican a las técnicas de regresión en general, cada uno con un énfasis y propósito diferente: directo (es decir, completo, estándar o simultáneo), secuencial (es decir, jerárquico) y paso a paso (es decir, estadístico). Estas estrategias de construcción no son necesariamente intercambiables, ya que pueden producir diferentes medidas de ajuste del modelo y diferentes estimaciones puntuales para las variables independientes a partir de los mismos datos. Por lo tanto, identificar el modelo apropiado para los objetivos del estudio es extremadamente importante.

El enfoque directo es una especie de valor por defecto, ya que introduce todas las variables independientes en el modelo al mismo tiempo y no hace suposiciones sobre el orden o la importancia relativa de dichas variables.[24] Por ejemplo, al analizar la mortalidad a 30 días en pacientes sépticos admitidos por el departamento de emergencias (ED), si se identifican 10 variables independientes para incluir, entonces las 10 se introducen en el modelo simultáneamente y tienen la misma importancia al inicio del análisis.

El enfoque directo es más adecuado si no existen hipótesis previas sobre cuáles variables tienen mayor relevancia que otras. De lo contrario, se puede considerar el uso de regresión secuencial/jerárquica, en la cual las variables se añaden secuencialmente para evaluar si mejoran el modelo de acuerdo a un orden predeterminado de prioridad.[25] Por ejemplo, se podría iniciar introduciendo la edad en el modelo, suponiendo que es el predictor más fuerte de mortalidad a 30 días en pacientes admitidos por sepsis, seguido de edad más comorbilidades, luego edad, comorbilidades y volumen de casos de sepsis en el ED, y así sucesivamente. Aunque este enfoque es útil para clarificar patrones causales entre variables independientes y resultados, puede volverse complejo conforme aumentan los patrones causales, dificultando

[21] Referencias 9,10
[22] Referencia 11
[23] Referencia 3
[24] Referencias 1,2
[25] Referencias 1,2

así la obtención de conclusiones definitivas sobre los datos en algunos casos.[26]

En contraste con los dos métodos anteriores, la regresión paso a paso identifica variables independientes que deben mantenerse o eliminarse del modelo con base en criterios estadísticos predefinidos que están influenciados por las características únicas de la muestra analizada.[27] Existen distintos tipos de técnicas paso a paso, incluyendo selección hacia adelante (por ejemplo, edad, comorbilidades, volumen de casos de sepsis en el ED, y otras variables independientes son introducidas una por una en el modelo para mortalidad por sepsis a 30 días, hasta que no se identifiquen más variables adicionales que contribuyan significativamente al resultado) y eliminación hacia atrás (por ejemplo, edad, comorbilidades, volumen de casos de sepsis en el ED, y otras variables se introducen todas simultáneamente en el modelo, y luego se eliminan una a una aquellas con contribuciones no significativas). con una contribución no significativa al resultado son eliminadas una por una hasta que sólo queden las variables estadísticamente significativas).[28] Otra estrategia de construcción del modelo que es conceptualmente similar a la regresión por pasos se llama "selección del mejor subconjunto", en la que se comparan modelos separados con diferentes números de variables independientes (por ejemplo, edad sola, edad más comorbilidades, comorbilidades más volumen de casos de sepsis en urgencias) para determinar el mejor ajuste según lineamientos preestablecidos.[29]

Una Nota de Precaución

Aunque la regresión por pasos se usa frecuentemente en la investigación clínica, su uso es algo controvertido porque se basa en una selección automatizada de variables que tiende a aprovechar factores aleatorios en una muestra dada.[30] Además, la regresión por pasos puede producir modelos que no parecen completamente razonables desde una perspectiva biológica.[31] Ante estas preocupaciones, algunos argumentan que la regresión por pasos se reserva mejor para el tamizaje preliminar o únicamente para pruebas de hipótesis,[32] como en casos de resultados novedosos y una comprensión limitada de las contribuciones de las variables independientes.[33] Sin embargo, otros señalan que los métodos por pasos no son en sí el problema (y de hecho pueden ser bastante efectivos en ciertos contextos); en cambio, el verdadero problema es una interpretación descuidada de los resultados sin valorar completamente los pros y contras de este enfoque. Por tanto, si uno elige crear un modelo por pasos, es importante validar posteriormente los resultados antes de sacar conclusiones. No obstante, debe destacarse que todos los tipos de modelos requieren validación formal antes de que se consideren definitivos para uso futuro, ya que se espera que los modelos funcionen mejor con la muestra original que con muestras subsiguientes.[34]

