Rev Esp Cardiol. 2011;64(6):501–507

Enfoque: Métodos contemporáneos en bioestadística (I)

Estrategias para la elaboración de modelos estadísticos de regresión

Eduardo Núñez[a,b,*], Ewout W. Steyerberg[c] y Julio Núñez[a]

[a] Servicio de Cardiología, Hospital Clínico Universitario, INCLIVA, Universitat de Valencia, España
[b] Cuore International, Reading, Pennsylvania, Estados Unidos
[c] Department of Public Health, Erasmus MC, Rotterdam, Países Bajos

Historia del artículo:
On-line el 29 de abril de 2011

Palabras clave:
Sobresaturación
Número de eventos por cada variable
Calibración
Discriminación

RESUMEN

Actualmente los modelos multivariables de regresión son parte importante del arsenal de la investigación clínica, ya sea para la creación de puntuaciones con fines pronósticos o en investigación dedicada a generar nuevas hipótesis. En la creación de estos modelos, se debe tener en cuenta: *a)* el uso apropiado de la técnica estadística, que ha de ser acorde con el tipo de información disponible; *b)* mantener el número de variables por evento no mayor de 10:1 para evitar la sobresaturación del modelo, relación que se puede considerar una medida grosera de la potencia estadística; *c)* tener presentes los inconvenientes del uso de los procesos automáticos en la selección de las variables, y *d)* evaluar el modelo final con relación a las propiedades de calibración y discriminación. En la creación de modelos de predicción, en la medida de lo posible se debe evaluar estas mismas medidas en una población diferente.

Regression Modeling Strategies

Keywords:
Overfitting
Number of events per variable
Calibration
Discrimination

ABSTRACT

Multivariable regression models are widely used in health science research, mainly for two purposes: prediction and effect estimation. Various strategies have been recommended when building a regression model: *a)* use the right statistical method that matches the structure of the data; *b)* ensure an appropriate sample size by limiting the number of variables according to the number of events; *c)* prevent or correct for model overfitting; *d)* be aware of the problems associated with automatic variable selection procedures (such as stepwise), and *e)* always assess the performance of the final model in regard to calibration and discrimination measures. If resources allow, validate the prediction model on external data.

Full English text available from: www.revespcardiol.org.

INTRODUCCIÓN

Los modelos de regresión multivariables se utilizan ampliamente en la investigación de ciencias de la salud. Con frecuencia, el objetivo en la recolección de datos obedece al afán de explicar las interrelaciones que existen entre ciertas variables o a determinar los factores que afectan a la presencia o ausencia de un episodio adverso determinado. Es ahí donde los modelos de regresión multivariables pasan a ser un instrumento útil, al suministrar una explicación matemática simplificada de dicha relación. El objetivo final será obtener un modelo simplificado que tenga sentido desde una perspectiva biológica, se atenga estrechamente a los datos disponibles y aporte predicciones válidas al aplicarlo a datos independientes.

Actualmente, investigadores con una formación no muy extensa en estadística a menudo se dedican al análisis de sus datos, en gran parte debido a la disponibilidad y la facilidad de uso de los paquetes informáticos más comunes. El objetivo de esta revisión es presentar recomendaciones prácticas sobre la forma de crear un modelo multivariable simplificado y efectivo. En la tabla 1 se indican los pasos generales que deben seguirse en la elaboración de cualquier modelo de regresión. Dada la limitación de espacio, esta revisión se centra sólo en aspectos prácticos.

ESTRUCTURA DE LOS DATOS Y TIPO DE ANÁLISIS DE REGRESIÓN

Los modelos de regresión tienen en general una estructura común que debe resultar familiar a la mayoría; generalmente siguen este patrón: respuesta = ponderación$_1$ × predictor$_1$ + ponderación$_2$ × predictor$_2$ + ... ponderación$_k$ × predictor$_k$ | término de error normal. La variable a explicar se denomina variable dependiente (o variable de evaluación). Cuando la variable dependiente es binaria, la literatura médica se refiere a ella en términos de eventos clínicos o episodios adversos. Los factores que explican la variable dependiente se denominan variables independientes, e incluyen la variable de interés (o variable explicativa) y las demás variables, a las que se denomina de manera genérica covariables. No es infrecuente que la función específica de estas covariables sea proporcionar un ajuste estadístico que

* Autor para correspondencia: Epidemiology and Statistical Services, Cuore International, 2914 Leiszs Bridge Rd. Reading, PA 19605, Estados Unidos.
Correo electrónico: enunezb@gmail.com (E. Núñez).

0300-8932/$ – see front matter
doi:10.1016/j.recesp.2011.01.019

Abreviaturas

DRA: diferencia de riesgo absoluto
EPV: eventos por variable
KM: método de Kaplan-Meier
MAR: ausencia de datos aleatoria
MCAR: ausencia de datos completamente aleatoria
MFP: polinomio fraccional multivariable
NMAR: ausencia de datos no aleatoria
NNT: número necesario a tratar

minimice el desequilibrio que pueda haber entre estos factores pronósticos y la variable explicativa con el evento. Sin embargo, a veces, la identificación de los factores predictores de la variable de evaluación constituye el objetivo principal del estudio, en cuyo caso cada variable independiente pasaría a ocupar la función de variable de interés. Los modelos de regresión pueden ser creados con diferentes finalidades (tabla 1), aunque en general estas pueden resumirse como objetivos de predicción y/o estimación del tamaño/significación del efecto observado[2]. En la tabla 2 se resumen las diferentes estrategias utilizadas en los modelos de predicción y de estimación del efecto.

