

13. Multicausalidad – enfoques de análisis

Conceptos y métodos para analizar datos epidemiológicos que involucran más de dos variables, control del fenómeno de confusión a través del análisis estratificado y del modelado matemático.

Análisis multivariado

En los dos capítulos precedentes profundizamos en las distintas circunstancias y situaciones que surgen en un contexto multivariado. Introdujimos los modelos aditivo y multiplicativo para los efectos conjuntos de variables de exposición múltiples y empleamos el análisis estratificado para examinar los efectos de una variable controlando para los valores de las otras. En este capítulo consideraremos enfoques analíticos para examinar las relaciones entre un resultado o daño y múltiples variables explicativas. Estas últimas pueden ser un factor de estudio y potenciales variables de confusión, un factor de estudio y potenciales modificadoras, o varias exposiciones que son todas de interés.

Fenómeno de confusión:

Para plantearlo nuevamente brevemente, el fenómeno de confusión es una situación en que un factor o combinación de factores distintos del factor de estudio es responsable de por lo menos una parte de la asociación observada entre el factor de estudio y el resultado. Si no controlamos para el fenómeno de confusión, podríamos atribuir equivocadamente un efecto al factor de estudio cuando la asociación en realidad refleja el efecto de otra variable. En una situación de confusión, los datos crudos nos dan la imagen equivocada de la relación entre la exposición y la enfermedad. Otros factores pueden estar exagerando la fuerza de la relación u ocultando una parte o la totalidad de ella. Para ver la imagen correcta, necesitamos tomar en cuenta los efectos de otros factores.

En una analogía con la ley, la exposición es el sospechoso de un asalto de banco y los otros factores son los delincuentes conocidos con los cuales él se asocia. Necesitamos establecer la culpabilidad del sospechoso. El sospechoso puede ser completamente inocente, puede haber tenido algún rol en el crimen, o puede haber tenido un rol mucho mayor que el que, en principio, aparentemente jugó. Para poder determinar la culpabilidad del sospechoso, necesitamos examinar la imagen total de las acciones de todos los individuos. En esta analogía, el fenómeno de confusión ocurriría si acusáramos al sospecho con un crimen que no cometió o con un rol en el crimen mayor o menor que el que correspondería a sus acciones. Por ejemplo, sería un fenómeno de confusión acusar al sospechoso con un asalto al banco si simplemente estaba pasando por allí y uno de los ladrones lo llamó para que entrara. También sería fenómeno de confusión si acusáremos al sospechoso como cómplice cuando en realidad era el principal organizador del robo.

El método más común de decidir si existe o no el fenómeno de confusión es el de comparar los resultados crudos (no controlados) con los resultados controlados. Si estos dos conjuntos de resultados son significativamente diferentes, si dan un "mensaje" diferente o sugieren una conclusión diferente

sobre la asociación en estudio, el fenómeno de confusión está presente; los resultados crudos están sufriendo "el fenómeno de confusión". La conclusión sobre la presencia del fenómeno de confusión, sin embargo, es secundaria a nuestro propósito principal, que es obtener una estimación válida de la existencia y la fuerza de la asociación entre la exposición de interés y la enfermedad resultante. Cuando determinamos que existe el fenómeno de confusión, presentamos ya sea los hallazgos específicos por estrato o calculamos una medida ajustada de asociación (p.ej., una razón de tasas estandarizada) que controla para los efectos de las variables de confusión.

Modificación del efecto

La modificación del efecto es una situación en que ni la medida cruda ni la ajustada provee una imagen adecuada de la relación en estudio. La imagen no está **equivocada**, pero igual puede engañarnos. Podemos no cumplir con nuestra responsabilidad de presentar la imagen completa. Modificación del efecto significa que hay importantes diferencias entre los grupos (o a distintos niveles de alguna variable modificadora) en la relación entre la exposición y la enfermedad en nuestra escala de medición. Cuando está presente la modificación del efecto, la relación entre la exposición y la enfermedad no es susceptible a formularse tan sencillamente como que "D y E están asociados con un riesgo relativo de aproximadamente 2". Más bien, la respuesta a la pregunta "¿cuál es el riesgo relativo para D dado E?", tiene que ser "depende". Por ejemplo, cualquier discusión de riesgos de enfermedad cardíaca en mujeres que toman anticonceptivos orales sería gravemente incompleta si no explicara que la situación es muy diferente en mujeres que fuman que en las que no fuman, sobretodo después de los 35 años de edad.

Cuando la modificación del efecto está presente, la medida de resumen es un promedio de componentes dispares, de manera que el resumen es demasiado poco informativo por sí mismo. Si Carlos mide 90 cm de altura, Shizue 120 cm de altura, y Rhonda 150 cm de altura puede ser útil saber que su altura promedio es 120 cm, pero probablemente no sea una buena idea comprar tres uniformes para el colegio de tamaño mediano (120 cm).

Enfoques analíticos

Hay dos enfoques analíticos para analizar los datos que involucran más de dos variables: análisis estratificado y modelado. Ya nos hemos enfrentado a los dos. En el análisis estratificado dividimos las observaciones en un grupo para cada nivel o combinación de niveles de las variables de control. Analizamos la asociación entre el factor de estudio y el resultado separadamente dentro de cada grupo. De esta manera tal vez podamos observar la asociación que involucra el factor de estudio sin interferencia de las variables de estratificación.

La comparación de la medida cruda de asociación con las medidas específicas por estrato o su promedio ponderado demostrará si la medida de asociación cruda sufre del fenómeno de confusión o no. El examen de los datos dentro del estrato individual revelará si la medida de asociación varía de manera tan importante que una medida de resumen por si misma puede engañar. Para profundizar en el análisis podemos estratificar por cada una de las co-variables y por varias combinaciones de ellas. El análisis estratificado nos da una imagen más completa que podemos examinar en detalle.

En cierto punto, sin embargo, el detalle es más un obstáculo que una ventaja. El modelado es una estrategia para sumergir el detalle y enfocar en las relaciones. La observación de los datos a través del marco del modelo nos beneficia con potencia analítica y conveniencia. Más que confundirnos nosotros mismos y a nuestra público presentando una plétora de tablas, empleamos la elegante simplicidad del modelo y sus parámetros, a través de los cuales podemos estimar las medidas de asociación que buscamos. Si hemos elegido bien nuestro modelo y evaluado su idoneidad, podemos realizar un análisis óptimo de los datos. Pero igual que el piloto que puede volar muy lejos sólo con instrumentos pero necesita ver la pista cuando aterriza, un análisis por modelado debe estar complementado con algunos análisis estratificados. Por otro lado, en un análisis estratificado, el cálculo de medidas de resumen para muchos estratos involucra un marco para el modelo, por lo menos implícito.

No importando que enfoques utilicemos, no podemos evadir el hecho de que la forma en la que procedemos y como interpretamos los resultados que observamos depende de nuestro modelo conceptual sobre las relaciones entre el daño, la exposición, y las variables de estratificación. Si los factores en estudio son totalmente desconocidos, talvez tengamos que proceder de una manera totalmente empírica. Pero si tenemos algún conocimiento, servirá como guía. Por ejemplo, supongamos que vemos una asociación que involucra nuestro factor de estudio y el daño, pero cuando controlamos para otro factor, la asociación desaparece. El hecho de concluir que hay un "fenómeno de confusión" y desechar la asociación cruda por ser un artefacto, o no, depende de que consideremos la variable de estratificación como una "verdadera" causa del daño más que un factor de estudio. Si la variable de estratificación es un factor intermedio en la cadena causal entre el factor de estudio y el daño, la situación no es una de fenómeno de confusión aunque pueda ser numéricamente idéntica.

Análisis Estratificado — interpretación

El análisis estratificado es conceptualmente simple. Involucra la desagregación de un conjunto de datos en subgrupos definidos por uno o más factores que son los que deseamos controlar. Por ejemplo, al estudiar el efecto del uso de la reserpina sobre el riesgo de cáncer de mama, podríamos estratificar por obesidad. El análisis dentro de cada estrato puede ser considerado entonces libre del efecto del fenómeno de confusión por ese factor de riesgo, en la medida que los estratos sean suficientemente estrechos. (Si los estratos son amplios, p.ej., "índice de masa corporal de 2.2 a 3.2" o "presión arterial mayor de 95 mmHg", podemos tener un "fenómeno de confusión residual" debido a la heterogeneidad de la variable de estratificación dentro de uno o más estratos.)

Ya hemos encontrado análisis estratificados, notoriamente en los capítulos sobre el fenómeno de confusión y la modificación del efecto. En este capítulo obtendremos una comprensión más profunda del análisis estratificado y como se relaciona con los otros conceptos que hemos aprendido. También veremos como y cuando obtener una medida global de resumen que tome en cuenta la estratificación.