[26]Referencia 1
[27]Referencias 2,3
[28]Referencias 1,3
[29]Referencia 3
[30]Referencia 2
[31]Referencia 3
[32]Referencia 2
[33]Referencia 3
[34]Referencia 3

Validación Interna y Externa del Modelo

Al validar modelos de regresión logística, existen numerosos métodos entre los cuales elegir, cada uno más o menos apropiado según los parámetros del estudio como el tamaño de muestra. Para establecer la validez interna (confirmación de resultados del modelo con el mismo conjunto de datos), los métodos comunes incluyen: 1) el método de retención, o división de la muestra en dos subgrupos antes de la construcción del modelo, con el grupo de "entrenamiento" usado para crear el modelo de regresión logística y el grupo de "prueba" usado para validarlo;[35] 2) validación cruzada k-fold o división de la muestra en k subgrupos de igual tamaño para propósitos de entrenamiento y validación;[36] 3) validación cruzada "uno fuera" (leave-one-out), una variante del método k-fold donde el número de particiones es igual al número de sujetos en la muestra;[37] y 4) diferentes formas de bootstrapping (es decir, obtener submuestras repetidas con reemplazo de toda la muestra).[38]

Además de validar internamente el modelo, uno debería intentar validarlo externamente en un nuevo entorno de estudio como una prueba adicional de su viabilidad estadística y utilidad clínica.[39] Si los resultados de la validación interna o externa presentan alguna alerta (por ejemplo, el modelo tiene bajo rendimiento para cierto subgrupo de pacientes), se recomienda hacer ajustes al modelo según sea necesario, o definir explícitamente cualquier restricción para el uso futuro del modelo.[40]

Interpretación de los Resultados del Modelo

1. Evaluación del Ajuste General del Modelo. Una vez que se ha creado el modelo de regresión logística, se determina qué tan bien se ajusta a los datos de la muestra en su totalidad. Dos de los métodos más comunes para evaluar el ajuste del modelo son la prueba de chi-cuadrado de Pearson y la desviación residual. Ambas miden la diferencia entre los resultados observados y los resultados predichos por el modelo, donde un mal ajuste del modelo se indica mediante valores de prueba elevados, lo que señala una diferencia mayor. Sin embargo, la precisión de estas medidas depende de contar con un número adecuado de observaciones para los diferentes patrones de variables independientes.[41]

Otra medida comúnmente utilizada del ajuste del modelo es la prueba de bondad de ajuste de Hosmer-Lemeshow, que divide a los sujetos en grupos iguales (a menudo de 10) según su probabilidad estimada del resultado. El decil más bajo está compuesto por aquellos que tienen menor probabilidad de experimentar el resultado. Si el modelo tiene buen ajuste, los sujetos que experimentaron el resultado principal (por ejemplo, mortalidad por sepsis a los 30 días) caerán en su mayoría en los deciles de mayor riesgo. Un modelo con mal ajuste resultará en sujetos distribuidos de manera más uniforme a lo largo de los deciles de riesgo para ambos resultados binarios.[42]

[35] Referencias 12,13
[36] Referencia 13
[37] Referencia 13
[38] Referencias 13,14
[39] Referencias 12,15
[40] Referencia 15
[41] Referencias 3, 16, 17
[42] Referencias 2, 3

Las ventajas de las pruebas de Hosmer-Lemeshow incluyen su aplicación sencilla y facilidad de interpretación.[43] Las limitaciones incluyen la dependencia de las pruebas sobre cómo se definen los puntos de corte de los grupos[44] y los algoritmos computacionales utilizados,[45] así como una menor capacidad para identificar modelos con mal ajuste en ciertas circunstancias.[46] Otras alternativas menos comunes para evaluar el ajuste del modelo son descritas por Hosmer et al.[47] y Kuss.[48]