Modelos de predicción

El objetivo principal de estos modelos es cuantificar la probabilidad de que ocurra el criterio de valoración (o episodio adverso), dados las condiciones o factores incluidos en el modelo, e idealmente reproducir estos resultados en poblaciones diferentes de la usada para su creación. Como ejemplo de ello, las reglas de predicción clínica derivadas del modelo ajustado a los datos de Framingham han demostrado, tras múltiples validaciones externas, predecir el riesgo absoluto de enfermedad coronaria en la población general[3]. Para estos tipos de modelos, el investigador debe establecer un equilibrio entre el grado de complejidad (y exactitud) y su simplicidad; en otras palabras, balancear la exactitud con que el modelo se ajusta matemáticamente a los datos usados para su derivación frente a su capacidad de generalizar las predicciones a poblaciones externas. Modelos complejos, por ejemplo aquellos con interacciones múltiples, número excesivo de predictores o predictores continuos que muestran un patrón de riesgo no lineal, tienden a reproducirse pobremente en poblaciones diferentes de la usada en su creación.

Se han propuesto varias recomendaciones para la elaboración de este tipo de modelos[2,4,5], las más importantes: *a)* incorporar la mayor cantidad posible de datos exactos, con distribución amplia en los valores de los predictores; *b)* imputar datos si es necesario, ya que mantener un adecuado tamaño de la muestra es de vital importancia; *c)* especificar de antemano la complejidad o el grado de no linealidad que deberá permitirse para cada predictor; *d)* limitar el número de interacciones e incluir solamente las preespecificadas y basadas en cierta plausibilidad biológica; *e)* seguir la regla de 10-15 eventos por variable dependiente (EPV) para criterios de valoración binarios, con el fin de evitar la sobresaturación del modelo, y si esto no es posible, utilizar técnicas para la simplificación (o reducción) de los datos; *f)* tener presentes los problemas asociados al uso de las estrategias de selección escalonada; en caso de utilizarlas, preferir la eliminación retrógrada y establecer el valor de p = 0,157 que es equivalente a usar el criterio AIC como regla de detención; en caso de muestras pequeñas, relajar aún más la regla de detención (p = 0,25-0,5) con el fin de no ignorar predictores importantes; utilizar el conocimiento previo como guía en la selección de las variables siempre que sea posible; *g)* verificar el grado de colinealidad entre los predictores importantes y utilizar la experiencia y la informacion que se tenga del tema para decidir qué predictores colineales deben ser incluidos en el modelo final; *h)* validar el modelo final con relación a parámetros de calibración y discriminación, preferiblemente utilizando técnicas de remuestreo *(bootstrapping),* e *i)* utilizar

Tabla 1
Pasos generales en el establecimiento de modelos de regresión multivariables

Determinación del objetivo del modelo
Predicción (modelos pronósticos)
Magnitud del efecto (o modelos explicativos)
Identificación del criterio de valoración verdadero
Minimización del error de clasificación de la variable de valoración
Se prefieren criterios de valoración «duros» para los modelos pronósticos
Si se usa una variable de valoración combinada, hay que asegurar que el sentido del efecto sea el mismo para todos los componentes
Considerar el uso de nuevos resultados como el número de días con vida y sin estar hospitalizado en los estudios de la insuficiencia cardiaca
Elección del método estadístico apropiado en función del resultado y el tipo de predicción
Continuo: regresión lineal
Binario: regresión logística
Binario con observaciones censuradas
Regresión proporcional de Cox
Regresión de supervivencia paramétrica
Riesgos en competencia
Variable de valoración longitudinal y tiempo hasta que ocurre el criterio de valoración: enfoque de modelado conjunto
Datos longitudinales con interés en variables de valoración intermedias: modelos de Markov multiestado
Creación del modelo adecuado, incluida la validación interna
Simplicidad frente a complejidad
Selección de las variables correctas. Precaución con el uso inapropiado de procedimientos escalonados. Uso de procedimientos retrógrados en vez de anterógrados. Ajuste de la regla de detención según el tamaño muestral[1]
Evitación de la sobresaturación (regla de EPV)
No dar por supuesta la linealidad de las variables continuas; transformarlas en caso necesario. Utilizar FPF o RCS para las funciones no lineales complejas
Evaluar el rendimiento del modelo
Validación interna (preferiblemente remuestreo). Parámetros a evaluar
Medidas de rendimiento global (R^2, puntuación de Brier)
Capacidad de discriminación (AUC, estadístico C, IDI, NRI)
Calibración (prueba de bondad de ajuste de Hosmer-Lemeshow, gráfico de calibración, pendiente de calibración, prueba de Gronnesby y Borgan, calibración general)
Validación externa. Los mismos parámetros, pero con datos externos
Necesidad de reducción del coeficiente de regresión
Si la evaluación de calibración muestra coeficientes excesivamente optimistas
Aplicar una reducción de ajuste a la pendiente de calibración o
Utilizar métodos de penalización más complejos como LASSO y MLE
Presentación de los resultados
Sin ajustar frente a ajustados
Medidas relativas (OR, HR)
Medidas absolutas (DRA, NNT)

DRA: diferencia de riesgo absoluto; AUC: área bajo la curva ROC; estadístico C: equivalente al AUC para datos censurados; EPV: número de eventos por variable; FPF: función polinómica fraccional; HR: razón de riesgos; IDI: índice de discriminación integrado; LASSO: *least absolute shrinkage and selection operator;* MLE:: estimación de máxima verosimilitud; NNT: número que es necesario tratar; NRI: índice de reclasificación neta; OR: *odds ratio;* R^2: medida de variación explicada; RCS: *splines* cúbicos restringidos.

