Rev Chil Anest 2015; 44: 106-111

Análisis de variables múltiples

RICARDO FUENTES H.[1]

[1] Profesor Asistente, División de Anestesiología, Pontificia Universidad Católica de Chile.

Key words: Evidence medicinne based, anesthesia.

- El análisis de variables múltiples se usa cada vez más en medicina. Es una técnica que analiza en forma simultánea varias variables que son sometidas a investigación.
- La técnica a usar va a depender de si las variables independientes y dependientes son nominales y/o numéricas.
- Las técnicas más usadas en medicina son la regresión logística y el modelo de riesgo proporcional de Cox.
- La regresión logística se usa frecuentemente en estudios de riesgo cuando las variables independientes incluyen medidas numéricas y nominales y el resultado es binario.
- El modelo de riesgo proporcional de Cox se usa especialmente para el análisis de sobrevida.
- En estos estudios se debe verificar, entre otras: que se haya hecho un análisis de residuales, la apropiada definición y no confusión de variables predictivas y resultados, las presunciones y validación del modelo elegido y el cálculo y mención de R^2.

El análisis multivariable es una herramienta estadística para determinar la contribución única de varios factores a un evento o resultado simple. Las técnicas estadísticas dirigidas a las variables múltiples cada vez se usan más en medicina, dado su gran ductilidad y por la creciente disponibilidad de programas computacionales. Sin embargo, junto a ello, ha aumentado su mal uso e interpretación.

En general, el análisis de variables múltiples se refiere a todos los métodos estadísticos que analizan en forma simultánea varias medidas (más de dos variables) de cada individuo u objeto que es sometido a investigación. En este tipo de análisis todas las variables deben ser aleatorias y sus diferentes efectos no debieran ser interpretados separadamente con algún sentido. El propósito de este tipo de análisis es medir, explicar y predecir el grado de relación de los valores teóricos. Así el carácter multivariable reside en los múltiples valores teóricos y no sólo en el número de variables u observaciones.

El análisis de los datos implica la separación, identificación y medición de la variación en un conjunto de variables, tanto entre ellas como entre una variable dependiente y una o más variables independientes. La variación debe ser mesurable.

El investigador debe identificar la escala de medida de cada variable empleada (nominales, ordinales y numéricas), tanto para las variables dependientes como las independientes, para poder decidir qué técnica de variable múltiple es la más conveniente para los datos.

Existen diferentes métodos a utilizar cuando hay investigaciones que comprenden dos o más variables independientes. Si el término variable independiente se usa para designar aquellas va-

Correspondencia a:
Dr. Ricardo Fuentes H.
rfuente@med.puc.cl

riables predictoras de pertenencia a un grupo, o la variable *X*, y el término dependiente se utiliza para designar las variables predichas cuyas medidas son comparadas, o la variable *Y*, podemos señalar los métodos como se listan en la Tabla 1.

Regresión múltiple

La regresión múltiple es el método de análisis apropiado cuando el problema del investigador incluye una única variable numérica dependiente que estaría relacionada con una o mas variables numéricas independientes. Su objetivo es predecir los cambios en la variable dependiente en respuesta a cambios en las variables independientes. Por ejemplo: interesa saber cuál es la influencia sobre el IMC (índice de masa corporal) de ingesta diaria de gramos de hidratos de carbono, lípidos, proteínas y alcohol; aquí la variable dependiente es el IMC y la ingesta de cada componente las independientes. Es útil de realizar cuando el investigador esté interesado en predecir la cantidad o magnitud de la variable dependiente. El modelo de regresión lineal simple es:

$$Y = a + bX$$

donde Y=variable dependiente, X=variable independiente, "a" es el punto de intersección y "b" es el coeficiente de regresión, la pendiente de la recta que se ajusta a los datos.

La extensión de una regresión simple a una múltiple de dos o más variables independientes es directa. Por ejemplo, si son cuatro variables independientes bajo estudio, el modelo de regresión múltiple es:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4$$

donde X_1 es la primera variable independiente y b_1 el primer coeficiente de regresión asociado con ella y así sucesivamente.

En la regresión múltiple, un coeficiente de regresión dado indica qué tanto cambia el valor predicho de Y cada vez que X aumenta su valor en una unidad, siempre que se mantengan constantes los valores de todas las demás variables de la ecuación de regresión. Esta característica permite que sea un método para controlar las diferencias basales y las variables de confusión.