Ejemplo

Supongamos que cuatro estudios caso-control han investigado una posible asociación entre la reserpina y el cáncer de mama (una interrogante que surgió en la década de los 70) y cada uno controló para obesidad dividiendo los datos en dos estratos. La tabla que sigue muestra los odds ratio crudos y

específicos por estrato de estos cuatro estudios (hipotéticos). ¿Cómo describiríamos los resultados de cada estudio?

**Asociación entre reserpina y cáncer de mama
controlando para peso corporal (odds ratios)
Datos hipotéticos**

Estudio	Obesidad	No obesos	Resumen (ajustado)	Total (crudo)
A	2.0	2.2	2.1	4.0
B	4.0	2.2	3.1	3.0
C	2.0	2.2	2.1	2.0
D	4.0	2.2	3.1	1.5

En el estudio A, vemos que el OR de cada categoría de peso corporal es de aproximadamente 2.0, mientras que el OR crudo es 4.0. El estudio A, por lo tanto, ilustra una situación de fenómeno de confusión: la medida cruda de asociación cae por fuera del rango de las medidas específicas por estrato. El OR crudo es significativamente diferente del OR ajustado y ningún otro método de ajuste cambiaría eso, dado que cualquier promedio ponderado de los OR específicos por estrato tendría que caer entre 2.0 y 2.2.

En los estudios B y C, por otro lado, el OR crudo podría ser igual (o casi igual) a un promedio ponderado de las medidas específicas por estrato (como lo son los OR ajustados mostrados), porque cae (casi) dentro del rango de esas medidas. Por lo tanto, el fenómeno de confusión no es una característica de los datos en ninguno de estos dos estudios. En el estudio B, si el número de participantes en cada estrato es suficientemente grande para que consideremos la diferencia entre los OR específicos por estrato significativa (no debida simplemente al "ruido"), la diferencia indica entonces una modificación del efecto del OR. Era importante para el estudio informar los OR específicos por estrato y no confiar completamente sobre las medidas crudas o ajustadas.

Si los estratos son suficientemente grandes y los OR se consideran razonablemente libres de sesgo, podríamos preguntarnos si de alguna manera la obesidad potencia el efecto de la reserpina (por lo menos en la escala del OR.) Si se considera que la relación es causal y estos OR son la mejor estimación de la fuerza de la relación, entonces el mayor OR para los pacientes obesos sugiere que ellos en particular deben evitar tomar reserpina si no pueden perder peso (el criterio habitual para "interacción de salud pública" e "interacción del manejo del riesgo individual" es el alejamiento del modelo aditivo de efectos conjuntos esperados. Sin embargo, si la asociación observada es "supra-multiplicativa" [más fuerte que la esperada por el modelo multiplicativo], también sería "supra-aditiva" [más fuerte que la esperada por un modelo aditivo].) En cambio, en el estudio C, la pequeña diferencia entre los dos estratos, aún si no es atribuible a la variación al azar, es insuficiente para merecer atención. Cualquier promedio ponderado de las dos medidas específicas por estrato sería una medida de resumen satisfactoria.

El estudio D ilustra tanto el fenómeno de confusión como la modificación del efecto, dado que el OR crudo cae fuera del rango de los OR específicos por estrato y por lo tanto no puede igualar el promedio ponderado de los dos. Al mismo tiempo, los OR específicos por estrato parecen ser notoriamente diferentes (suponiendo tamaños adecuados de los estratos.) No sería suficiente presentar solamente una medida de resumen (en la escala de OR.)

Resumiendo las relaciones

A menudo estamos interesados en obtener una evaluación global del rol de un factor de estudio, controlando para otros factores de riesgo. La utilidad de una medida global de asociación obviamente será diferente en estos cuatro estudios. En los estudios A y C, una única medida global podría resumir adecuadamente los OR de los dos estratos de manera que no sería esencial presentarlos también. En los estudios B y D sin embargo, claramente necesitamos presentar los OR específicos por estrato, aunque para algunos propósitos una medida de resumen puede también ser útil.

La estimación global más conveniente, si no sufre del fenómeno de confusión, es la medida basada en los datos agregados, la estimación cruda. El análisis estratificado en el estudio C indica que no hay fenómeno de confusión por obesidad. Si esa es la única variable para la cual necesitamos controlar, podemos usar el OR crudo para resumir la relación.

Sin embargo, en el estudio A y en el estudio D, el fenómeno de confusión está presente. Sería claramente engañoso confiar en el OR crudo como la medida de resumen de los resultados estratificados. Por lo tanto, necesitamos una medida de resumen que "ajuste para" la obesidad. La medida de resumen que derivamos es un promedio ponderado de las medidas específicas por estrato. Las medidas de resumen que encontramos en el capítulo sobre estandarización (la REM y la razón de tasas estandarizada) son ejemplos de dichas medidas de resumen.

Relación entre el análisis estratificado y los modelos para los efectos conjuntos

Los modelos aditivo y multiplicativo presentados anteriormente expresan la incidencia conjunta o efecto de dos (o más) factores en términos de la incidencia separada o efecto de cada una. El modelo multiplicativo, por ejemplo, expresa el RR conjunto como:

$$RR_{11} = RR_{10} \times RR_{01}$$

y el riesgo (o tasa) conjunto como:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

donde los subíndices en el primer y segundo lugar indican la presencia (1) o ausencia (0) del primer y segundo factor, respectivamente. Resulta que si los datos se ajustan a este modelo, en un análisis

estratificado que controla para cualquiera de los dos factores, los RR específicos por estrato para el otro factor serán iguales entre ellos.

Para demostrar esto, simplemente dividimos ambos lados de la segunda forma del modelo por R_{01} :

$$\frac{R_{11}}{R_{01}} = \frac{R_{10} \times R_{01}}{R_{00} \times R_{01}} = \frac{R_{10}}{R_{00}}$$

Examinemos el término de la izquierda y el término de la derecha. En ambos términos, el primer factor está presente en el numerador de la tasa pero ausente en el denominador. Así, cada uno de estos términos es una razón de tasas para el efecto del primer factor.

$$\begin{array}{ccc} \text{RR para el 1er factor} & = & \text{RR para el 1er factor} \\ \text{(2o. factor presente)} & & \text{(2o. factor ausente)} \end{array}$$

Mientras tanto, el segundo factor está presente tanto en el numerador como en el denominador de la tasa a la izquierda, y ausente de ambos componentes a la derecha. Dado que cada tasa requiere un número de casos y un denominador en personas o tiempo persona, cada RR debe venir de una tabla 2 x 2 que contenga casos expuestos, casos no expuestos, sanos expuestos o tiempo-persona, y sanos no expuestos o tiempo-persona.

Así, estos dos RR corresponden a un análisis estratificado que controla para el segundo factor como presente vs. ausente. Su igualdad significa que el RR para el daño con respecto al primer factor es igual en ambos estratos del segundo factor. Si hubiésemos dividido originalmente por RR_{01} , en vez de RR_{10} , hubiésemos encontrado que el RR para el segundo factor es igual en ambos estratos del primer factor.

Para ver la relación con algunos números familiares, aquí presentamos una parte de los datos de Mann y cols. presentados anteriormente:

Incidencia de infarto agudo de miocardio (IAM) en usuarias de anticonceptivos orales (ACO) de 40-44 años de edad, por 100,000 mujeres-años

Cigarrillos/día	ACO*	$\overline{\text{ACO}}^*$	RR**	RA***
0-14	47 (R ₀₁)	12 (R ₀₀)	4	35
15 +	246 (R ₁₁)	61 (R ₁₀)	4	185

* Tasa por 100,000 mujeres-años

** RR=riesgo relativo (razón de tasas)

*** RA= riesgo atribuible (diferencia de tasas, diferencia absoluta)

Vimos en el capítulo sobre la modificación del efecto que la tabla entera se adaptaba bastante bien a un modelo multiplicativo. Si volvemos a mirar la tabla vemos que los RR de las dos primeras filas (3) son iguales y los de las segundas dos filas (4, mostradas arriba) son iguales.

Supongamos que las cuatro tasas de la tabla son representadas por R₀₀, R₁₀, R₀₁, y R₁₁, siendo que el primer subíndice representa el hábito de fumar y el segundo ACO. Entonces podemos escribir:

$$R_{11} = \frac{R_{10} \times R_{01}}{R_{00}}$$

y

$$246 \approx \frac{61 \times 47}{12}$$

La igualdad anterior es sólo aproximada, pero en realidad las razones de las tasas no eran exactamente iguales (3.92 versus 4.03.) Por lo tanto, la aseveración de que el RR es igual en todos los estratos es equivalente a decir que los datos se adaptan a un modelo multiplicativo.

Podríamos haber demostrado este hecho de igual manera usando el OR (¡inténtalo!). Si hubiésemos usado la diferencia de tasas o de riesgos como el parámetro de interés, hubiésemos encontrado (por una resta más que una división) que la igualdad de las medidas de diferencia específicas por estrato es equivalente a que los datos se adapten a un modelo aditivo (prueba esto también!).