Tabla 2
Diferencias de estrategia según la finalidad asignada al modelo

Objetivos	Consideraciones	Validación
Predicción	Simplicidad por encima de complejidad; lo que importa es la información integrada derivada de todos los predictores	Calibración y discriminación con el empleo de remuestreo *bootstrapping* (validación interna)
Predicción de un resultado de interés (puntuación pronóstica)	Evitar el uso excesivo de valores de corte	Reducción de los efectos principales (reducción lineal o penalización)
Identificación de predictores importantes	Imputación múltiple con valores perdidos > 5%	Calibración y discriminación de datos independientes (validación externa)
Estratificación según el riesgo	No basarse en exceso en procedimientos escalonados. Si hay demasiados predictores o un conocimiento insuficiente del área en cuestión, utilizar un algoritmo MFP	Actualizar los coeficientes en los nuevos datos en caso necesario
	Utilizar la regla de EPV para limitar el número de predictores. Agrupar los predictores en caso necesario (p. ej., con el uso de puntuación de propensión)	Convertir los coeficientes de regresión a puntuaciones con el empleo de algoritmos apropiados
	Aceptar grados de libertad adicionales para las interacciones y modelizar relaciones no lineales de predictores continuos si el conocimiento previo lo indica. Limitar las pruebas de términos individuales; considerar en su lugar la significación general	Evaluar la utilidad clínica de la puntuación derivada (nomograma, gráfico de puntuación pronóstica, análisis de curva de decisión)
		Si se piensa en la toma de decisiones clínicas, presentar también medidas absolutas (DRA o NNT)
Estimación del efecto	La importancia relativa de simplicidad y complejidad varía en función de la pregunta investigada	Cuando menos, debe presentarse la calibración y la capacidad de discriminación del modelo; la validación interna es un valor adicional
Explicativo: comprender el efecto de los predictores	Minimizar siempre la categorización de los predictores continuos	Si se piensa en la toma de decisiones, presentar también medidas absolutas (DRA o NNT)
Ajuste para los predictores en el diseño experimental con objeto de aumentar la precisión estadística	Imputación múltiple si los valores perdidos son > 5%	
Centrarse en la información independiente aportada por un predictor (o variable explicativa)	Uso del algoritmo MFP como método preferido de selección de variables	
	Tener siempre presente la regla básica de EPV. Agrupar los predictores en caso de tamaño muestral pequeño (p. ej., con el uso de puntuación de propensión)	
	Si el tamaño muestral lo permite, modelizar la relación no lineal de predictores continuos con no más de 4-5 grados de libertad (FPF o RCS). La decisión final debe basarse en la significación general (valor de p general)	
	Como estudio de generación de hipótesis, realizar pruebas de interacciones. Mantener los términos de interacción solamente cuando el valor de p general sea significativo	

DRA: diferencia de riesgo absoluto; EPV: número de eventos por variable; FPF: función polinómica fraccional; MFP: polinomio fraccional multivariable; NNT: número que es necesario tratar; RCS: *splines* cúbicos restringidos.

métodos para la simplificación o reducción de datos si la validación interna muestra predicciones excesivamente optimistas.

Modelos para la estimación del efecto

El fin de estos modelos es servir como instrumentos para la evaluación del tamaño y significación estadística del efecto en cuestión (expresado a través de la variable explicativa); de esta manera, se convierten en instrumentos útiles para la verificación de nuevas hipótesis. La mayoría de los artículos biomédicos publicados se basan en este tipo de modelos. Dado que hay poca preocupación por la simplicidad, el balance se decanta por desarrollar un modelo más exacto y complejo que refleje los datos disponibles. Sin embargo, aun en este tipo de modelos de regresión, hay que tener presentes las medidas recomendadas para evitar la sobresaturación e incluso, de ser necesario, aplicar métodos de simplificación (o reducción). Además, siempre es recomendable validar el modelo final con base en parámetros de calibración y discriminación.

Tipos de modelos en función de la estructura de los datos

Otra consideración que tener en cuenta al elaborar un modelo de regresión es elegir el método estadístico apropiado que se corresponda con el tipo de variable dependiente. Existen muchas variantes en la forma de obtener los datos, y no es infrecuente que los mismos datos puedan ser analizados con más de un método de regresión. En la tabla 1 se muestran los diferentes métodos de regresión recomendados para las variantes más comunes en relación a la estructura de los datos (basándose en un supuesto de error normal). Una explicación detallada de cada uno de estos métodos queda fuera del alcance de este artículo, por lo que sólo se presentarán los aspectos más importantes.

Análisis de regresión lineal

El principal supuesto es la linealidad de la relación entre la variable dependiente, que debe ser continua, y los predictores. Si esto no se cumple, se linealiza la relación ya sea transformando la variable o aplicando métodos no paramétricos.

