Capítulo 8: Identificación de factores de riesgo

En ocasiones estamos interesados en conocer la influencia que una serie de variables tienen en una variable respuesta. Cuando la misma era numérica una herramienta que estaba a nuestra disposición era la regresión múltiple. Pero, ¿qué podemos hacer cuando la respuesta es dicotómica? Por ejemplo, ¿qué podemos hacer si la respuesta observada es el desarrollo o no de una enfermedad?

Este tipo de sitiaciones aparecen de manera natural en investigaciones médicas. Citemos unos cuantos ejemplos:

- Se cree que la apnea del sueño obstructiva es un factor de riesgo para la hipertensión arterial. Podríamos reformularlo en lenguaje estadístico como que la variable independiente padecer *apnea obstructiva del sueño*, está asociado (es factor de riesgo) para el la ocurrencia del evento *hipertensión arterial*.
- Se cree que la intervención por laparoscopia para tratar la hernia de hiato ofrece menor riesgo de complicaciones postoperatorias que otra técnica tradicional. La variable respuesta sería *padecer complicaciones* (sí o no), y la variable independiente sería el *tipo de operación*.
- Se cree que fumar es un factor de riesgo para la muerte fetal tardía. Esto se podría formular de varias maneras:
    - Podemos considerar una variable independiente que es la "*madre fuma"* (sí o no) y una variable respuesta (dependiente) que es el "*feto muere*" (sí o no). Nos interesará evaluar cuánto aumenta el riesgo de que se pruduzca el evento de interés (muerte del feto) cuando está presente el factor de riesgo (la madre fuma).
    - Otra aproximación podría ser considerar como variable numérica "número medio de cigarrillos que fuma la madre". En este caso nos puede interesar estimar cuánto aumenta el riesgo de muerte del feto, por cada cigarrilo adicional que fuma la madre diariamente.
    - Si el aumento del riesgo (se evalúe como se evalúe) no parece tener una tendencia constante con el número de cigarrillos, sino que mas bién se puede dividir a las madres en tres categorías "No fuma", "Fuma poco", "Fuma mucho", nos interesará evaluar cómo aumenta el riesgo en las dos últimas categorías con respecto a las madres del primer grupo (grupo de control, o de referencia).

El modelo de regresión logística es muy útil para abordar este tipo de cuestiones bajo la condición de que hayamos tenido en cuenta al realizar el estudio todas las variables importantes para explicar la variable respuesta, y que hagamos el estudio con una muestra suficientemente numerosa y bien distribuida.

Antes de pasar directamente al modelo de regresión logística vamos a refrescar rápidamente una serie de conceptos relacionados con la comparaciones de riesgos.

8.1 Riesgo, Oportunidad, Riesgo Relativo y Odds Ratio

En medicina es frecuente encontrar todos estos términos. Vamos a repasarlos pues será necesario manejarlos con cierta soltura en el resto del capítulo.

En 1 de cada 200 nacimientos ocurre un parto gemelar. Por tanto la **probabilidad** o **riesgo** de que elegido un embarazo al azar éste de lugar a gemelos es de $R_1=1/200$. Esto es simplemente, *el número de casos en que el evento ocurre dividido por el total de casos*.

Hay otra forma de decir lo mismo, que seguramente ha sido tomada del lenguaje usado por los anglosajones en las apuestas. Consiste en la **oportunidad** (del inglés *odds*). Podemos decir que de 200 partos, 1 es gemelar y 199 no lo son. Las apuestas están 1 a 199. Se denomina oportunidad a la cantidad $O_1=1/199$, es decir, *el número de casos en los que el evento ocurre dividido por el número de casos en que no ocurre*. En el fondo es sólo una manera anglosajona de decir lo mismo que con probabilidades. Hasta aquí no hay nada especial.

Compliquemos la cosa un poquito, introduciendo un factor de riesgo. Se observó que entre las mujeres que han tomado ácido fólico para disminuir la probabilidad de espina bífida en sus hijos, ocurrió algo no esperado: 3 de cada 200 partos eran gemelares. Esto corresponde a un riesgo de $R_2=3/200$, o si lo preferimos, a una oportunidad (*odds*) de 3 a 197, $O_2=3/197$.

¿Cómo podemos expresar numéricamente el aumento del riesgo de embarazo gemelar? Hay dos maneras. Una de ellas es más fácil de entender, y la otra tiene mejores propiedades matemáticas.