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$R_{11} - R_{01} = R_{10} + R_{01} - R_{00} - R_{01} = R_{10} - R_{00}$$

Esta relación entre los modelos multiplicativo y aditivo por un lado y el análisis estratificado por el otro es esencialmente banal, pero también esencial, de manera que vale la pena dedicarle un poco más de tiempo.

El análisis estratificado como "tablas" o "columnas"

Un análisis estratificado que involucra un daño dicotómico, una exposición dicotómica, y una variable de estratificación dicotómica requiere dos tablas 2 x 2, cada una con dos columnas de casos y sanos (o tiempo-persona). Si visualizamos los datos como columnas, más que tablas, casi podemos "ver" la estructura del modelo multiplicativo o aditivo en la estratificación. Por ejemplo, presentamos a continuación dos tablas 2 x 2 creadas con números hipotéticos que producen tasas similares a los datos ya citados de Mann *y cols.* y presentados en la forma de nuestros análisis estratificados anteriores.

Datos hipotéticos sobre la incidencia de infarto agudo de miocardio (IAM) en usuarias de anticonceptivos orales (ACO) por 100,000 mujeres-años, controlados para el hábito de fumar (tomado de Mann y cols.)

Cigarrillos / día	15+	15+	0-14	0-14
uso ACO	ACO	ACO	ACO	ACO
ECC	49	11	19	8
Mujeres-año*	20	18	40	66
Tasa**	245	61	48	12
	R ₁₁	R ₁₀	R ₀₁	R ₀₀

* (en miles)

** por 100,000 (algunos difieren levemente de los de Mann *y cols.*)

La tabla 2 x 2 de la izquierda muestra la relación entre los ACO y ECC entre mujeres que fuman 15+ cigarrillos / día; la tabla a la derecha muestra la relación entre mujeres que fuman menos de 15 cigarrillos / día. **De manera equivalente**, las cuatro columnas muestran el número de casos, mujeres-años de riesgo, y la tasa de ECC en, de izquierda a derecha:

15+ cigarrillos / día usuarias de ACO	(R ₁₁ , = 49/20,000 = 245/100,000 m-a)
15+ cigarrillos/ día no usuarias de ACO	(R ₁₀ , = 11/18,000 = 61/100,000 m-a)
0-14 cigarrillos/ día usuarias de ACO	(R ₀₁ , = 19/40,000 = 48/100,000 m-a)

0-14 cigarrillos/ día no usuarias de ACO ($R_{00} = 8/66,000 = 12/100,000$ m-a)

De igual manera, todas las estimaciones de RR relevantes pueden ser obtenidas formando razones de las tasas apropiadas, p.ej.:

Razones de tasas

Ambos factores (versus ninguno) $RR_{11} = R_{11} / R_{00} = 245/12 = 20$

Hábito de fumar (1er factor) actuando sólo $RR_{10} = R_{10} / R_{00} = 61/12 = 5$

Hábito de fumar (1er factor) en presencia de ACO (2o factor) $RR_{F|A} = R_{11} / R_{01} = 245/48 = 5$

ACO (2o. factor) actuando sólo $RR_{01} = R_{01} / R_{00} = 48/12 = 4$

ACO (2o. factor) en presencia del hábito de fumar (1er factor) $RR_{A|F} = R_{11} / R_{10} = 245/61 = 4$

De manera que el modelo multiplicativo para los efectos conjuntos, introducido en el capítulo sobre modificación del efecto, es equivalente a análisis estratificados en que la medida de razón es la misma en todos los estratos. Lo mismo puede ser demostrado para el modelo aditivo y la medida de diferencia, aunque no con estos datos dado que no se ajustan a un modelo aditivo.

"Homogeneidad" y "heterogeneidad" vs. "sinergia" o "antagonismo"

En la terminología usada cuando discutimos las medidas resumen de asociación, se dice que las medidas específicas por estrato son "homogéneas" cuando son iguales y "heterogéneas" cuando son significativamente diferentes. Obviamente, una medida resumen es mejor en una situación en que la medida que está siendo resumida es homogénea en los estratos. En el caso habitual, para una medida de razón de efecto, la homogeneidad entre los estratos es equivalente a tasas, odds o razones que se adaptan a un modelo multiplicativo de efectos conjuntos. En el caso de las medidas de diferencias (absolutas), la homogeneidad es equivalente a un modelo aditivo de efectos conjuntos. "Modificación del efecto" (o modificación de la medida del efecto", en la terminología nueva de Greenland y Rothman) significa heterogeneidad para esa medida.

Los análisis epidemiológicos de factores de riesgo típicamente emplean medidas de efecto de tipo razón. En escala de razón, las medidas de resumen de los análisis estratificados (y como veremos pronto las obtenidas de los modelos matemáticos) son derivadas bajo la premisa de la homogeneidad de los efectos a través de los estratos, lo que es equivalente a un modelo multiplicativo de los efectos conjuntos esperados, y también consistente en general con un modelo aditivo. De manera que el término "modificación del efecto" se aplica más habitualmente a situaciones en que la medida del efecto de razón es heterogénea en los estratos aún si ocurriera (como la excepción por cierto) que los datos no se adecuan a un modelo aditivo! En contraste, "sinergia" desde la perspectiva de salud pública se considera hoy día, como un mayor efecto observado que el esperado de un modelo **aditivo**. De manera que cuando hay una "modificación del efecto del riesgo relativo" hay generalmente "interacción desde la perspectiva de salud pública".

Dicha inconsistencia es sin duda una indicación de que estos conceptos fueron diseñados por mortales, y no por poderes superiores, y también resalta el hecho de que "la modificación del efecto" es relativa a la escala de medida o el modelo esperado para los efectos conjuntos. Podemos esperar que a medida que la disciplina evoluciona, aparecerá una nueva síntesis que evitará este enfoque "esquizofrénico". Mientras tanto, tal vez ayude la siguiente tabla resumen.

Homogeneidad, heterogeneidad, y modificación del efecto con relación a los modelos aditivo y multiplicativo

	Perspectiva del impacto de salud	Perspectiva de la medida de resumen
1. Los datos se ajustan a un modelo aditivo (homogeneidad de la medida de diferencia en los estratos)	No hay interacción (no hay sinergismo)	No hay modificación del efecto (de la medida de diferencia), <u>la medida de diferencia resumen es adecuada</u> Modificación del efecto (de una medida de razón), una <u>razón</u> resumen <u>no</u> es una medida adecuada
2. Los efectos conjuntos exceden las expectativas bajo un modelo aditivo ("supra-aditivo" – puede o no ser igual o exceder el modelo multiplicativo)	Interacción de Salud Pública (efecto sinérgico)	Modificación del efecto (de la medida de diferencia, tal vez también medida de razón), una medida de resumen de diferencias no es adecuada (tal vez también una medida de resumen de razón)
3. Los datos se ajustan a la expectativa bajo un modelo multiplicativo (homogeneidad de la medida de razón en los estratos)	Interacción de Salud Pública (efecto sinérgico)	No hay modificación del efecto (de una medida de razón), la medida resumen de razón es adecuada
4. Los efectos conjuntos exceden las expectativas bajo un modelo multiplicativo ("supra-multiplicativo")	Interacción de Salud Pública (efecto sinérgico)	Modificación del efecto (de medidas de diferencia y razón), no son adecuadas las medidas de diferencia y razón

Tipos de medidas resumen globales

Cuando las medidas crudas y específicas por estrato son todas similares, la medida cruda sirve adecuadamente de medida de resumen. Cuando hay una heterogeneidad significativa, será necesario presentar las propias medidas específicas por estrato. Queda la situación en que las medidas específicas por estrato son suficientemente homogéneas que una medida de resumen de algún tipo es de interés pero, debido al fenómeno de confusión, la medida cruda no puede cumplir este rol. En dichos casos la medida cruda cae por fuera del rango de las medidas específicas por estrato o tan lejos del centro del rango que sería un resumen engañoso. Estas circunstancias requieren una medida ajustada, generalmente algún tipo de promedio ponderado de las medidas específicas por estrato.

Supongamos que todas las medidas específicas por estrato se acercan mucho unas a otras (i.e., son homogéneas), de manera que tendemos a considerarlas todas como estimaciones del mismo parámetro poblacional (la "verdadera" medida de asociación), más o menos alguna distorsión de la variabilidad de muestreo (si queremos cuantificar la compatibilidad de los datos con este supuesto, podemos emplear una prueba estadística, como la prueba de chi cuadrada de homogeneidad de Breslow-Day, para evaluar el rango esperado de variabilidad al azar). Si hay un "verdadero" valor subyacente, ¿cuál es la mejor manera de estimarlo? Obviamente con algún tipo de promedio ponderado, ¿pero de qué tipo?