Análisis de regresión logística

El modelo de regresión logística es apropiado cuando se trata de un criterio de valoración binario, sin tener en cuenta el momento en que esta variable ocurre. Lo único que necesitamos conocer acerca del criterio de valoración es si está presente o ausente en cada individuo al final del estudio. La estimación del efecto del tratamiento (o variable explicativa) se expresa mediante la estimación de la *odds ratio* (OR) ajustada por otros factores incluidos en el modelo como covariables. A veces la regresión logística se ha utilizado de manera inadecuada para analizar datos donde el tiempo hasta que ocurre el criterio de valoración representa una característica importante del diseño. Annesi et al[6] pusieron de manifiesto que, en comparación con el método de regresión de Cox, la regresión logística producía estimaciones similares y con una eficiencia relativa asintótica cercana a 1 solamente en estudios con seguimiento corto y tasa del criterio de valoración baja. En consecuencia, la regresión logística debe considerarse una alternativa a la regresión de Cox solamente cuando la duración del seguimiento de la cohorte sea corta o cuando la proporción de observaciones censuradas es mínima y similar en los dos niveles de la variable explicativa.

Diseño de tiempo hasta que ocurre el criterio de valoración

Los métodos de regresión de supervivencia se han diseñado para compensar la presencia de observaciones censuradas y, por lo tanto, constituyen el método de elección para el análisis de datos con este tipo de diseño. El método de riesgos proporcionales de Cox es el más comúnmente utilizado. La magnitud del efecto se expresa en unidades relativas, como razón de riesgos (HR). El supuesto principal es que la diferencia de riesgo instantáneo se mantenga constante durante el seguimiento. Se han propuesto varias alternativas no paramétricas a la regresión de Cox cuando el supuesto de proporcionalidad no se cumple[7].

Los métodos de supervivencia paramétricos se recomiendan en los siguientes casos: *a)* cuando la función de supervivencia o de riesgo basal es de interés principal; *b)* para obtener estimaciones más exactas en situaciones en que la forma de la función de riesgo basal es conocida por el investigador; *c)* como forma de estimar la diferencia de riesgo absoluto ajustada (DRA) y el número que es necesario tratar (NNT) en puntos específicos de tiempo a lo largo del seguimiento; *d)* cuando no se cumple el supuesto de proporcionalidad para la variable explicativa[8], y *e)* cuando es necesario extrapolar los resultados más allá de los datos observados.

En el análisis de supervivencia, cada sujeto, a lo largo del seguimiento, puede experimentar uno de varios tipos diferentes de criterios de valoración. Si la presencia de uno de ellos influye o impide la aparición del criterio de valoración de interés, se produce una situación de riesgos en competencia. Por ejemplo, en un estudio de pacientes con insuficiencia cardiaca aguda, se impide el reingreso hospitalario (como evento de interés) si el paciente fallece durante el seguimiento. La muerte es en este caso un episodio adverso competitivo, ya que impide que el paciente pueda llegar a reingresar. En escenarios de riesgos competitivos, varios estudios han demostrado que los métodos de supervivencia tradicionales, como la regresión de Cox y el método de Kaplan-Meier (KM) son inapropiados[9]. Se han propuesto otros métodos alternativos, como las funciones de incidencia acumulativa y el modelo de riesgos de subdistribución proporcional de Fine y Gray[10].

En resumen, estos métodos deben considerarse: *a)* cuando el criterio de valoración de interés es una variable de valoración intermedia y la muerte del paciente impide su aparición; *b)* cuando es necesario introducir un ajuste de una forma específica de muerte (como muerte cardiovascular) en función de otras causas de muerte, y *c)* para introducir un ajuste respecto a eventos que se encuentran en la vía causal entre la exposición y el evento en cuestión (generalmente un evento terminal). Por ejemplo, las intervenciones de revascularización pueden considerarse incluidas en esta categoría porque modifican la historia natural de la enfermedad en el paciente y, por lo tanto, influyen en la incidencia de mortalidad.

Mediciones repetidas con evento de supervivencia

En la investigación biomédica, muchos estudios tienen diseños que utilizan mediciones de variables continuas, repetidas y en el mismo sujeto. Por ejemplo, en los registros cardiovasculares, la información de cada paciente se obtiene en cada visita y se sigue recogiéndola durante el seguimiento; esta información puede incluir marcadores biológicos continuos, como el BNP, la fracción de eyección ventricular izquierda, etc. En este contexto, el investigador puede estar interesado en determinar la trayectoria del marcador a lo largo del tiempo, e identificar los factores que la explican, o quizá el interés principal sea determinar el efecto de la secuencia de mediciones del marcador en la mortalidad o simplemente utilizar la secuencia de mediciones del marcador como factor de ajuste para un tratamiento instaurado al inicio del estudio. Un problema frecuente y serio en esos estudios es que la información del marcador es incompleta, en muchos casos debido a la muerte del paciente, lo que ocasiona la interrupción prematura de las mediciones repetidas. A este mecanismo de valores perdidos se lo denomina censura informativa (o abandono informativo) y requiere una metodología estadística especial para su análisis[11–14]. El enfoque analítico para este tipo de diseño se denomina regresión de modelado conjunto, y se está empezando a aplicar en programas informáticos comerciales de estadística[15,16].

En un contexto diferente, cuando el objetivo es describir la evolución de un proceso en el que un sujeto pasa por una serie de estadios, los modelos de Markov multiestado se convierten en el instrumento de análisis adecuado[17,18]. La evolución natural de muchas enfermedades crónicas puede representarse mediante una serie de etapas sucesivas, que terminan en un «estado absorbente» (generalmente la muerte). Dentro de este modelo, los pacientes pueden avanzar a fases más avanzadas de la enfermedad, regresar a estadios menos severos o fallecer, lo cual permite al investigador determinar probabilidades de transición entre las distintas fases, determinar los factores que influyen en esas transiciones y estimar el valor predictivo de cada estadio con respecto al estadio final de muerte.