Cuando los estudios incluyen diversas variables, algunas observaciones de ciertos individuos pueden faltar o perderse. En estos casos es importante determinar el porcentaje de la información que se perdió y, si la pérdida fue imprevista o se perdió por algún factor causante. El potencial de las observaciones perdidas aumenta en estudios con variables múltiples. Las soluciones pueden ser exclusión de individuos en quienes se han perdido observaciones, eliminación de variables que han perdido valores del estudio o sustitución de algunos valores (como por ejemplo, promedio del valor esperado). Este tema será analizado en

Tabla 1. Resumen de métodos a utilizar en investigaciones que comprenden dos o más variables independientes

Variables independientes (X)	Variables dependientes (Y)	Método (s)
Nominal	Nominal	Logarítmica lineal
Nominal y numérica	Nominal (binaria)	Regresión logística
Nominal y numérica	Nominal (dos o más categorías)	Regresión logística Análisis discriminativo Análisis grupal Árbol de clasificación y regresión
Nominal	Numérica	Análisis de varianza (ANOVA) Análisis de varianza multivariada (MANOVA)
Numérica	Numérica	Regresión múltiple
Nominal y numérica	Numérica	Modelo de riesgo proporcional de Cox
Factores de confusión	Numérica	Análisis de covarianza (ANCOVA) Análisis de varianza multivariada (MANOVA) Ecuación de estimación generalizada

otro artículo.

Algunas de las variaciones observadas en cualquier variable se presentan sólo por casualidad. La ecuación de regresión no puede distinguir entre variación real y casual. La ecuación puede validarse con una segunda muestra, lo que es llamado validación cruzada. La ecuación de regresión se utiliza para predecir el resultado en la segunda muestra, y los resultados esperados se comparan con los reales.

Un problema común en el análisis de regresión es la selección de qué variables predictivas deben ser incluidas en el modelo. Un método de selección de variables es por protocolo, es decir, el investigador decide a *priori* qué y cuántas variables predictoras incluirá en el modelo basado habitualmente en investigaciones previas que las identificaron como relevantes. Otra manera es usar a *posteriori*, una vez que se recolectó la información, métodos para seleccionar qué variables se incluirán; al respecto se usan dos, los llamados "*forward stepwise*" o "*backward stepwise*". En la selección "forward", las variables son ingresadas al modelo de una por vez y aquellas que no aumentan significativa o considerablemente el valor predictivo del modelo son excluidas. En la selección "backward", todas las variables predictivas son ingresadas al modelo; luego el predictor más débil es removido y se recalcula la regresión. Si el modelo se debilita significativamente, la variable es reingresada; en caso contrario es eliminada.

Las presunciones varían según el tipo de regresión y los autores deben discutir si esas presunciones fueron verificadas y cómo. Una causa frecuente de inestabilidad de los modelos, vale decir cambios importantes en los resultados ante variaciones menores en las variables predictoras, es un tamaño de muestra insuficiente. Se sugiere como mínimo 10 casos con el resultado que se analiza por cada variable predictora considerada en el modelo. Por ejemplo, si se quiere estudiar los factores predictores de muerte postoperatoria y se tiene una muestra de 1.000 pacientes seguidos por 30 días con 33 pacientes fallecidos el modelo no debe incluir más de 3 variables predictoras.

Antes de aceptar los resultados, se debe intentar validar el modelo para así poder generalizar los resultados al conjunto de la población. Algunos métodos para ello son: 1) división de la muestra en dos, usando una submuestra para estimar el modelo y la segunda submuestra para estimar la precisión predictiva; 2) análisis de "*bootstrapping*" (técnica de remuestreo de datos), que permite resolver problemas relacionados con la estimación de intervalos de confianza o la prueba de significación estadística) y 3) conseguir una muestra distinta. Esto puede implicar hacer otro estudio en la misma institución, otro estudio en otra institución distinta u otro en múltiples instituciones. La fortaleza predictiva del modelo aumenta cuando se sostiene con muestras diversas.

Análisis de covarianza (ANCOVA)

El ANCOVA es la técnica estadística para controlar las influencias de una variable de confusión, la cual se presenta cuando los individuos no pueden asignarse a diferentes grupos por azar. Se puede usar para ajustar el defecto por variables de confusión en más de dos grupos. También se puede ajustar para más de una variable de confusión en el mismo estudio; las variables de confusión pueden ser nominales o numéricas.

El ANCOVA supone que las relaciones entre la covariante (variable X) y la variable dependiente (Y) es la misma en los grupos y, por ende, las pendientes de regresión son las mismas. El ANCOVA investiga si existe o no una diferencia entre las intersecciones, asumiendo que las pendientes son iguales, enfoca la atención en el punto de intersección.