Si hay sólo una "verdadera" medida de asociación y cada uno de los estratos da una estimación de esa verdadera medida, entonces queremos poner más atención en los estratos que presentan las "mejores" estimaciones (más precisas). El procedimiento de promediar que empleemos debe ponderar más las estimaciones de esos estratos. Podemos cumplir con ese objetivo usando como pesos la precisión estimada de cada estimación específica por estrato. Dicho promedio ponderado da la mejor estimación de la "verdadera" medida de asociación bajo los supuestos sobre los cuales hemos estado trabajando. (Rothman se refiere a las estimaciones resumen derivadas de esta manera como estimaciones "directamente puestas en común". Sin embargo, el término "puestas en común" (en inglés pooled") a veces se usa para denominar el total crudo de un conjunto de estratos o estudios.

[Nota: el cálculo de medidas de asociación resumen a menudo no es incluido en los cursos introductorios. Los principales conceptos que los estudiantes principiantes deben retener de toda esta discusión de las medidas de resumen son: (1) las medidas de resumen son típicamente promedios ponderados; (2) si las medidas específicas por estrato son significativamente diferentes unas de otras, cualquier medida de resumen (cruda o ajustada) da una imagen incompleta de la relación, por lo menos en esa escala de medida (p.ej., diferencias o razones), de manera que el investigador debe informar los resultados específicos por estrato y tomar esa heterogeneidad en cuenta al interpretar una medida de resumen; (3) si la medida cruda cae por fuera del rango de las medidas específicas por estrato, entonces hay fenómeno de confusión y debe utilizarse una medida ajustada de asociación para resumir la relación; (4) si la medida de asociación cruda cae cómodamente dentro del rango de las medidas específicas por estrato, no hay fenómeno de confusión y por lo tanto es una opción aceptable para una medida de resumen; (5) dada la "no colapsabilidad" del OR (ver Apéndice del capítulo sobre el Fenómeno de Confusión), si el daño no es raro una comparación del odds ratio crudo con los odds ratio específicos por estrato puede no brindar información sobre el fenómeno de confusión. La siguiente discusión se presenta para los más avanzados o los más aventureros. Otros pueden desear volver a esta sección durante su próximo curso de métodos epidemiológicos.]

Estimaciones de medidas de resumen ponderadas por precisión– tema opcional

La **imprecisión** de una estimación puede ser definida como la amplitud del intervalo de confianza alrededor de ella. Dado que estamos acostumbrados a estimar intervalos de 95% de confianza sumando y restando 1.96 veces el error estándar de la estimación, la amplitud total es $2 \times 1.96 \times$ error estándar. Dado que todas estas amplitudes incluirán el 2×1.96 , toda la variabilidad de la precisión está incluida en los errores estándar. Cuanto **menor** el error estándar, mayor el grado de precisión, de manera que los pesos que consisten en los recíprocos de los errores estándar cumplirán con la ponderación por precisión. De hecho, los pesos usados son los cuadrados de estos recíprocos y se denominan "pesos inversos de varianza".

Medida de diferencia – la diferencia de intervalos de confianza (DIC)

La varianza de la DIC es fácil de calcular, dado que la DIC es simplemente la diferencia de dos proporciones. Cuando hay por lo menos 5 "éxitos", la varianza de una proporción (p) puede ser estimada simplemente como $p(1-p)/n$, donde n es el tamaño de la muestra. La varianza de una suma o una diferencia de dos variables aleatorias independientes es la suma de sus varianzas. De manera que la varianza (cuadrado del error estándar) de un DIC es:

$$\begin{aligned} \text{var(DIC)} &= \text{var(IC}_1) + \text{var(IC}_0) \\ [\text{e.e.}(\text{DIC})]^2 &= \frac{\text{IC}_1 (1 - \text{IC}_1)}{n_1} + \frac{\text{IC}_0 (1 - \text{IC}_0)}{n_0} \end{aligned}$$

Usando la notación de nuestras tablas 2×2 , donde "a" representa los casos expuestos y "b" representa los casos no expuestos, podemos escribir esta fórmula como sigue:

$$\begin{aligned} [\text{e.e.}(\text{DIC})]^2 &= \frac{a/n_1 (c/n_1)}{n_1} + \frac{b/n_0 (d/n_0)}{n_0} \\ [\text{e.e.}(\text{DIC})]^2 &= \frac{ac}{n_1^3} + \frac{bd}{n_0^3} = \frac{n_0^3 ac + n_1^3 bd}{n_1^3 n_0^3} \end{aligned}$$

cuyo recíproco (y el peso específico por estrato, w) es:

$$w = \frac{1}{[e.e.(DIC)]^2} = \frac{n_1^3 + n_0^3}{n_0^3ac + n_1^3bd}$$

Este valor se calcula para cada estrato y se utiliza como peso para el DCI de ese estrato. Para dos estratos (indicados por los subíndices 1 y 2):

$$\text{DIC resumen} = \frac{w_1 \text{DIC}_1 + w_2 \text{DIC}_2}{w_1 + w_2}$$

Dado que acabamos de calcular las varianzas de las estimaciones de DIC específicas por estrato y como la varianza de la estimación de la DIC resumen es simplemente la suma de las varianzas, la varianza de esta estimación de DIC resumen es simplemente $1/w_1 + 1/w_2$, y el intervalo de confianza del 95% para la estimación de la DIC resumen es:

$$\text{IC del 95\% para DIC (resumen)} = \text{DIC} \pm 1.96 \sqrt{1/w_1 + 1/w_2}$$

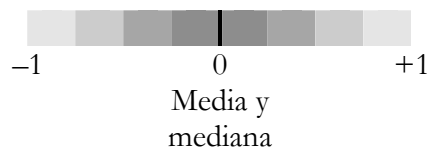
Medidas de razón

Una variable aleatoria uniforme que es una proporción tiene una distribución simétrica, dado que sus posibles valores caen entre 0 y 1, y la media de la distribución (0.5) es igual que su mediana. De manera similar, la distribución de la DIC, basada en la diferencia de dos proporciones aleatorias uniformes, es simétrica, dado que cae entre -1 y 1 y tiene una media y una mediana a nivel de su valor nulo, 0.

Distribución de una proporción:

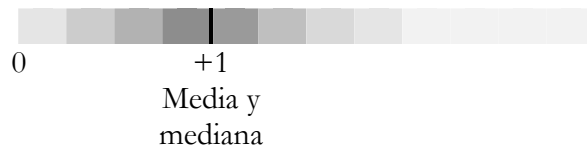


Distribución de una diferencia de dos proporciones:



Dada esta simetría, se podrían usar las estimaciones de las varianzas basadas en una distribución aproximadamente normal. Las medidas de razón, sin embargo, no tienen distribuciones simétricas. La razón de intervalos de confianza (RIC) (una razón de dos proporciones) y el OR (una razón de odds, que a su vez son razones de dos proporciones no-independientes) tienen ambas un límite inferior de 0, una mediana (y valor nulo) de 1.0, y no tienen límite superior.

Distribución de RIC, RDI, OR



Esta asimetría hace que el uso de una aproximación normal sea más problemático. Sin embargo, el logaritmo de una medida de razón sí tiene una distribución simétrica, de manera que se puede usar la aproximación normal.

Distribución de $\ln(\text{RIC})$, $\ln(\text{RDI})$, $\ln(\text{OR})$:



Por lo tanto, las varianzas de RIC, RDI y OR se estiman usando una transformación logarítmica.

Medidas de razón – RIC:

El logaritmo natural de la RIC es:

$$\ln(\text{RIC}) = \ln \left[\frac{\text{IC}_1}{\text{IC}_0} \right] = \ln(\text{IC}_1) - \ln(\text{IC}_0)$$

Si cada IC específico por estrato es una proporción aleatoria independiente, la varianza del logaritmo de la estimación de la RIC específica por estrato es la suma de las varianzas de los logaritmos de las estimaciones de los IC específicos por estrato.

$$\text{Var}(\ln(\text{RIC})) = \text{Var}(\ln(\text{IC}_1)) + \text{Var}(\ln(\text{IC}_0))$$

La varianza de estos logaritmos se obtiene usando una aproximación de series de Taylor como sigue: (Kleinbaum, Kupper, and Morgenstern; Rothman and Greenland):

$$\text{Var}(\ln(\text{RIC})) \approx \frac{c}{an_1} + \frac{d}{bn_0} = \frac{bcn_0 + adn_1}{abn_1n_0}$$

de manera que los pesos específicos por estrato son:

$$w = \frac{1}{\text{Var}(\ln(\text{RIC}))} = \frac{abn_1n_0}{adn_1 + bcn_0}$$

Entonces para dos estratos, la medida resumen ponderada por precisión $\ln(\text{RIC})$ es:

$$(\ln(\text{RIC})) \text{ resumen} = \frac{w_1 \ln(\text{RIC}_1) + w_2 \ln(\text{RIC}_2)}{w_1 + w_2}$$