MANIPULACIÓN DE LOS DATOS

No es infrecuente que los datos requieran una depuración antes de iniciar el análisis estadístico. Hay tres puntos importantes que considerar en este caso:

1. Valores perdidos. Este es un problema universal en la investigación en ciencias de la salud. Se han diferenciado tres tipos de mecanismos[19]: valores perdidos completamente aleatorios (MCAR), valores perdidos aleatorios (MAR) y valores

Tabla 3
Mecanismos para la ausencia de datos

Mecanismo	Descripción	Ejemplo	Efectos
MCAR	Probabilidad de ausencia de datos no relacionada con datos observados ni con los no observados	Pérdida accidental de registros de pacientes por un incendio. Pérdida del seguimiento de un paciente porque ha cambiado de trabajo	Pérdida de potencia estadística. No hay sesgo en los parámetros estimados
MAR	Dados los datos observados, la probabilidad de ausencia de datos no depende de datos no observados	Ausencia de datos relacionada con características conocidas de los pacientes, tiempo, lugar o resultados	Pérdida de potencia estadística y sesgo en el análisis CC para datos con más del 25% de ausencia[20]. El sesgo puede minimizarse con el empleo de imputación múltiple (con ausencia de datos de entre el 10 y el 50%)
NMAR	La probabilidad de ausencia de datos depende de variables no observadas y/o de valores no disponibles de los datos	Ausencia de datos relacionada con el valor del predictor o características no disponibles en el análisis	Pérdida de potencia estadística. El sesgo no puede reducirse en este caso y deben realizarse análisis de sensibilidad en diversos supuestos de NMAR

CC: caso completado; MAR: ausencia de datos aleatoria; MCAR: ausencia de datos completamente aleatoria; NMAR: ausencia de datos no aleatoria.

perdidos no aleatorios (NMAR) (tabla 3). La imputación múltiple se desarrolló para abordar la ausencia de datos en los supuestos de MAR y MCAR mediante la sustitución de los valores no disponibles por una serie de valores plausibles basados en información auxiliar presente en otras variables. A pesar de que, en la mayoría de los casos, el mecanismo por el que se han producido los valores perdidos no es identificable y rara vez es MCAR, la mayor parte de los estadísticos contemporáneos se muestran proclives a imputar dichos valores utilizando algoritmos de imputación múltiple complejos, en especial cuando los valores perdidos llegan a ser el 5% o más.

2. Codificación de variables. Las variables a incluir en los modelos de regresión deben estar codificadas adecuadamente e intentar acoplar categorías adyacentes de una variable ordinal si fuese necesario reducir la dimensionalidad de los datos. En la medida de lo posible, hay que mantener las variables en forma continua, puesto que su categorización (o, peor aún, su dicotomización) implicaría una pérdida importante de la información contenida en la variable; además de la arbitrariedad que representa el punto de corte elegido. Por lo tanto, cuando se dicotomize una variable, hay que presentar argumentos respecto a la elección del punto de corte elegido o si se ha basado en un valor de corte ya aceptado en la literatura médica.
3. Verificación de observaciones con influencia manifiesta. Al evaluar qué tan bien se ajustan los resultados del modelo a los datos, es importante realizar análisis de influencia o de apalancamiento para determinar el efecto de ciertas observaciones con influencia desproporcionada en los parámetros de regresión. Lamentablemente, no hay una guía sólida respecto a cómo tratar estas observaciones con influencia manifiesta. En dichas circunstancias, a veces es necesario examinar cuidadosamente dichos valores en los documentos fuente con el fin de determinar el origen de dichas observaciones.

ESTRATEGIAS DE CREACIÓN DE MODELOS

La selección de variables es un paso crucial en el proceso de creación del modelo (tabla 1). La inclusión de variables adecuadas es un proceso intensamente influido por el equilibrio preespecificado entre complejidad y simplicidad (tabla 3). Los modelos predictivos deben incluir las variables que reflejen el patrón de la asociación en estudio en la población representada en los datos. En este caso, lo que importa es la información que el conjunto del modelo representa. Por otra parte, en modelos donde el fin es la estimación del efecto en estudio, un modelo ajustado que refleje la idiosincrasia de los datos es aceptable en tanto se corrijan los parámetros estimados respecto a la sobresaturación.

Se utiliza el término sobresaturación *(overfitting)* para describir un modelo ajustado con un número excesivo de grados de libertad respecto al número de observaciones (o eventos en los modelos binarios). Esto sucede generalmente cuando el modelo incluye demasiados predictores y/o asociaciones complejas entre los predictores y la respuesta (como interacciones, efectos no lineales complejos, etc.) que pudiesen estar presentes en la muestra pero no en la población. Como consecuencia, es poco probable que las predicciones del modelo sobresaturado se repliquen en una población diferente, algunos de los predictores seleccionados pueden mostrar una asociación falsa con el evento, y los coeficientes de regresión estarán sesgados en dirección opuesta a la hipótesis nula (exceso de optimismo). En otras palabras, si se incluyen demasiados predictores en un modelo, es muy probable que se obtengan resultados aparentemente importantes, independientemente de su existencia o no en la población. Se han propuesto reglas básicas que establecen el número de predictores que incluir en el modelo según el tamaño muestral. En la regresión múltiple lineal, se ha recomendado un mínimo de 10 a 15 observaciones por predictor[21]. Para los modelos de supervivencia, el número de eventos en el criterio de valoración es el factor limitante (10 a 15)[22]. Para la regresión logística, si el número de no eventos es inferior al número de eventos, este pasará a ser el número a utilizar. En los estudios de simulación, entre 10 y 15 eventos por variable fue la proporción óptima[23,24].