Aunque los índices de ajuste del modelo son componentes esenciales de la regresión logística, también se deben usar estadísticas de diagnóstico antes de sacar conclusiones sobre la adecuación del modelo final. Estas estadísticas ayudan a determinar si el modelo permanece intacto en todas las configuraciones posibles de las variables independientes.[49] Aunque una visión detallada de los métodos de diagnóstico excede el alcance de este artículo, se puede consultar a Hosmer y Lemeshow[50] para obtener más información.

Como forma de ampliar los resultados del ajuste del modelo y de las estadísticas diagnósticas, también se puede evaluar la capacidad del modelo para discriminar entre grupos. Las formas comunes de hacer esto incluyen 1) tablas de clasificación, donde la pertenencia a un grupo dentro de una categoría binaria del resultado se predice usando probabilidades estimadas y puntos de corte predefinidos,[51] y 2) el área bajo la curva característica operativa del receptor (AUROC), donde un valor de 0.5 significa que el modelo no es mejor que el azar para discriminar entre los sujetos que tienen el resultado y los que no, y un valor de 1.0 indica que el modelo discrimina perfectamente entre sujetos. El AUROC se usa a menudo cuando se desean considerar diferentes puntos de corte para la clasificación y así maximizar tanto la sensibilidad como la especificidad.[52]

2. Interpretación de los Resultados de Variables Individuales. Dentro del contexto del modelo de regresión logística, las variables independientes usualmente se presentan como razones de momios (ORs, por sus siglas en inglés).[53] Las ORs revelan la fuerza de la contribución de la variable independiente al resultado y se definen como las probabilidades de que ocurra el resultado ($\hat{Y}$) frente a que no ocurra...
$(1 - \hat{Y})$ para cada variable independiente. La relación entre la razón de momios (OR) y el coeficiente beta estimado de la variable independiente se expresa como OR = e^{β_i}. Con base en esta fórmula, un cambio de una unidad en la variable independiente multiplica la probabilidad del resultado por la cantidad contenida en e^{β_i}.[54]

Para un modelo de regresión logística con solo una variable independiente, la OR se considera "no ajustada" porque no hay otras variables cuya influencia deba ser ajustada o restada. Para fines ilustrativos, supongamos que el resultado es mortalidad intrahospitalaria

[43]Referencias 3, 16
[44]Referencias 10, 11
[45]Referencia 17
[46]Referencias 3, 16
[47]Referencias 16, 17
[48]Referencia 17
[49]Referencia 3
[50]Referencia 3
[51]Referencias 3, 21
[52]Referencias 3, 18
[53]Referencia 3
[54]Referencias 2,3

después de una lesión traumática, y que la única variable independiente es la edad del paciente, clasificada en mayores o menores de 65 años, con la categoría más reciente como grupo de referencia (o el grupo con el que se comparan todas las demás categorías de variables independientes). Una OR de 1.5 significa que para los pacientes mayores, las probabilidades de morir son 1.5 veces mayores que para los pacientes más jóvenes (grupo de referencia). Expresado de otro modo, hay un aumento del $(1.5-1.0)\times 100\% = 50\%$ en las probabilidades de morir en el hospital después de una lesión traumática para pacientes mayores frente a los más jóvenes.

En contraste, si el modelo de regresión logística incluye múltiples variables independientes, las OR ahora son "ajustadas" porque representan la contribución única de la variable independiente después de ajustar (o restar) los efectos de las otras variables en el modelo. Por ejemplo, si el escenario de mortalidad intrahospitalaria posterior a un trauma incluye edad más sexo, IMC y comorbilidades, la OR ajustada para la edad representa su contribución única a la mortalidad intrahospitalaria cuando las otras tres variables se mantienen constantes. Como resultado, las OR ajustadas suelen ser menores que sus contrapartes no ajustadas.