Para obtener una estimación resumen para la RIC, el $\ln(\text{RIC})$ debe ser convertida ahora a la escala natural a través de la exponenciación:

$$\text{RIC resumen} = \exp(\text{resumen } \ln(\text{RIC}))$$

De nuevo, podemos usar w_i para obtener la varianza de la estimación global de la RIC, aunque de nuevo necesitaremos una transformación de la escala. La varianza de la estimación del $\ln(\text{RIC})$ resumen es simplemente $1/w_1 + 1/w_2$, de manera que el intervalo de confianza del 95% es:

$$\text{intervalo de confianza del 95\% para } \ln(\text{RIC}) = \ln(\text{RIC}) \pm 1.96 \sqrt{1/w_1 + 1/w_2}$$

$$\text{intervalo de confianza del 95\% para RIC} = \exp[\ln(\text{RIC}) \pm 1.96 \sqrt{1/w_1 + 1/w_2}]$$

Medidas de razón – OR:

Una estimación aproximada de la varianza del $\ln(\text{OR})$ en el i ésimo estrato es:

$$\text{Var}(\ln(\text{OR})) = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

De manera que el peso del i ésimo estrato es:

$$w_i = \frac{1}{\left(\frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i} \right)}$$

(Nota: un número pequeño en cualquiera de las celdas hace que la varianza sea grande y, por lo tanto, el peso pequeño.) El $\ln(\text{OR})$ global se estima entonces como:

$$\ln(\text{OR}) = \frac{w_1 \text{OR}_1 + w_2 \text{OR}_2}{w_1 + w_2}$$

y el OR global como:

$$\text{OR} = \exp(\ln(\text{OR}))$$

La varianza del $\ln(\text{OR})$ es $1/\sum w_i$ y puede ser utilizado para obtener un intervalo de confianza del 95% para el $\ln(\text{OR})$, lo cual puede ser exponentiado para obtener un intervalo de confianza para el OR, igual que para la RIC.

Medidas de resumen de Mantel-Haenszel:

Nathan Mantel y William Haenszel, en su trabajo clásico de 1959, introdujeron un OR resumen que es particularmente fácil de calcular:

$$\text{OR}_{\text{MH}} = \frac{\sum [a_i d_i / n_i]}{\sum [b_i c_i / n_i]}$$

Rothman muestra que el OR_{MH} es un promedio ponderado, con pesos específicos de $b_i c_i / n_i$. Estos pesos también son basados en la precisión, dado que son inversamente proporcionales a la varianza del logaritmo de los OR específicos por estrato. La diferencia entre estos pesos y los de la fórmula anterior es que para el OR_{MH} los pesos se basan en varianzas que se aplican en el supuesto de que los OR son 1.0, mientras que los de la fórmula anterior no requerían ese supuesto. Sin embargo, las dos medidas de resumen producen resultados similares y son esencialmente equivalentes cuando los OR específicos por estrato no se alejan mucho de 1.0. Una ventaja del OR_{MH} es que puede ser usado con datos escasos, incluyendo una celda "ocasional" con contenido cero (ver Rothman).

Las fórmulas para estas y otras medidas de asociación resumen (DDI, RDI), intervalos de confianza, y pruebas globales de significancia estadística pueden hallarse en los textos de Kleinbaum, Kupper, y Morgenstern; Hennekens y Buring; Schlesselman; y Rothman. El texto de Rothman incluye una discusión sobre los métodos de máxima verosimilitud para estimar las medidas resumen.

Aunque esta discusión ha puesto el énfasis sobre la utilidad de las medidas de resumen en los análisis en que hay poca heterogeneidad entre los estratos, a veces un investigador puede desear presentar una medida de resumen aún cuando hay una importante heterogeneidad presente. En

estas situaciones se usan medidas estandarizadas (más que ajustadas) (ver Rothman y/o Kleinbaum, Kupper, y Morgenstern).

[Nota: Ahora es el momento de reintegrarse si saltaste la sección sobre esquemas de ponderación para medidas de asociación resumen. Por otro lado, si ya tienes familiaridad con los modelos matemáticos puedes saltar esta sección o leerla por arriba

Diseños apareados

Como vimos en el capítulo sobre el fenómeno de confusión, cuando el diseño de estudio utiliza el apareamiento, puede ser necesario utilizar el control en el análisis para las variables apareadas. En un estudio de seguimiento, el análisis de los datos sin tener en cuenta el apareamiento puede no brindar las estimaciones más precisas, pero las estimaciones no serán sesgadas. Un estudio caso-control con controles apareados, sin embargo, puede brindar estimaciones sesgadas si el apareamiento no es considerado en el análisis. Así, las variables apareadas deben ser siempre controladas al analizar los datos de un estudio caso-control apareado. Si el resultado no es diferente al del análisis no apareado, entonces, por simplicidad, se puede utilizar en análisis no apareado.

La forma más directa de controlar para las variables de apareamiento es a través del análisis estratificado como se presentó anteriormente. Si el apareamiento por categorías es utilizado (i.e., apareamiento por frecuencia, p.ej., por grupos de edad y sexo), el procedimiento es un análisis estratificado controlando para esas variables. Si se emplea un apareamiento individual (p.ej., apareamiento en pares, de a tres, etc.), cada par o conjunto de valores apareados es considerado un estrato.

Supongamos que los datos de un estudio caso-control usando apareamiento en pares se presentan en la siguiente tabla:

Par	Caso	Control	Tipo
6	n	n	A
9	n	n	A
10	n	n	A
1	s	n	B
2	s	n	B
5	s	n	B
3	n	s	C

8	n	s	C
4	s	s	D
7	s	s	D

Si cada par es un estrato, el análisis estratificado de los datos anteriores consiste en 10 tablas, cada una con un caso y un control. Habrá 3 tablas como la tabla A, 3 como la tabla B, 2 como la tabla C y 2 como la tabla D.

	Exp	No Exp	Exp	No Exp	Exp	No Exp	Exp	No Exp
Caso	0	1	1	0	0	1	1	0
Control	0	1	0	10	1	0	1	0
Tipo	A		B		C		D	

Aunque no podemos calcular ninguna medida de asociación específica por estrato, podemos calcular un odds ratio resumen de Mantel-Haenszel usando la fórmula:

$$OR_{MH} = \frac{\sum[a_i d_i / n_i]}{\sum[b_i c_i / n_i]}$$

donde a_i , b_i , c_i , d_i son las celdas de la tabla i , y n_i es el número de participantes de la tabla i . Esta fórmula general se hace más simple para los datos apareados en pares, porque todos los n_i son 2 y muchos de los términos desaparecerán debido a las celdas con contenido cero. Cuando quitamos estos términos y multiplicamos el numerador y el denominador por 2 (n_i), nos quedamos con (a) uno ($a_i d_i$) en el numerador de cada tabla en que el control es expuesto y el caso no lo es (tabla tipo B); y b) uno ($b_i c_i$) en el denominador de cada tabla en que el caso es expuesto y el control no lo es (tabla tipo C). Para los datos anteriores:

$$OR_{MH} = \frac{1 + 1 + 1}{1 + 1} = \frac{3}{2} = 1.5$$

De esta manera la fórmula se convierte simplemente en $OR=B/C$, donde B es el número de pares discordantes en que el caso es expuesto y C es el número de pares en que el control es expuesto. Señalamos que los pares concordantes (tipos A y D) no tienen efecto sobre el OR.

Modelos matemáticos

Anteriormente en el capítulo mostramos que cuando el RR es igual en todos los estratos de un análisis estratificado, los datos se ajustan a un modelo multiplicativo, y vice-versa. También dijimos que para las medidas de diferencia, la igualdad de las medidas de diferencia específicas por estrato es equivalente a que los datos se ajusten a un modelo aditivo. De hecho, estos modelos simples pueden servir de punto de lanzamiento para comprender los modelos matemáticos utilizados para controlar el fenómeno de confusión.

Volviendo al tema del cáncer de mama con relación a la obesidad y /o el uso de reserpina, supongamos que la siguiente tabla muestra los datos de un estudio de cohortes. (Señalemos que esto es hipotético – alguna vez se sospechó que la reserpina estaba relacionada con el riesgo del cáncer de mama, pero esa evidencia ha sido rechazada hoy en día).

**Riesgo de cáncer de mama a los diez años, por obesidad y según uso de reserpina
(datos hipotéticos)**

Factores de riesgo	Numérico(ilustrativo)	Algebraico
Ninguno (riesgo de base)	.01	R ₀₀
Sólo obesidad	.03	R ₁₀
Sólo reserpina	.02	R ₀₁
Reserpina y obesidad	.04	R ₁₁

Así:

R₀₀ indica el riesgo de base (sin reserpina, no obesos)

R₁₀ indica el riesgo para obesidad (sin reserpina)

R₀₁ indica el riesgo para reserpina (sin obesidad)

R₁₁ indica riesgo para reserpina y obesidad

En este ejemplo, el riesgo conjunto se ajusta a un modelo aditivo:

$$RD_{11} = RD_{10} + RD_{01} \quad (\text{las diferencias de riesgo son aditivas})$$

$$R_{11} - R_{00} = (R_{10} - R_{00}) + (R_{01} - R_{00})$$

$$(.04 - .01) = (.03 - .01) + (.02 - .01)$$

$$0.03 = 0.02 + 0.01$$

o, en forma equivalente:

$$R_{11} = R_{10} + R_{01} - R_{00}$$

$$0.04 = 0.03 + 0.02 - 0.01$$

También podemos expresar los diferentes riesgos en términos del riesgo de fondo y el "efecto" de los factores de riesgo:

$$R_{10} = R_{00} + RD_{10} \quad (.03 = .01 + .02) \quad (\text{"efecto" de la obesidad})$$

$$R_{01} = R_{00} + RD_{01} \quad (.02 = .01 + .01) \quad (\text{"efecto" de la reserpina})$$

$$R_{11} = R_{00} + RD_{01} + RD_{10} \quad (.04 = .01 + .02 + .01) \quad (\text{Ambos})$$

Señalemos que la palabra "efecto" se usa aquí por convención y por conveniencia, más que para sugerir causalidad.