Se han propuesto otras medidas adicionales para corregir el fenómeno de sobresaturación: *a)* utilizar la experiencia en el campo para eliminar variables no importantes; *b)* eliminar variables con distribución demasiado estrecha; *c)* eliminar predictores con un número elevado de valores perdidos; *d)* aplicar técnicas de reducción de los datos y penalización de los coeficientes de regresión, y *e)* intentar agrupar varias variables en una, utilizando medidas de similitud, ya sea con el uso de técnicas estadísticas multivariables, con puntuaciones pronósticas ya validadas, o con una puntuación de propensión estimada.

Selección automática de variables

La mayoría de los programas informáticos de estadística ofrecen la opción de seleccionar automáticamente el «mejor modelo» mediante la introducción y/o exclusión secuencial de variables predictoras. En la selección anterógrada, el modelo inicial incluye solamente una constante, y en cada paso posterior se añade al modelo la variable que aporta el máximo (y significativo) ajuste del modelo. En la eliminación retrógrada, el modelo inicial es el modelo que incluye todas las variables, y en cada paso se excluye una de ellas con base en la menor reducción (no significativa) en el

ajuste del modelo. También es posible un enfoque «combinado» que empieza con una selección anterógrada pero, tras la inclusión de la segunda variable, valora en cada paso si una de las variables ya incluidas puede ser eliminada del modelo sin que se produzca una reducción significativa del ajuste de este. El modelo final de cada uno de estos procedimientos escalonados idealmente debería incluir las variables predictoras que mejor expliquen la respuesta.

El empleo de procedimientos escalonados ha sido criticado por diversos motivos[2,24–27]. Es frecuente que los métodos escalonados no logren incluir todas las variables que realmente influyen en la variable de evaluación o que seleccionen variables que no tienen influencia alguna. Al relajar el valor de p = 0,05 utilizado como regla de detención, mejora la selección de variables importantes en series de datos pequeñas[1]. Los procedimientos escalonados se han asociado también a un aumento de la probabilidad de identificar al menos una de las variables significativas por azar (error de tipo I), a causa de la realización de pruebas múltiples sin la introducción simultánea de ajuste por el número de comparaciones. Una selección anterógrada con 10 predictores realiza 10 pruebas de significación en el primer paso, 9 pruebas de significación en el segundo paso, y así sucesivamente, y en cada ocasión incluye una variable en el modelo cuando alcanza el criterio especificado. Además, los resultados de los procedimientos escalonados tienden a ser inestables, en el sentido de que cambios tan sólo ligeros de los datos pueden conducir a resultados diferentes en cuanto a las variables que conformen el modelo final y la secuencia en la que se incorporan a él. Por ello este tipo de procedimiento no es apropiado para jerarquizar la importancia relativa del predictor en modelo.

Para superar los inconvenientes asociados a los procedimientos escalonados, Royston[28] ha desarrollado un procedimiento denominado polinomios fraccionales multivariables (MFP), que incluye dos algoritmos para la selección de un modelo polinómico fraccional, que en ambos casos combina una eliminación retrógrada con la selección de una función polinómica fraccional para cada predictor continuo. Se empieza transformando la variable a partir de los polinomios fraccionales más complejos y luego, para intentar simplificar el modelo, se reducen los grados de libertad hacia 1 (lineal). Durante este proceso de linealización, se establece la transformación óptima de las variables continuas mediante pruebas estadísticas. Se ha afirmado que el algoritmo utilizado por defecto se parece a un procedimiento de prueba cerrada, al mantener el error general de tipo I en el nivel nominal preespecificado (generalmente del 5%). El algoritmo alternativo disponible en MFP, el algoritmo secuencial, se asocia a un error general de tipo I de aproximadamente el doble del procedimiento de prueba cerrado, aunque se cree que este error de tipo I aumentado proporciona una mayor potencia estadística para la detección de relaciones no lineales.

Se han propuesto otras alternativas sofisticadas para mejorar el proceso de selección de variables, como la regresión de subgrupo mejor[29], las técnicas escalonadas con datos imputados múltiples[30,31] y el empleo de algoritmos de selección automáticos que hacen las correcciones apropiadas en los coeficientes estimados, como el método LASSO *(least absolute shrinkage and selection operator)*[32] y la penalización de probabilidad máxima[33].

Recientemente, se ha recomendado el uso de técnicas de remuestreo *(bootstrap)* como medio de evaluar el grado de estabilidad de los modelos obtenidos mediante procedimientos escalonados[34–37]. Estas técnicas de remuestreo simulan la estructura de los datos de que se dispone. La frecuencia de las variables seleccionadas en cada muestra, a la que se denomina fracciones de inclusión de remuestreo *bootstrap* (BIF), podría interpretarse como criterio de la importancia de una variable. Una variable que tenga una correlación débil con otras y sea significativa en el modelo completo deberá seleccionarse en alrededor de la mitad de las pruebas de remuestreo (BIF $\geq$ 50%). Con valores de p inferiores, la BIF aumenta hacia el 100%.