- **Riesgo Relativo (RR)**: Este es el más simple de entender. Claramente el riesgo ha aumentado por 3, lo que corresponde a un Riesgo Relativo (RR) que es el cociente entre el riesgo de los embarazos expuestos al ácido fólico (factor de riesgo) y los que no han sido expuestos, $RR = R_2/R_1 = (3/200)/(1/200) = 3$.
- **Odds Ratio (OR):** En español se traduce a veces en textos académicos como Oportunidad Relativa, aunque en las publicaciones aparece más frecuentemente con el término inglés. Es parecido al RR, pero usando oportunidades (odds). Es el cociente entre la oportunidad de los embarazos expuestos al ácido fólico (factor de riesgo) y los que no han sido expuestos, $OR = O_2/O_1 = (3/197)/(1/199) = 3.03$.

Desde luego no es tan fácil de interpretar una OR como lo es un RR, aunque en este caso poseen valores muy similares. Esto ocurre siempre que la probabilidad de que ocurra un evento sea cercana a cero, como en el caso de un embarazo gemelar. Cuando las probabilidades del evento no son cercanas a cero, ambas cantidades no son iguales y hay que tener cuidado con no confundirlas.

A pesar de no ser un concepto tan natural la OR como el RR, podemos acostumbrarnos a ella recordando lo siguiente:

- Un valor de OR=1 se interpreta como que no hay tal factor de riesgo, ya que la oportunidad para los expuestos es la misma que para los no expuestos.
- En epidemiología es frecuente intentar localizar factores dañinos. Eso corresponde a buscar valores de OR mayores que uno. Se interpreta como qué se ha localizado un factor de riesgo, pues es mayor la oportunidad de que ocurra el evento a los expuestos al factor que a los controles.

- En los ensayos clínicos, se persigue encontrar tratamientos que reduzcan la frecuencia de un evento (por ejemplo, la muerte del enfermo). En este caso se buscan valores de OR menores que uno. Es decir, que sea menor la oportunidad de que ocurra el evento en los individuos expuestos al tratamiento que en los controles.

Por otro lado la OR tiene muy buenas propiedades matemáticas:

- OR toma valores entre cero e infinito. Esto lo hace muy adecuado para ser modelado matemáticamente. Sobre todo si tomamos su logaritmo, ya que en ese caso cualquier valor es posible. El modelo que consideraremos posteriormente será el *de regresión logística*.
- El modelo logístico de regresión puede usarse para determinar intervalos de confianza para la OR:
    - Si dichos intervalos contienen al valor OR=1, no puede rechazarse que el factor de riesgo (o el tratamiento) no sea tal.
    - En otro caso decimos que aumenta o disminuye la oportunidad del evento en función de que el intervalo de confianza sea de valores mayores o menores que uno respectivamente.
- Cuando se evalúa la eficacia de una prueba diagnóstica es razonablemente simple conocer la sensibilidad y especificidad de la misma, pero los valores predictivos requieren del conocimiento de la prevalencia, que no está siempre disponible. Si realizamos un estudio caso-control, donde la prevalencia de la enfermedad es desconocida, aunque no podamos calcular índices predictivos, siempre podremos estimar la OR. Si la enfermedad (el evento de interés) es rara, podemos considerarla como una aproximación del RR, que tiene una interpretación muy natural.

8.2 Regresión logística

Si tenemos una variable que describe una respuesta en forma de dos posibles eventos (vivir o no, enfermar o no), y queremos estudiar el efecto que otras variables (independientes) tienen sobre ella (fumar, edad), el modelo de regresión logística binaria puede resultarnos de gran utilidad para:

- Dado los valores de las variables independientes, estimar la probabilidad de que se presente el evento de interés (por ejemplo, enfermar.)
- Podemos evaluar la influencia que cada variable independiente tiene sobre la respuesta, en forma de OR. Una OR mayor que uno indica aumento en la probabilidad del evento y OR menor que uno, implica disminución.

Para construir un modelo de regresión logística necesitamos:

- Un conjunto de variables independientes o predictoras, muy en el estilo de la regresión lineal múltiple.
- Una variable respuesta dicotómica. Aquí se diferencia del modelo de regresión múltiple, donde la variable respuesta era numérica.