Interpretar las OR también depende de si la variable independiente es continua o categórica. Para las variables continuas, primero se debe identificar una unidad de medida significativa que exprese mejor el grado de cambio en el resultado asociado con esa variable independiente.[55] Usando la ilustración anterior de mortalidad intrahospitalaria con la edad mantenida en su escala continua original y seleccionando incrementos de 10 años como la unidad de cambio, uno interpretaría los resultados de la siguiente manera: "Por cada 10 años que envejece un paciente, las probabilidades de morir en el hospital después de una lesión traumática aumentan 1.5 veces, o un 50%".

Finalmente, los intervalos de confianza (IC) al 95% se informan rutinariamente junto con las OR como una medida de precisión (es decir, si los hallazgos probablemente se mantendrán en la población no observada). Si el IC cruza 1.00, es posible que no haya una diferencia significativa en esa población. Por ejemplo, si la OR de 1.5 para la edad tiene un IC del 95% de 0.85 a 2.3, no se puede afirmar de manera concluyente que la edad sea un contribuyente significativo a la mortalidad intrahospitalaria tras una lesión traumática.

4.2 Resumen

Las técnicas de regresión son versátiles en su aplicación a la investigación médica porque pueden medir asociaciones, predecir resultados y controlar los 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 al cuantificar la contribución única de cada variable independiente. Utilizando componentes de la regresión lineal reflejados en la escala logit, la regresión logística identifica de manera iterativa la combinación lineal más fuerte de variables con la mayor probabilidad de detectar el resultado observado.

Consideraciones importantes al realizar regresión logística incluyen la selección de variables independientes, asegurarse de que se cumplan los supuestos relevantes y elegir una

[55] Referencia 3

estrategia adecuada para la construcción del modelo. Para la selección de variables independientes, se deben considerar factores como la teoría aceptada, investigaciones empíricas previas, consideraciones clínicas y análisis estadísticos univariantes, reconociendo las posibles variables de confusión que deben ser tenidas en cuenta.

Los supuestos básicos que deben cumplirse para la regresión logística incluyen: independencia de los errores, linealidad en el logit para variables continuas, ausencia de multicolinealidad y ausencia de valores atípicos altamente influyentes. Además, debe haber un número adecuado de eventos por variable independiente para evitar un modelo sobreajustado, con una regla general recomendada que oscila entre 10 y 20 eventos por covariable.

Respecto a las estrategias de construcción del modelo, existen tres tipos generales: directa/estándar, secuencial/jerárquica y por pasos/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 (i.e., replicabilidad dentro del mismo conjunto de datos) y la validez externa (i.e., generalización más allá de la muestra actual).

El ajuste general del modelo de regresión logística a los datos de muestra se evalúa utilizando diversas medidas de bondad de ajuste, siendo mejor el ajuste cuanto menor sea la diferencia entre los valores observados y los valores predichos por el modelo. También se recomienda el uso de estadísticas de diagnóstico para evaluar adecuadamente el modelo. Finalmente, los resultados para las variables independientes se reportan típicamente como razones de momios (odds ratios, OR) con intervalos de confianza del 95% (ICs).

4.3 Tipos de regresión y fundamentos de la regresión logística

Existen diferentes tipos de regresión según los objetivos de la investigación y el formato de las variables, siendo la regresión lineal una de las más utilizadas. La regresión lineal analiza resultados continuos (es decir, aquellos que pueden sumarse, restarse, multiplicarse y dividirse significativamente, como el peso) y asume que la relación entre el resultado y las variables independientes sigue una línea recta (por ejemplo, a medida que aumentan las calorías consumidas, aumenta el peso).

Para evaluar el efecto de una sola variable independiente sobre un resultado continuo (por ejemplo, el efecto del consumo de calorías sobre el aumento de peso), se realizaría una regresión lineal simple. Sin embargo, normalmente es más deseable determinar la influencia de múltiples factores al mismo tiempo (por ejemplo, calorías consumidas, días de ejercicio por semana y edad en el aumento de peso), ya que esto permite ver las contribuciones únicas de cada variable después de controlar los efectos de las demás. En este caso, la regresión lineal multivariada es la opción adecuada.