EVALUACIÓN DEL MODELO FINAL

Un elemento central en el proceso de crear un modelo de regresión es su evaluación en cuanto al rendimiento. En este sentido, se han propuesto diversas medidas, que pueden agruparse en dos categorías principales: medidas de calibración y de discriminación (tablas 1 y 3). Independientemente del objetivo para el que se ha creado el modelo, estas dos medidas del rendimiento deben derivarse de los datos que le han dado origen, y preferiblemente deben estimarse utilizando técnicas de remuestreos (o *bootstrap*), lo que se conoce como validez interna. Con el remuestreo, es posible cuantificar el grado de exceso de optimismo en los coeficientes de regresión y, por lo tanto, la cantidad de simplificación (reducción) que es necesaria para corregirlo. Sin embargo, si el objetivo es evaluar la validez externa del modelo (que es un aspecto crucial en los modelos de predicción), estas medidas del rendimiento deben estimarse en una población diferente. Desafortunadamente, esto no siempre es posible, debido a la falta de recursos. Para una revisión detallada de este tema, consúltese «Clinical Prediction Models» de Steyerberg[38].

La calibración es una medida que expresa la concordancia entre los resultados observados y las predicciones del modelo. En otras palabras, es la capacidad del modelo de producir estimaciones no sesgadas de la probabilidad del evento o variable de valoración. Las medidas de calibración más habituales son la calibración general, la pendiente de calibración (ambas derivadas de los gráficos de calibración) y la prueba de Hosmer-Lemeshow (o su equivalente para la regresión de Cox, la prueba de Gronnesby y Borgan[39]).

La discriminación es la capacidad del modelo de asignar el resultado correcto a un par de sujetos seleccionados al azar; en otras palabras, permite al modelo clasificar a los sujetos en un contexto de criterio de valoración con predicción binario. El área bajo la curva ROC (AUC) es la medida de discriminación más frecuentemente utilizada para modelos de error normal y resultado binario. Su equivalente para los datos con observaciones censuradas es el estadístico C[40].

En el contexto de la investigación traslacional (ómica y biomarcadores), la evaluación del valor predictivo añadido que aporta un predictor es tal vez igual de importante que la validación de la exactitud de predicción del modelo en conjunto. Se han propuesto varios enfoques, los más importantes de los cuales son la mejora de reclasificación neta, la mejora de discriminación integrada[41,42] y el análisis de curvas de decisión[43].

En resumen, un modelo de regresión con buenas propiedades de calibración y discriminación, idealmente evaluadas mediante remuestreo, se considera actualmente un requisito importante para su aplicación en predicción pronóstica, siempre y cuando su evaluación con datos independientes no sea posible.

PRESENTACIÓN DE LOS RESULTADOS

Las consideraciones finales en el proceso de creación de un modelo corresponden a la forma en la que se presentarán los parámetros estimados. Con frecuencia, los programas informáticos de estadística expresan la magnitud del efecto de la variable explicativa en unidades relativas, al comparar dos grupos respecto a un resultado binario. Para la regresión logística y la regresión de Cox, la OR y la HR son las unidades tradicionales utilizadas para indicar el grado de asociación entre una variable y el evento (o variable de evaluación). Dado que se trata de cocientes de proporciones, la información aportada acerca de la magnitud del

efecto es de carácter relativo. De igual modo, en los ensayos controlados y aleatorizados, el riesgo relativo, que no es más que un cociente de proporciones, se utiliza a menudo para expresar la asociación entre la intervención y el evento binario. La principal limitación inherente a estas medidas relativas es que no se ven influidas por las diferencias absolutas observadas. Por ejemplo, en una regresión de Cox, la HR no tiene en cuenta la función de riesgo basal y, por lo tanto, debe interpretarse como un efecto constante de la exposición (o intervención) durante el seguimiento. En consecuencia, existe la preocupación de que las medidas relativas aporten una información clínica limitada, mientras que las medidas absolutas, al incorporar la información correspondiente al riesgo basal del evento, serán más relevantes para la toma de decisiones clínicas.

Las medidas absolutas más frecuentes son: la DRA y el NNT. El NNT es simplemente el recíproco de la DRA. Como consecuencia de la aleatorización, en los ensayos controlados y aleatorizados es relativamente sencillo estimar la DRA, simplemente como la diferencia en las proporciones del evento entre los individuos tratados y los no tratados al final del ensayo. A pesar de que el resultado sea del tipo de tiempo hasta que ocurre el criterio de valoración (estudios de supervivencia), la diferencias de proporciones pueden estimarse a tiempos de seguimiento específicos mediante las curvas de supervivencia de KM[44]. Sin embargo, en los estudios observacionales, a menudo los sujetos que corresponden a los dos grupos de la variable de exposición difieren sistemáticamente en covariables basales con importancia pronóstica, lo cual conduce a su vez a la aplicación de métodos estadísticos que permiten el cálculo de la DRA ajustada (y del NNT)[45].

En resumen, en los estudios de cohorte observacionales, la DRA y el NNT pueden obtenerse a partir de modelos de regresión logística y regresión de Cox, y en la medida de lo posible estas medidas deben complementar la presentación de las estimaciones de regresión tradicionales.

CONFLICTO DE INTERESES

Ninguno.