8.2.1 Codificación de las variables

Para simplificar la interpretación del análisis del modelo de regresión logística es conveniente llegar a cierto convenio en la codificación de las variables. Realmente compensa seguir las siguientes recomendaciones:

- En la **variable dependiente** se codifica como 1 la ocurrencia del evento de interés y como 0 la ausencia.
- Las **variables independientes** pueden ser varias y cada una de un tipo diferente. Analicemos cada caso:
  - **Caso dicotómico**: Se codifica como 1 el caso que se cree favorece la ocurrencia del evento. Se codifica como 0 el caso contrario. Por ejemplo con 0 codificamos típicamente a los individuos no expuestos a un posible factor de riesgo (casos de referencia, controles), y como 1 a los expuestos.
  - **Caso categórico**: Cuando la variable independiente puede tomar más de dos posibles valores podemos codificarlas usando variables indicadoras (*dummy*), como se hacía con el modelo de regresión lineal múltiple. Si estamos usando SPSS, el programa nos ayuda a hacerlo sobre la marcha. Es necesario destacar una modalidad que represente al caso de referencia, y al que le corresponde la codificación con todas las variables indicadoras puestas a 0.
  - **Caso de variable numérica**: pueden darse dos situaciones:
    - Si creemos que por cada unidad que aumente la variable, la OR aumenta en un factor multiplicativo constante, podemos usar la variable tal cual en el modelo. Si tenemos dudas de que esto sea así, o no sepamos ni siquiera lo que significa la frase anterior, mejor olvidamos esta posibilidad y consideramos la siguiente;
    - Si creemos que la variable numérica puede afectar a la respuesta, pero no tenemos muy claro de qué manera, podemos "categorizar" la variable. Esto consiste por ejemplo en estratificar la variable en valores pequeños, medianos y grandes. Los puntos de corte los podemos elegir nosotros manualmente, o usar cortes automáticos basados en que cada categoría tenga el mismo número de observaciones (usando percentiles). La opción de menú "Transformar – Categorizador visual…" de SPSS nos puede ser de gran ayuda.

8.2.2 Requisitos y limitaciones

Además de las mencionadas en cuanto a los criterios para codificar la variables debemos tener en cuenta muchas otras cuestiones para confiar en la validez del modelo. De entre ellas destacamos:

- Los parámetros del modelo se calculan usando una *estimación de máxima verosimilitud*. Estas sólo son válidas cuando para cada combinación de variables independientes tenemos un número suficientemente alto de observaciones. Si los parámetros estimados en el modelo son anormalmente grandes, posiblemente esta condición sea violada. Tal vez se solucione el problema agrupando categorías (donde tenga sentido).
- No debemos introducir variables innecesarias. Ver el punto anterior.

- Ninguna variable relevante debe ser excluida. Si identificamos variables confusoras, tengámoslas en cuenta introduciéndolas en el modelo o estratificando el estudio en submuestras.
- La colinealidad es un problema como ocurría en la regresión lineal múltiple. Si los errores típicos en la estimación de los coeficientes, o los intervalos de confianza son anormalmente grandes, es posible que esta situación se esté dando.

8.2.3 Interpretación del modelo

El modelo de regresión logística puede escribirse como:

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + \cdots + b_n x_n$$

donde p es la probabilidad (riesgo) de que ocurra el evento de interés, las variables independientes están representadas con la letra x, y los coeficientes asociados a cada variable con la letra b. Tal vez con esa expresión el modelo no resulte muy elocuente, pero tras unas transformaciones, que nos mostramos para ahorrar espacio mostramos lo que resulta de mayor interés:

- Dado el valor de las variables independientes, podemos calcular directamente la estimación del riesgo de que ocurra el evento de interés:

  $p = \frac{e^{suma}}{1+e^{suma}}$, donde $suma = b_0 + b_1 x_1 + \cdots + b_n x_n$
- La oportunidad (*odds*) para los individuos de referencia o control (aquellos para los que x_i vale cero, si estamos siguiendo el convenio fijado anteriormente) es $\exp(b_0) = e^{b_0}$
- Si nos fijamos en cualquier otro coeficiente del modelo, la cantidad $\exp(b_i) = e^{b_i}$ coincide con la OR del aumento del valor de x_i en una unidad con respecto a aquellos individuos que presentan los valores de todas las demás variables iguales. Si hemos seguido el criterio de codificación recomendado, y la variable de la que hablamos es dicotómica, esto corresponde a la OR del factor de riesgo x_i. Si la variable es numérica como el número de *bypass* coronarios, estima la OR del factor de riesgo "tener un *bypass* más".