Otra manera de apreciar estas diversas ecuaciones de riesgos es tratando de ponerlas todas juntas en una única ecuación con "interruptores" para los efectos que se "encienden". El riesgo de base R_{00} está siempre presente, de manera que requerimos sólo dos "interruptores", uno para el efecto de la obesidad y uno para el efecto de la reserpina:

Riesgo	=	R_{00}	+	Efecto de obesidad	×	"Interruptor" de obesidad	+	Efecto de reserpina	×	"Interruptor" de reserpina
Riesgo	=	R_{00}	+	RD_{10}	×	<input style="width: 100px; height: 20px;" type="text"/>	+	0.01	×	<input style="width: 100px; height: 20px;" type="text"/>
Riesgo	=	0.01	+	0.02	×	<input style="width: 100px; height: 20px;" type="text"/>	+	0.01	×	<input style="width: 100px; height: 20px;" type="text"/>

Cuando un "interruptor" es encendido (=1) pasa a importar el 0.02 (efecto de obesidad) o 0.01 (efecto de reserpina), haciendo mayor el riesgo calculado por el modelo.

Riesgo	=	R ₀₀	+	Efecto de obesidad	"Interruptor" de obesidad	+	Efecto de reserpina	"Interruptor" de reserpina	
Riesgo	=	0.01	+	0.02	× <input type="text" value="0"/>	+	0.01	× <input type="text" value="0"/>	= 0.1
Riesgo	=	0.01	+	0.02	× <input type="text" value="1"/>	+	0.01	× <input type="text" value="0"/>	= 0.03
Riesgo	=	0.01	+	0.02	× <input type="text" value="0"/>	+	0.01	× <input type="text" value="1"/>	= 0.02
Riesgo	=	0.01	+	0.02	× <input type="text" value="1"/>	+	0.01	× <input type="text" value="1"/>	= 0.04

Ahora tenemos un "modelo" que podemos usar para calcular el riesgo para cualquier combinación de los dos factores de riesgos. Aunque este ejemplo es trivial, además de inventado, la estructura del modelo es la misma que en la regresión lineal múltiple. Para ver nuestro modelo de una manera más sofisticada, simplemente tenemos que reemplazar los "interruptores" con variables que puedan tomar valores de 0 o 1.

Modelos lineales:

Si permitimos que:

B = 1 si la mujer es obesa y 0 si no lo es

E = 1 si la mujer usa reserpina y 0 si no la usa

nuestro modelo se convierte en:

$$R(B,E) = R_{00} + (RD_{10})B + (RD_{01})E$$

Sustituyendo con valores obtenidos de la tabla:

$$R(B,E) = .01 + (0.02)B + (0.01)E$$

Nuestras dos variables dicotómicas (B=1 o 0, E=1 o 0) nos dan cuatro posibles combinaciones para el uso de reserpina y la obesidad, igual que nuestro modelo de interruptores. Ahora tenemos un modelo lineal de aspecto profesional para el riesgo de cáncer de mama, y la diferencia de riesgos (o aumento de riesgo) atribuible a cada factor. Las diferencias de riesgos (0.02, 0.01) se denominan "coeficientes" y a menudo se representan con la letra griega β; el riesgo de base se representa a menudo con la letra griega α.

Talvez te preguntes cuál es el valor de todos los cálculos anteriores, dado que no obtenemos más información de nuestro modelo de la que teníamos en nuestra tabla de riesgos (i.e., en nuestro análisis estratificado.) El logro está en la capacidad de estimar las diferencias de riesgos para cada factor, controlando para el otro(s), a través de la estimación de los coeficientes en el modelo. El poder del modelado está en la capacidad de utilizar los datos de estudio para estimar los coeficientes del modelo usando una técnica estadística conocida como análisis de regresión. Los coeficientes estimados brindan medidas epidemiológicas que son ajustadas para los efectos de las demás variables en el modelo.

Podemos hacer que nuestro modelo parezca más complejo y profesional agregando una tercer variable y notación adicional:

$$\text{Riesgo} = \Pr(\text{Enf}=1 | X_1, X_2, X_3) = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Aquí, expresamos el riesgo como la probabilidad de que la variable de enfermedad es igual a 1 (en oposición a 0) basado en los valores de X_1, X_2, X_3 . Cada β representa la diferencia de riesgo, o aumento de riesgo, para el factor correspondiente (X). La estimación de cada β se basa en la diferencia de riesgo observada en cada estrato de las demás variables.

Este modelo, por supuesto, es igual al que acabamos de desarrollar, salvo que impresiona más, se usan α y β en vez de diferencia de riesgos, X en vez de las letras más familiares, y se ha agregado un tercer término. Por ejemplo, si X_1 es obesidad, X_2 es reserpina, y X_3 paridad, (también codificada como variable dicotómica, p.ej., nulípara vs. múltipara) el coeficiente para X_1 será el promedio ponderado de la diferencia de riesgos para obesidad entre los cuatro subgrupos definidos por los otros dos factores de riesgo:

1. no reserpina-mujeres nulíparas
2. no reserpina-mujeres múltiparas
3. reserpina- mujeres nulíparas
4. reserpina- mujeres múltiparas.

Por lo tanto, cada coeficiente (diferencia de riesgo) se ajustará para los efectos de las otras variables en el modelo, más o menos como si hubiésemos calculado una medida global ajustada en un análisis estratificado.

Igual que en un análisis estratificado, la bondad del coeficiente como una diferencia de riesgos ajustada depende del hecho de que la diferencia de riesgo para la reserpina sea esencialmente igual en los cuatro grupos. El modelo está diseñado para manejar la variabilidad al azar en las diferencias de riesgo, pero no la realidad biológica (o sociológica, etc.) De manera que, igual que con toda medida de resumen, la bondad del coeficiente de regresión lineal (i.e., la estimación de la diferencia global de riesgos) puede estar comprometida por una significativa heterogeneidad de la diferencia de riesgos de otras variables entre los estratos (i.e., por el grado de interacción estadística o modificación del efecto de la diferencia de riesgos.)

Si fuera necesario, el modelo puede incluir alguna heterogeneidad con la ayuda de un término de "interacción" que represente la "diferencia en las diferencias de riesgo". Los términos de interacción generalmente se crean como producto de dos (o más) factores que "interactúan" dado que dicho término es cero si cualquiera de los factores está ausente y uno, solamente cuando ambos están presentes. Por el precio de una letra griega más (γ , gamma) podemos escribir el modelo como sigue:

$$\text{Riesgo} = \Pr(\text{Enf}=1 | X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 X_1 X_2$$

Nos da la posibilidad de que el efecto de X_1 dependa de si está o no presente X_2 (además de que el efecto de X_2 dependa de si está o no presente X_1). Pero si incorporamos términos de interacción para todos los pares o tripletes,... de variables, nos encontraríamos de nuevo en el mismo lugar de donde partimos – una análisis estratificado completo y sin medida de resumen para usar.

El modelo lineal que acabamos de ver tiene muchas características muy atractivas, siendo importante su simplicidad y la facilidad con que se lleva a cabo la estimación estadística de sus coeficientes. Es más, aunque hemos desarrollado e ilustrado el modelo usando sólo variables dicotómicas, o "binarias", el modelo puede utilizar variables discretas y continuas, y con ciertas precauciones, variables ordinales. (Para una variable no-dicotómica, el coeficiente es la diferencia de riesgo para una unidad de aumento en la variable.)

Pero los modelos lineales tienen varias desventajas. Primero, por supuesto, los datos pueden no adaptarse a un modelo aditivo, tal vez más allá del punto en que un único término de interacción podría alcanzar para "arreglar" los datos. En segundo lugar, es posible obtener estimaciones de coeficientes que resultarán en "riesgos" que son menores que cero o mayores de uno. El modelo lineal en los deberes hará eso para algunas combinaciones de factores de riesgo, aunque esta es más una objeción técnica. En tercer lugar, la regresión lineal estima diferencias de riesgo, pero los epidemiólogos habitualmente se interesan en la estimación de medidas de razón de asociación.