Regresión logística

La regresión logística junto al modelo de riesgo proporcional de Cox (ver más adelante), son las técnicas de análisis de variables múltiples más utilizadas actualmente en medicina. Se usa cuando las variables independientes incluyen medidas numéricas y nominales y el resultado es binario (dicotómico: sí o no, por ejemplo: fallecimiento).

La regresión logística modela la probabilidad de un resultado y cómo la probabilidad cambia con un cambio en las variables predictivas. El supuesto básico es que cada aumento en una unidad en el predictor multiplica la probabilidad del resultado por un cierto factor (la *odds ratio* del predictor) y que el efecto de varias variables es el

producto multiplicativo de sus efectos individuales. La función regresión logística produce una probabilidad de resultado limitada por 0 y 1.

Una ventaja de la regresión logística es que no es necesario hacer suposiciones sobre la distribución de las variables independientes. El coeficiente de regresión puede interpretarse en función del riesgo relativo en estudios de cohortes o como *odds ratio* en estudios de caso-control.

Los resultados globales de la regresión logística pueden probarse con la prueba de calidad de ajuste de Hosmeer y Lomeshow.

Análisis logarítmico lineal

Es análogo al modelo de regresión, donde todas las variables dependientes e independientes se miden en una escala nominal. Esta técnica implica utilizar el logaritmo de las frecuencias observadas en el cuadro de contingencia.

También puede utilizarse cuando no existe distinción entre las variables dependientes e independientes. Así, el coeficiente de regresión no se interpreta en el análisis logarítmico lineal.

Modelo de riesgo proporcional de Cox

Esta técnica es apropiada cuando las observaciones esperadas que se incluyen dependen del tiempo. Se usa especialmente para el análisis de sobrevida.

Los coeficientes de regresión de Cox pueden emplearse para determinar la *odds ratio* o el riesgo relativo asociado con cada variable independiente y el resultado variable, ajustado por el efecto de otras variables incluidas en la ecuación. El modelo de Cox da el riesgo relativo ajustado. Más detalles serán analizados en otro artículo.

Análisis de varianza multivariada (MANOVA)

El MANOVA es una técnica que puede ser usada para ver las relaciones entre diversas categorías de variables independientes y dos o más variables métricas dependientes. Por ejemplo, analizar el efecto sobre la densidad ósea y calcio sérico de cantidad de calcio en la dieta, consumo de alcohol, tipo de dieta (vegetariana o no). Representa una extensión del ANOVA. Es útil cuando el investigador diseña una manipulación de varias variables de tratamiento no numéricas para comprobar hipótesis concernientes a la varianza de respuestas de grupos sobre dos o más variables numéricas dependientes, aunque todos los sujetos tengan las mismas cifras en las otras variables.

¿Cómo verificar el uso de un análisis de variables múltiples en una investigación?

Antes de estimar el modelo multivariante, se deben evaluar los supuestos subyacentes. Todos los modelos multivariantes tienen supuestos subyacentes, que afectan su capacidad para representar relaciones multivariantes. Así se deben tener los supuestos de normalidad multivariante, linealidad, independencia de los términos de error e igualdad de las varianzas en una relación de dependencia.

Cuando un objetivo del estudio es determinar el riesgo relativo de un evento adverso en dos o más grupos de pacientes, es muy importante que los pacientes sean estratificados en grupos que sean relativamente homogéneos con respecto al procedimiento, de lo contrario podría ser difícil generalizar los resultados. Para evitar este problema se debe incluir en el análisis sólo los grupos que están muy bien definidos.

En un estudio clínico se pueden definir dos tipos de variables: las variables resultados y las variables predictivas. Las variables deben ser comúnmente definidas y fácilmente comprendidas, sin dejar la más mínima posibilidad de error de interpretación por parte de los observadores o colectores de datos. A mayor exactitud en definir las variables, mayor será el poder predictivo del modelo. Por ejemplo en un estudio quirúrgico, las variables predictivas representan las condiciones del paciente antes del procedimiento quirúrgico, tales como edad, antecedente de hipertensión o hematocrito previo. En cambio, ejemplo de las variables resultados son duración de la cirugía o episodios de hipotensión.

Para evitar el problema de muchas variables, se sugiere analizar cada variable predictiva separadamente por su influencia sobre los resultados. La probabilidad de encontrar factores falsamente significativos está aumentada cuando varias variables se evalúan en forma conjunta. Además si se evalúan varias variables significativas en forma conjunta, puede resultar que sólo una o dos

resulten significativas. Otra sugerencia es colapsar o agrupar varias variables en un índice. Al utilizar tales índices se puede reducir el número de variables predictivas y evitar el problema de inestabilidad del modelo.