Hay mucho que interpretar en una salida de ordenador en un cálculo de regresión logística. Aquí vamos a mencionar sólo algunas de las que consideramos más interesantes en una primera aproximación:

- **Significación de cada coeficiente del modelo** (basada en el estadístico de Wald): Ofrece el equivalente a la significación de los coeficientes de regresión lineal múltiple. Si una variable independiente resulta no significativa podemos considerar eliminarla del modelo (a menos que esté confundida con otra variable independiente significativa, claro está).
  La significación del estadístico de Wald para el coeficiente b_i es la que corresponde a contrastar la hipótesis nula de que éste vale cero. O lo que es lo mismo que la OR asociada, $\exp(b_i)=1$, es decir, que la variable x_i no es factor de riesgo y por ello podemos olvidarnos de ella a menos que la encontremos confundida con otra que sí lo sea.

- **Exp(b_i)=OR** estimada para el factor x_i. Aparecen en la columna "B" de la salida de SPSS. Más interesante aún son los intervalos de confianza para la Exp(b_i)=OR (si hemos marcado la opción para que los calcule). Si no contienen al valor uno, es señal de que la variable es de interés en el modelo. Estadísticamente es lo mismo que lo mencionado en el punto anterior, pero de este modo tiene más significado clínico.

Ejemplo: Se realizó en un estado de EE.UU, una revisión de los juicios por asesinato con culpable condenado. Se sospecha que cuando el acusado es de raza negra, la probabilidad de ser condenado a pena de muerte es mayor. Es decir, se cree que la raza del acusado es un *factor de riesgo* para el evento *ser condenado a pena de muerte*. En una primera aproximación, ambas variables parecen ser independientes, al menos así lo muestra la prueba de independencia basada en el contraste ji-cuadrado. Se obtiene una significación p=0.194 para la hipótesis de independencia (no puede declararse significativa al nivel de significación habitual de 0.05).

Tabla de contingencia Raza del condenado * Pena de Muerte

Recuento

		Pena de Muerte		
		No	Sí	Total
Raza del condenado	Blanco	432	54	486
	Negro	178	15	193
Total		610	69	679

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	1,687[b]	1	,194

b. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 19,61.

Sin embargo parece que hay una variable confusora que no se ha tenido en cuenta, que es la raza de la víctima. Si observamos con atención la siguiente tabla parece que generalmente los condenados han matado generalmente a gente de su propia raza (relación entre variables independientes). Además cuando la víctima es de raza blanca se observa una frecuencia mayor de condenas a pena de muerte (relación entre la variable raza de la víctima y la variable dependiente). Estos son los ingredientes típicos de la confusión de variables.

Tabla de contingencia Raza del condenado * Pena de Muerte * Raza de la víctima

Recuento

			Pena de Muerte		
Raza de la víctima			No	Sí	Total
Negro	Raza del condenado	Blanco	17	0	17
		Negro	140	4	144
Blanco	Raza del condenado	Blanco	415	54	469
		Negro	38	11	49

Por supuesto hay que tener mucha vista para darse cuenta de todo eso observando la tabla de contingencia, a menos que sepamos lo que vamos buscando. Vamos a utilizar un modelo de regresión logística. La capacidad expresiva del mismo es enorme si se sabe interpretar.

Aplicamos el criterio de codificación de las variables que se adapta a una interpretación sencilla de los resultados:

- **Variable dependiente**: *Condena a pena de muerte*.
  - No hay condena, se codifica como 0.
  - Sí hay condena, se codifica como 1 (es el evento que queremos detectar).
- **Variables independientes**:
  - *Raza del acusado*:
    - Raza blanca se codifica como 0 (caso de referencia o control). No se considera factor de riesgo.
    - Raza negra se codifica como 1. Se considera factor de riesgo.
  - *Raza de la víctima*:
    - Raza negra se codifica como 0 (caso de referencia o control). No se considera factor de riesgo.
    - Raza blanca se codifica como 1. Se considera factor de riesgo.