Modelos logísticos:

Más ampliamente utilizado en el análisis epidemiológico es el modelo logístico (también llamado un modelo logístico múltiple o el modelo de análisis de logits). En nuestro modelo lineal, anteriormente descrito, elegimos modelar el riesgo sobre una función lineal de dos factores de riesgos. En el modelo logístico, modelamos el "logit" como función lineal de los factores de riesgo:

$$\text{Logit}(\text{Enf}=1 | X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

El logit es el logaritmo natural del odds, $\ln(\text{odds})$ o $\ln[p/(1-p)]$. Puede parecer algo descabellado trabajar con el logit, más que con el riesgo, pero recuerda nuestra explicación para el uso de una transformación logarítmica para estimar la varianza de una medida de razón.

Mientras que el riesgo tiene valores de entre 0 y 1, una situación restrictiva para los matemáticos, el logit no tiene límites. Mientras que la razón de riesgos y el OR tienen su valor nulo (1.0) ubicado en un extremo del rango de posibles valores (cero a infinito), el $\log(\text{OR})$ tiene un rango sin límites, con un

valor nulo (cero) justo en el centro (i.e., tiene una distribución simétrica.) Generalmente usamos los logaritmos Neperianos o "naturales" (base e), abreviados como ln.

Más aún, el modelo logístico que veremos, corresponde al modelo multiplicativo que vimos anteriormente como el supuesto por el análisis estratificado basado en el OR o la razón de riesgos. Es más, los coeficientes que estimamos usando la regresión logística pueden ser convertidos en OR, de manera que obtenemos una medida de asociación de razón.

Es fácil descubrir lo que son los coeficientes logísticos. Dado que el logit es el logaritmo de los odds, la diferencia de dos logits es el logaritmo de un OR (porque la resta de logaritmos corresponde a la división de sus argumentos – ver el apéndice del capítulo 5 sobre Medición de la Enfermedad y la Exposición.)

Supongamos que X_3 es una variable dicotómica (0-1) que indica la ausencia (0) o presencia (1) de una exposición. Primero escribamos el modelo con la exposición "presente" ($X_3=1$), y debajo escribamos el modelo con la exposición "ausente" ($X_3=0$).

$$\begin{array}{l} \text{logit(Enf=1 | } X_1, X_2, X_3=1) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \quad (X_3=1, \text{ presente)} \\ - \quad \text{logit(Enf=1 | } X_1, X_2, X_3=0) = \alpha + \beta_1 X_1 + \beta_2 X_2 + 0 \quad (X_3 = 0, \text{ ausente)} \end{array}$$

Cuando restamos el segundo modelo al primero, se eliminan todos los términos de la derecha salvo el coeficiente X_3 . A la izquierda encontramos una diferencia de dos logits (bastante desordenada), una para cuando X_3 presente y otra cuando X_3 está ausente:

$$\text{logit(Enf=1 | } X_1, X_2, X_3=1) - \text{logit(Enf=1 | } X_1, X_2, X_3=0) = \beta_3$$

Desarrollando los logits:

$$\ln(\text{odds(Enf=1 | } X_1, X_2, X_3=1)) - \ln(\text{odds(Enf=1 | } X_1, X_2, X_3=0)) = \beta_3$$

y dado que una diferencia de logaritmos es el logaritmo de una razón:

$$\ln \left[\frac{\text{odds(Enf=1 | } X_1, X_2, X_3=1)}{\text{odds(Enf=1 | } X_1, X_2, X_3=0)} \right] = \beta_3$$

Una razón de odds es simplemente un OR, en este caso, el OR de la enfermedad con respecto a la exposición representado por X_3 :

$$\ln [OR] = \beta_3$$

$$\exp (\ln [OR]) = \exp(\beta_3)$$

$$OR = \exp(\beta_3)$$

β_3 es la diferencia de logits, de ahí el logaritmo del OR para la exposición representado por X_3 . Por lo tanto $\exp(\beta_3)$ es el OR para una unidad de cambio en X_3 .

Nota: $\exp(\beta_1)$ significa el anti-logaritmo: e, la base de los logaritmos neperianos, elevado a la potencia β_1 . Dado que los coeficientes están en escala logarítmica, para ver el resultado en una escala de OR, necesitamos tomar el antilogaritmo. Por ejemplo, un coeficiente de modelo logístico de 0.7 corresponde a un OR de aproximadamente 2.0 para una variable dicotómica o 2.0 por una unidad de aumento en una variable continua.

De manera que el coeficiente de una variable explicativa dicotómica es el logaritmo del OR del daño con respecto a esa variable explicativa, controlando para los otros términos incluidos en el modelo. El término constante (α) en un modelo sólo con variables de factores de riesgo dicotómicas es el logit de base (log odds) para el daño – el logaritmo del odds de la enfermedad para una persona que no tiene ninguno de los factores de riesgo ($\ln[\Pr(CI_0)/(1-CI_0)]$).

Para un factor de riesgo no dicotómico, podemos comparar los odds en dos niveles diferentes. Por ejemplo, si la edad está expresada por una variable continua X_1 para el número de años, el $\exp(\beta_1)$ da el OR por año de edad y $\exp(10 \beta_1)$ da el OR por década de edad.

El modelo logístico también puede ser escrito en términos de riesgo (i.e., probabilidad) tomando los anti-logaritmos (exponentes) y empleando algo de álgebra. Dejamos la transformación como un ejercicio opcional para aquellos de ustedes que tengan interés en realizarla. El resultado es:

$$\Pr(\text{Enf}=1 | X_1, X_2, X_3) = \frac{1}{1 + \exp(-\alpha - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3)}$$

o, si $L = \text{logit} = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

$$\Pr(\text{Enf}=1 | X_1, X_2, X_3) = \frac{1}{1 + \exp(-L)}$$

De la formulación del riesgo podemos ver fácilmente que la función logística debe tener valores con rango entre cero y uno, una propiedad deseable para modelar el riesgo. Cuando L (el logit) es

"infinitamente negativo", el exponente (-L) es "infinitamente grande" y la estimación de probabilidad es cero. Cuando L es cero, exponente (-L) es 1, y la estimación de probabilidad es un medio.

Supuestos epidemiológicos claves en el modelo logístico

1. el logaritmo del odds de la enfermedad está relacionado en forma lineal con cada uno de los factores de riesgo (variables X), o en forma equivalente, el odds de enfermedad está exponencialmente relacionado a cada uno de los factores de riesgo, o en forma equivalente, el riesgo de enfermedad está relacionado con cada factor de riesgo por una curva logística (sigmoidea);
2. los efectos conjuntos de los factores de riesgo son multiplicativos sobre el odds de la enfermedad (p.ej., si un aumento de una unidad en X_1 por sí sólo multiplica el odds de incidencia dos veces y un aumento de una unidad en X_2 por sí sólo multiplica el odds de incidencia tres veces, un aumento simultáneo de una unidad en X_1 y X_2 multiplica el odds de incidencia seis veces) (Greenland, *AJPH*, 1989; Rothman, *Modern epidemiology*)

Además, para estimar los coeficientes usando procedimientos de regresión, debe suponerse que los sujetos son una muestra aleatoria de observaciones independientes de la población sobre la cual se desea hacer inferencias (Harrell, Lee, and Pollock, 1988.)

Así el modelo logístico corresponde al modelo multiplicativo del análisis estratificado que consideramos anteriormente. El verdadero OR se supone constante en todos los estratos. Igual que con el modelo lineal, es el supuesto de homogeneidad que nos permite estimar coeficientes que son sencillos de interpretar.

Podemos no ser tan estrictos con los supuestos incluyendo los términos de multiplicación, como se ilustró anteriormente para el modelo lineal. Pero entonces los coeficientes son más difíciles de interpretar. Además, si llevamos esta estrategia demasiado lejos nos llevará de vuelta a una situación de estratificación completa y agotará nuestro tamaño muestral, recursos de computación y nuestra imaginación.

Aunque hemos ilustrado ambos modelos con variables dicotómicas (uno-cero), pueden ajustarse fácilmente a variables continuas. De nuevo, la estructura del modelo se basa en un supuesto – el que la relación de la variable dependiente (riesgo, para el modelo lineal, o el logit, para el modelo logístico) con la variable independiente es lineal.

Para algunas relaciones, este supuesto es fácilmente sostenible, p.ej., riesgo de ECC y número de cigarrillos fumados. Para otros, p.ej., riesgo de muerte y peso corporal, la relación tiene forma de U, de manera que un modelo lineal simple o logístico no será adecuado (hay formas más complejas de modelos lineales y logísticas para las variables con forma de U a través de técnicas como la incorporación de los valores de las variables elevados al cuadrado.)

Otra limitación del modelo logístico es que el OR no es la medida epidemiológica de asociación preferida, y cuando el daño o resultado no es raro, no se mantiene la semejanza del OR con el riesgo relativo. Además el modelo no puede proveer lo que el estudio no da. Aunque el modelo logístico de la manera explicada puede ser usado con datos provenientes de estudios caso-control, las estimaciones de riesgo requieren datos de seguimiento. Las matemáticas sólo pueden sustituir los datos hasta cierto punto.