El análisis de los residuales es la mejor forma de evaluar si un modelo se ajusta a los datos. Los residuos son la diferencia entre los valores observados y los estimados; pueden ser considerados como el error en la estimación. Residuos grandes sugieren que el modelo no se ajusta a los datos. Lamentablemente las publicaciones rara vez informan sobre los residuos y ni siquiera mencionan que se hizo un análisis de ellos.

La calibración y la discriminación son dos términos que describen los componentes de la exactitud de la predicción. La calibración se refiere al grado del sesgo. La discriminación mide la capacidad de un predictor de separar pacientes con diferentes respuestas. Si un modelo predictivo tiene pobre discriminación, ningún ajuste o calibración puede corregir el modelo. Si la discriminación es buena, el predictor puede ser calibrado sin sacrificar la discriminación. Todo trabajo debiera mencionar y comentar estos aspectos.

Para una variable de respuesta continua, la discriminación está relacionada al error cuadrático esperado y a la correlación entre las respuestas observada y la predicha. En el caso de una regresión lineal múltiple, la discriminación puede ser medida por el coeficiente de correlación múltiple cuadrado? (R^2). Cuando $R^2 = 1$, el modelo es perfectamente capaz de separar todas las respuestas del paciente basado en la variable predictiva y el error cuadrático de predicción será igual a 0. Es decir, que en un modelo con un R^2 cercano a 1, las variables dependientes predicen con precisión el resultado y lo inverso si R^2 se acerca a 0.

Cuando la variable de resultado es dicotómica y la predicción está basada en la probabilidad que un evento ocurrirá, la calibración y la discriminación proporcionan más información en la exactitud de la medición que el error cuadrático esperado.

Fiabilidad de un modelo

Una importante amenaza para la fiabilidad es un tamaño de muestra insuficiente. Ya hemos mencionado que debería haber al menos 10 resultados por cada variable independiente. El tamaño muestral requerido está basado sobre el estado menos frecuente de los resultados dicotómicos. Un intervalo de confianza ancho es el resultado de un tamaño muestral insuficiente.

Aún si un estudio tiene un número suficientemente grande de eventos por variable independiente, las estimaciones de la asociación entre un factor de riesgo y un resultado todavía pueden ser inexactos si el factor de riesgo es raro.

Consejos para el lector crítico

A continuación se presenta una lista de verificación en que se señalan los puntos que debe chequear un lector crítico al estar en presencia de un trabajo que ha utilizado algún tipo de análisis de variables múltiples:

- El artículo debe hacer mención y ojalá presentar el análisis de residuales. Un residual elevado sugiere que el modelo no se ajusta bien a los datos.
- Número de variables: recordar que deben existir al menos 10 muestras o resultados por cada variable considerada en el estudio.
- Revisar una clara definición de las variables predictivas y las variables resultados. Además no se debe incluir como variables predictivas aquellas que son variables resultados (por ejemplo presencia de hipotensión o tiempo de duración de la cirugía).
- Chequear las presunciones del modelo elegido: que cumplan con las características de distribución de las variables, linealidad de las variables continuas (para cada variable predictiva "*X"* está linealmente relacionado a *"Y"*) y aditividad de los efectos de los predictores (ausencia de interacción).
- Existe una correcta calibración (o grado del sesgo) del modelo (observado versus predicho). Por ejemplo si la mortalidad predicha es en promedio 0,15 para un grupo particular de pacientes y el resultado de fallecidos es 0,15; la predicción está bien calibrada.
- Debe existir un análisis de discriminación. La discriminación mide la capacidad de un predictor de separar a los pacientes que tienen diferentes respuestas.
- Se debe mencionar el R^2.

Referencias

1. Dawson-Saunders B, Trapp R. Bioestadística médica. D.F., México: Manual Moderno; 2005.
2. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996 Feb;15(4):361-387.
3. Katz MH. Multivariable analysis: a primer for readers of medical research. Ann Intern Med 2003 Apr;138(8):644-650.
4. Reboldi G, Angeli F, Verdecchia P. Multivariable analysis in cerebrovascular research: practical notes for the clinician. Cerebrovasc Dis 2013;35(2):187-193.
5. Smith LR. Observational studies and predictive models. Anesth Analg 1990 Mar;70(3):235-239.