La ecuación básica para la regresión lineal con múltiples variables independientes es:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \tag{4}$$

- β_0 es la ordenada al origen, o el punto en el que la línea de regresión toca el eje vertical Y. Se considera un valor constante.

- $\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$ representa el valor de cada variable independiente (X_i) ponderado por su coeficiente beta (β). Estos coeficientes indican la pendiente de la

línea de regresión o cuánto aumenta el resultado por cada unidad adicional en el valor de la variable independiente. Cuanto mayor es el coeficiente beta, mayor es la contribución de su variable independiente correspondiente al resultado.

A pesar de su uso común, la regresión lineal no es adecuada para ciertos tipos de resultados médicos. Para eventos binarios, como la mortalidad, la regresión logística es el método habitual de elección. Al igual que la regresión lineal, la regresión logística puede incluir una o varias variables independientes, siendo generalmente más informativa la evaluación de múltiples variables porque permite ver las contribuciones únicas de cada una tras ajustar por las otras.

La identificación de estas contribuciones en la regresión logística comienza con la siguiente ecuación:

$$\text{Probabilidad del resultado } (\hat{Y}_i) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}} \tag{5}$$

Esta ecuación contiene configuraciones similares para las variables independientes (X) y sus coeficientes beta (β) que la regresión lineal. No obstante, hay diferencias clave:

1. En regresión logística, $\hat{Y}_i$ representa la probabilidad estimada de estar en una categoría de resultado binario (por ejemplo, tener la enfermedad) frente a no estar en ella, en lugar de un resultado continuo estimado.

2. La expresión $e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_i X_i}$ representa la ecuación de regresión lineal para las variables independientes expresadas en escala logit, y no en el formato lineal original.

Esta transformación a escala logit es esencial en el modelo de regresión logística, ya que un resultado binario expresado como probabilidad debe estar entre 0 y 1. En cambio, las variables independientes podrían asumir cualquier valor. Si no se rectifica esta discrepancia, los valores predichos del modelo podrían caer fuera del rango de 0 a 1. La escala logit resuelve este problema al transformar matemáticamente la ecuación original de regresión lineal para producir el logit o logaritmo natural de las probabilidades de estar en una categoría ($\hat{Y}$) frente a la otra ($1-\hat{Y}$):

$$\ln\left(\frac{\hat{Y}}{1-\hat{Y}}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i \tag{6}$$

En el contexto de estas ecuaciones, la regresión logística identifica, mediante ciclos iterativos, la combinación lineal más fuerte de variables independientes que aumente la probabilidad de detectar el resultado observado—un proceso conocido como estimación por máxima verosimilitud.

Introducción

El artículo revisa detalladamente el modelo de regresión logística (RL), una técnica multivariable esencial para analizar relaciones entre variables independientes y una variable dependiente categórica, especialmente en investigaciones médicas. A través del análisis de 37 artículos científicos publicados entre 2000 y 2018 y seis libros de texto especializados, los

autores exploran conceptos clave de la RL, su aplicación, problemas frecuentes y propuestas para mejorar su uso.

Antecedentes Históricos de la Regresión Logística

La regresión logística tiene sus raíces en el siglo XIX, con los trabajos de Pierre François Verhulst, quien introdujo la *curva logística* para modelar el crecimiento poblacional. Sin embargo, fue en el siglo XX cuando su aplicación estadística tomó forma. En 1944, Joseph Berkson introdujo el *modelo logit* en el contexto de bioestadística, proponiéndolo como alternativa al modelo probit. La RL fue adoptada ampliamente en estudios biomédicos a partir de la década de 1960, gracias a su capacidad para manejar variables dicotómicas y ofrecer interpretaciones claras a través del odds ratio. En décadas recientes, la regresión logística se ha convertido en una herramienta fundamental para el análisis de datos en epidemiología, medicina clínica, y ciencias sociales.