BIBLIOGRAFÍA

1. Steyerberg EW, Eijkemans MJ, Harrell Jr FE, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making. 2001;21:45–56.
2. Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer-Verlag; 2001.
3. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837–47.
4. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. BMJ. 2009;338:b604.
5. Steyerberg EW. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009.
6. Annesi I, Moreau T, Lellouch J. Efficiency of the logistic regression and Cox proportional hazards models in longitudinal studies. Stat Med. 1989;8:1515–21.
7. Martinussen T, Scheike TH. Dynamic regression models for survival data. New York: Springer-Verlag; 2006.
8. Lambert PC, Royston P. Further development of flexible parametric models for survival analysis. Stata J. 2009;9:265–90.
9. Pintilie M. Competing risks: a practical perspective. New York: John Wiley & Sons; 2007.
10. Fine JP, Gray RJ. A proportional hazard model for the subdistribution of a competing risk. J Am Stat Assoc. 1999;94:496–509.
11. Ibrahim JG, Chu H, Chen LM. Basic concepts and methods for joint models of longitudinal and survival data. J Clin Oncol. 2010;28:2796–801.
12. Rizopoulos D. Joint modelling of longitudinal and time-to-event data: challenges and future directions. En: 45th Scientific Meeting of the Italian Statistical Society. Padova: Università di Padova; 2010.
13. Touloumi G, Babiker AG, Kenward MG, Pocock SJ, Darbyshire JH. A comparison of two methods for the estimation of precision with incomplete longitudinal data, jointly modelled with a time-to-event outcome. Stat Med. 2003;22:3161–75.
14. Touloumi G, Pocock SJ, Babiker AG, Darbyshire JH. Impact of missing data due to selective dropouts in cohort studies and clinical trials. Epidemiology. 2002;13:347–55.
15. Rizopoulos D. JM: An R package for the joint modelling of longitudinal and time-to-event data. J Stat Soft. 2010;35:1–33.
16. Pantazis N, Touloumi G. Analyzing longitudinal data in the presence of informative drop-out: The jmre1 command. Stata J. 2010;10:226–51.
17. Meira-Machado L, De Una-Álvarez J, Cadarso-Suárez C, Andersen PK. Multi-state models for the analysis of time-to-event data. Stat Methods Med Res. 2009;18:195–222.
18. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26:2389–430.
19. Little RJA, Rubin DB. Statistical Analysis with Missing Data, 2nd ed., Hoboken: J.W. Wiley & Sons; 2002.
20. Marshall A, Altman DG, Holder RL. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. BMC Med Res Methodol. 2010;10:112.
21. Green SB. How many subjects does it take to do a regression analysis? Multivar Behav Res. 1991;26:499–510.
22. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol. 1995;48:1503–10.
23. 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.
24. Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol. 1999;52:935–42.
25. Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Stat Med. 1989;8:771–83.
26. Steyerberg EW, Eijkemans MJ, Harrell Jr FE, Habbema JD. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med. 2000;19:1059–79.
27. Austin PC, Tu JV. Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. J Clin Epidemiol. 2004;57:1138–46.
28. Royston P, Sauerbrei W. MFP: multivariable model-building with fractional polynomials. En: Royston P, editor. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables.. Chichester: John Wiley & Sons; 2008. p. 79–96.
29. Roecker EB. Prediction error and its estimation for subset-selected models. Technometrics. 1991;33:459–68.
30. Vergouwe Y, Royston P, Moons KG, Altman DG. Development and validation of a prediction model with missing predictor data: a practical approach. J Clin Epidemiol. 2010;63:205–14.
31. Wood AM, White IR, Royston P. How should variable selection be performed with multiply imputed data? Stat Med. 2008;27:3227–46.
32. Tibshirani R. Regression shrinkage and selection via the LASSO. J Roy Stat Soc B Stat Meth. 2003;58:267–88.
33. Steyerberg EW. Modern estimation methods. En: Steyerberg EW, editor. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009. p. 231–40.
34. Beyene J, Atenafu EG, Hamid JS, To T, Sung L. Determining relative importance of variables in developing and validating predictive models. BMC Med Res Methodol. 2009;9:64.
35. Royston P, Sauerbrei W. Model stability. En: Royston P, editor. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. Chichester: John Wiley & Sons; 2008. p. 183–99.
36. Royston P, Sauerbrei W. Bootstrap assessment of the stability of multivariable models. Stata J. 2009;9:547–70.
37. Vergouw D, Heymans MW, Peat GM, Kuijpers T, Croft PR, De Vet HC, et al. The search for stable prognostic models in multiple imputed data sets. BMC Med Res Methodol. 2010;10:81.
38. Steyerberg EW. Evaluation of performance. En: Steyerberg EW, editor. Clinical prediction models: a practical approach to development, validation, and updating. New York: Springer; 2009. p. 255–79.
39. May S, Hosmer DW. Advances in survival analysis. En: Balakrishnana N, Rao CR, editors. Hosmer and Lemeshow type goodness-of-fit statistics for the Cox proportional hazards model. Amsterdam: Elsevier; 2004. p. 383–94.
40. Harrell Jr FE, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
41. Pencina MJ, D'Agostino Sr RB, D'Agostino Jr RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157–72.
42. Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11–21.
43. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26:565–74.
44. Altman DG, Andersen PK. Calculating the number needed to treat for trials where the outcome is time to an event. BMJ. 1999;319:1492–5.
45. Laubender RP, Bender R. Estimating adjusted risk difference (RD) and number needed to treat (NNT) measures in the Cox regression model. Stat Med. 2010;29:851–9.