Otros modelos de regresión [Opcional para EPID 168]

Otras formas de modelos matemáticos que los epidemiólogos usan habitualmente para controlar el fenómeno de confusión y para obtener medidas ajustadas de los efectos son las de los modelos de riesgos proporcionales (en inglés "proportional hazards") y de Poisson.

Para un resultado con un período de riesgo prolongado, sobretodo para un resultado que no es raro, a menudo es deseable usar un enfoque de análisis, como la densidad de incidencia o la supervivencia, que tome en cuenta el tiempo hasta la ocurrencia del daño. El modelo de riesgos proporcionales desarrollado por David R. Cox, es un modelo matemático ampliamente usado para analizar los datos epidemiológicos donde el "tiempo hasta la ocurrencia" es importante. El "riesgo" ("hazard" en inglés) convencionalmente representado por la letra griega lambda, λ) es esencialmente el mismo concepto que la densidad de incidencia instantánea.

Para tres variables independientes, el modelo de riesgos proporcionales puede ser escrito como:

$$\log[DI(t|X_1, X_2, X_3)] = \log[DI_0(t)] + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

(i.e., el logaritmo natural de densidad de incidencia como función del tiempo es la suma del logaritmo de una densidad de incidencia subyacente o de fondo más un incremento por cada variable predictora.)

El modelo también puede ser planteado en términos de supervivencia:

$$S(t|X_1, X_2, X_3) = [S_0(t)] \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3)$$

donde $S(t)$ es la probabilidad de que el evento no ha ocurrido hasta el momento t .

El coeficiente de un predictor dicotómico es el logaritmo de la razón de densidad de incidencia [$\ln(RDI)$] para ese predictor:

$$\begin{aligned} \log[\text{DI}(t | X_1, X_2, X_3=1)] &= \log[\text{DI}_0(t)] + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \quad (X_3 \text{ presente}) \\ - \log[\text{DI}(t | X_1, X_2, X_3=0)] &= \log[\text{DI}_0(t)] + \beta_1 X_1 + \beta_2 X_2 + 0 \quad (X_3 \text{ ausente}) \end{aligned}$$

$$\log[\text{RDI}(t)] = \beta_3$$

$$\text{RDI}(t) = \exp(\beta_3)$$

Además de los supuestos requeridos por el modelo logístico, el modelo de riesgos proporcionales de Cox requiere que la razón de riesgos (la RDI) sea constante a través del tiempo, aunque modelos más sofisticados de supervivencia emplean "co-variables dependientes del tiempo" que permiten no ser tan estrictos con este supuesto.

El modelo de Poisson es similar al modelo logístico y al modelo de riesgos proporcionales en que los tres involucran una transformación logarítmica de la función de riesgo (i.e., odds, riesgo) que se estima y tienen una combinación lineal (i.e., una expresión del tipo: $a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots$) del lado derecho de la ecuación. El modelo de Poisson es de particular interés cuando los resultados son muy raros.

Puntos claves [Estudiantes de EPID 168 por favor retomen aquí.]

Algunos principios guía para el análisis multivariado son:

1. Hay que tener presente que nuestros objetivos principales son de describir e interpretar los datos que tenemos, usando criterios informados, percepción y extenso conocimiento además de técnicas.
2. El análisis estratificado es un enfoque potente. Aunque no mantiene su utilidad cuando tratamos de analizar muchas variables simultáneamente podemos controlar para dos o tres a la vez, usando distintos subconjuntos, y permitir que nuestro criterio compense lo que falta. Cabe siempre la posibilidad de que una asociación observada que no es eliminada cuando controlamos para el hábito de fumar, hipercolesterolemia, presión arterial, y comportamiento de Tipo A en forma individual, igual pueda ser debido a un efecto combinado de todos ellos. ¿Pero qué tan probable es, sobretodo si hemos controlado para cada par de factores de riesgo y aún no hemos encontrado la asociación?
3. Al llevar a cabo un análisis estratificado para una variable o combinación de variables, estamos preguntando "¿es responsable esa combinación de variables del resultado observado?". La pregunta tiene que ser una que razonablemente nos podríamos preguntar. Si unos cuantos factores principales individualmente no justifican el hallazgo observado, la probabilidad de que una combinación de ellos lo justifique parece menos probable. [Pero nadie ha demostrado empíricamente esta aseveración.]
4. El modelado matemático es un enfoque potente del análisis de datos. Pero siempre es una pregunta clave si la forma del modelo es apropiada para los datos, y las relaciones subyacentes,

que estamos analizando. La utilización de modelos inapropiados puede producir resultados sesgados. Hay técnicas estadísticas para evaluar la bondad estadística de los modelos empleados ("pregúntale a tu estadístico").

(Es recomendable (ver Greenland, *AJPH*, 1989; 79(3):340-349 y Vanderbroucke JP:Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13) antes de empezar los ejercicios sobre modelado que no pueden ser directamente validados con resultados de análisis estratificados, llevar a cabo análisis paralelos con las mismas variables para validar la selección de modelos y resultados con los datos estratificados.)

Lo esperado para los alumnos de EPID 168:

- Conocer la relación entre el modelo multiplicativo y el análisis estratificado, y (solamente) los conceptos básicos de los modelos de regresión lineal y los modelos de regresión logística. Lo esperado de comprensión del modelado matemático es modesto:
- Conocer las ventajas y desventajas del modelado (en comparación con el análisis estratificado, por ejemplo) como se presentó en el capítulo sobre el fenómeno de confusión.
- Conocer el significado epidemiológico del coeficiente de un término de exposición en un modelo de regresión lineal y saber como el modelo de regresión lineal se relaciona con los análisis estratificados y el modelo aditivo discutido en el capítulo sobre Modificación del Efecto.
- Conocer el significado epidemiológico del coeficiente de un término de exposición en un modelo logístico y saber como ese modelo se relaciona con el análisis estratificado y el modelo multiplicativo.
- Conocer el significado epidemiológico del coeficiente de un término de exposición en el modelo de riesgos proporcionales y que ese modelo es usado para el análisis en términos de densidad de incidencia [supervivencia].
- Para los tres modelos, los coeficientes en un modelo con varias variables están todas "ajustadas" para los efectos de las otras variables en el modelo.

Bibliografía

Rothman, Modern epidemiology, pp. 285-295; Schlesselman, Case-control studies, pp. 227-234.

Harrell, Frank E., Jr.; Kerry L. Lee, Barbara G. Pollock. Regression models in clinical studies. *JNCI* 1988; 80(15):1198-1202.

Godfrey, Katherine. Simple linear regression in medical research. *N Engl J Med* 1985;313:1629-36.

Silberberg, Jonathan A. Estimating the benefits of cholesterol lowering: are risk factors for coronary heart disease multiplicative. *J Clin Epidemiol* 1990; 43(9):875-879.

J. Paul Leigh. Assessing the importance of an independent variable in multiple regression: is stepwise unwise? *J Clin Epidemiol* 1988; 41:669-678.

Vandenbroucke, Jan P. Should we abandon statistical modeling altogether? *Am J Epidemiol* 1987; 126:10-13.

Greenland, Sander. Modeling and variable selection in epidemiologic analysis. *Am J Public Health* 1989; 79:340-349 (Advanced)

Kleinbaum, Kupper, and Morgenstern. *Epidemiologic research: Principles and quantitative methods*. Chapters 16-17.

Breslow and Day. *Statistical methods in cancer research. I. The analysis of case-control studies*. Chapters 3-7 (Primarily Chapter 3).

Wilcosky, Timothy C. and Lloyd E. Chambless. A comparison of direct adjustment and regression adjustment of epidemiologic measures. *J Chron Dis* 1985; 38:849-356.

See also: Flanders, W. Dana; and Philip H. Rhodes. Large sample confidence intervals for regression standardized risks, risk ratios, and risk differences. *J Chron Dis* 1987; 40(7):697-704. [includes SAS program]

Deubner, David C., William E. Wilkinson, Michael J. Helms, et al. Logistic model estimation of death attributable to risk factors for cardiovascular disease in Evans County, Georgia. *Am J Epidemiol* 1980;112:135-143, 1980.

Weinstein, Milton C.; Pamela G. Coxson, Lawrence W. Williams, Theodore M. Pass, et al. Forecasting coronary heart disease incidence, mortality, and cost: the Coronary Heart Disease Policy Model. *Am J Public Health* 1987; 77:1417-1426.

McGee, Daniel; Dwayne Reed, Katsuhika Yano. The results of logistic analyses when the variables are highly correlated: an empirical example using diet and CHD incidence. *J Chron Dis* 1984; 37:713-719.

Breslow and Storer. General relative risk functions for case-control studies. *Am J Epidemiol* 1985;

Szklo, Moyses; F. Javier Nieto. *Epidemiology: beyond the basics*. Gaithersburg MD, Aspen, 2000. Chapter 7 has an excellent presentation of stratified analysis and mathematical modeling at a basic and understandable level.