Fundamentos del Modelo de Regresión Logística

La RL es ideal para predecir la probabilidad de ocurrencia de un evento binario (sí/no) y se basa en la transformación logística del *odds ratio* (razón de probabilidades). A diferencia de la regresión lineal, no requiere que las variables independientes sigan una distribución normal ni que la relación con la dependiente sea lineal.

La función logística

La función logística transforma la probabilidad de un evento en odds, y posteriormente en log-odds (*logit*), acotando los valores entre 0 y 1. Esto asegura interpretaciones coherentes para eventos dicotómicos. El modelo toma la forma:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Selección de variables

Se discuten criterios para seleccionar variables dependientes (ej. enfermedad/salud) y predictoras (factores clínicos), advirtiendo sobre el impacto negativo del sesgo de selección, multicolinealidad y tamaño de muestra reducido. El artículo destaca la importancia del conocimiento previo para seleccionar variables relevantes y evitar el sobreajuste.

Evaluación del Modelo

Evaluación general

Se emplean dos pruebas principales:

- **Razón de verosimilitudes (likelihood ratio test)**: compara un modelo completo con uno nulo, evaluando si los predictores mejoran significativamente la predicción.
- **Prueba de Hosmer-Lemeshow**: mide el ajuste entre valores observados y esperados por deciles de riesgo. Un valor $p > 0.05$ indica buen ajuste.

Evaluación de predictores

La significancia individual de cada predictor se evalúa con el **estadístico Wald**, basado en la relación entre el coeficiente estimado y su error estándar. También se puede usar la razón de verosimilitudes para cada predictor.

Exactitud Predictiva y Discriminación

- **Tabla de clasificación**: compara predicciones contra observaciones reales, generando métricas como sensibilidad, especificidad, precisión y valor predictivo.
- **Curva ROC (Receiver Operating Characteristic)**: representa la sensibilidad frente a 1 - especificidad. El área bajo la curva (AUC) cuantifica la capacidad discriminativa del modelo. Un AUC de 0.5 indica clasificación aleatoria, mientras que 1.0 representa clasificación perfecta.

Validación del Modelo

Se destaca la importancia de validar los modelos, ya sea de manera **interna** (con subconjuntos del mismo conjunto de datos) o **externa** (con nuevos datos). Se discuten métodos como *bootstrap*, *jackknife* y validación cruzada.

También se mencionan medidas como:

- R^2 **de Cox & Snell**
- R^2 **de Nagelkerke**

Estas proporcionan información sobre el poder explicativo del modelo, aunque no son equivalentes al R^2 clásico de regresión lineal.

Aplicación Práctica

Se presenta un estudio de caso usando RL para investigar factores que influyen en la decisión de partos por cesárea en mujeres embarazadas. Se evidenció que:

- El peso del bebé menor a 3.5 kg reduce la probabilidad de cesárea.
- Mujeres con más de tres partos tienen menor probabilidad de requerir cesárea.

El modelo mostró buen ajuste (Hosmer-Lemeshow $p = 0.65$), alta capacidad explicativa (R^2 de Nagelkerke = 0.723) y precisión predictiva del 82.9%.

Conclusiones

Los autores concluyen que, aunque la RL es una herramienta poderosa, su uso en la investigación médica presenta deficiencias notables:

- Tamaños de muestra insuficientes.
- Falta de validación del modelo.
- Reportes incompletos sobre el ajuste y la significancia de predictores.

Recomiendan mayor rigor metodológico y transparencia en los reportes, así como comparar RL con métodos más recientes como redes neuronales o árboles de decisión en futuras investigaciones.

References

[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**:953–961.

[7] Newgard CD, Haukoos JS. Advanced statistics: missing data in clinical research—part 2: multiple imputation. *Acad Emerg Med*. 2007; **14**:669–678.

[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**:1373–1379.

[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**:453–473.

[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. 1995:1137–1143.

[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**:1213–1226.

[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**:965–980.

[17] Kuss O. Global goodness-of-fit tests in logistic regression with sparse data. *Stat Med*. 2002; **21**:3789–3801.

[18] Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007; **115**:654–657.