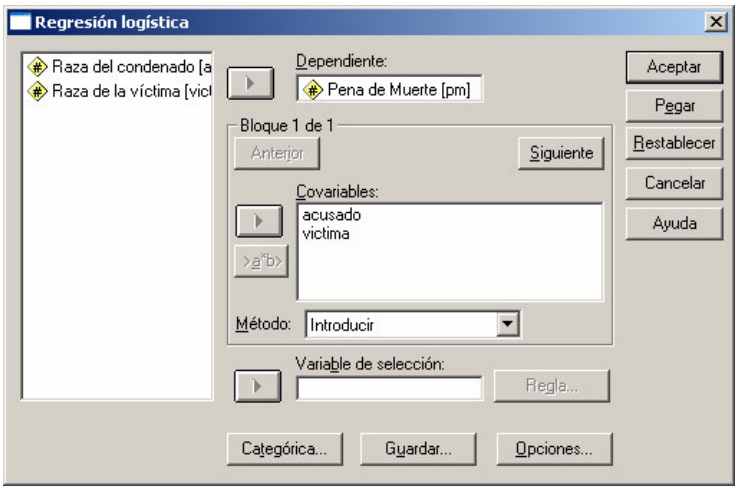
En primer lugar, sólo consideramos como en la prueba ji-cuadrado a la variable raza del acusado. Se aprecia que el coeficiente asociado a la variable "*Raza del acusado*" no es significativo (p=0.196, muy próximo a la significación del contraste ji-cuadrado). Por tanto la OR de este factor de riesgo se considera compatible con uno (no es factor de riesgo). Si observamos el valor Exp(B) correspondiente obtenemos la OR estimada, ¡que es menor que uno! Parece que ser un acusado de raza negra más bien es un factor de protección.

A la misma conclusión llegamos al examinar el intervalo de confianza. Se aprecia que con una confianza del 95%, la OR de este factor para el modelo propuesto está comprendida en [0.371, 1.226]. El valor 1 forma parte de dicho intervalo. Por tanto parece que nuestras sospechas son infundadas.

Variables en la ecuación

	B	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
				Inferior	Superior
acusado	-,394	,196	,674	,371	1,226
Constante	-2,079	,000	,125		

Veamos que ocurre si consideramos en el modelo la inclusión de la segunda variable independiente, "*raza de la víctima*". Usamos la opción de menú en SPSS "Analizar – Regresión – Logística binaria".



Variables en la ecuación

	B	Sig.	Exp(B)	I.C. 95,0% para EXP(B)	
				Inferior	Superior
acusado	,827	,024	2,286	1,118	4,677
victima	2,391	,000	10,928	3,378	35,355
Constante	-4,435	,000	,012		

Ahora sí tenemos que los coeficientes asociados a cada variable son significativos. Las OR son significativamente mayores que uno para el nivel de significación habitual de 0.05 (p=0.024 para la raza de acusado y p aproximadamente cero para la raza de la víctima). Las valores estimados para las OR los tenemos en la columna Exp(B). Como se aprecia los intervalos de confianza al 95% no contienen al valor 1.

Como ejercicio para que el lector practique con lo que es una OR, vamos a obtener algunas interpretaciones:

- Si en un juicio, un acusado únicamente tiene como factor de riesgo el que su raza sea negra, la oportunidad de ser condenado a pena de muerte es mayor en 2.286 veces que si fuese blanco.
- Si en un juicio, un acusado exclusivamente tiene como factor de riesgo el que la raza de la víctima sea blanca, la oportunidad de ser condenado a pena de muerte es mayor en 10.928 veces que si la víctima fuese negra.
- Si un individuo tiene los dos factores de riesgo (es negro y mata a un blanco), la oportunidad que tiene de ser condenado a pena de muerte es $2.286 \cdot 10.928 = 24.98$ veces mayor que la de un individuo que no tiene presente los factores de riesgo (blanco que mata a un negro).

Por supuesto, estas conclusiones podrían quedar completamente invalidadas si hubiera otras variables importantes que debieran haber sido tenidas en cuenta. Sobre todo si estas variables fueran confusoras con las ya incluídas. En el modelo logístico de regresión se supone que el investigador ha comprendido bien las variables que deben ser consideradas. Pruebas como la de bondad de ajuste de Hosmer-Lemeshov pueden ayudarle a evaluar el modelo establecido