Grau d'Estadística: Estadística Mèdica
Problemas del Bloque 4

Fecha de entrega: viernes 21 de diciembre 2018

Alumna: Laura Julià Melis

EJERCICIO 1

Dad dos ejemplos ficticios (y distintos a los de los apuntes) para el sesgo de selección en estudios epidemiológicos. En estos casos, ¿es de esperar una sobreestimación o una infraestimación de la medida de asociación de interés?

El cálculo del odds ratio, teniendo en cuenta las probabilidades de selección, es el siguiente:

$$OR = W \cdot \frac{A \cdot D}{B \cdot C} = \frac{w_A \cdot w_D}{w_B \cdot w_C} \cdot \frac{A \cdot D}{B \cdot C}$$

Donde $w_A, \dots, w_D$ son las probabilidades de selección y $A, \dots, D$, el número de individuos en cada grupo de exposición y enfermedad. Y por ello, si $W = 1$, la estimación del odds ratio no tendrá sesgo de selección.

Ejemplo 1.

Se considera un estudio de casos y controles para valorar la asociación entre la exposición al sol y la aparición de un melanoma. Los resultados obtenidos en el estudio y las probabilidades de selección son los siguientes:

	D	$\bar{D}$
E	100	200
$\bar{E}$	100	400

	D	$\bar{D}$
E	90%	90%
$\bar{E}$	90%	70%

Así pues, el odds ratio en la población, sin tener en cuenta la probabilidad de selección de cada grupo, es:

$$OR = \frac{100 \cdot 400}{100 \cdot 200} = 2$$

Este valor se interpreta diciendo que la probabilidad de tener un melanoma es 2 veces mayor entre las personas expuestas al sol que entre las no expuestas.

Pero, si tenemos en cuenta las probabilidades de selección, obtendremos unos valores diferentes:

	D	$\bar{D}$
E	100·0.9=90	200·0.9=180
$\bar{E}$	100·0.9=90	400·0.7=280

Se puede observar cómo el grupo control ($\bar{D}$) tiene una fracción de personas expuestas (180/280=0.64) más alta que en el estudio base, en el que no se utilizan las probabilidades (200/400=0.5).

Ahora, el odds ratio es:

$$OR = \frac{90 \cdot 280}{90 \cdot 180} = \frac{0.9 \cdot 0.7}{0.9 \cdot 0.9} \cdot \frac{100 \cdot 400}{100 \cdot 200} = 0.77 \cdot 2 = 1.6$$

Como $W = 0.77 \neq 1$, se confirma que hay sesgo de selección y, más concretamente, al ser $W < 1$, se puede afirmar que el odds ratio infraestima. También lo podemos ver comparando el primer OR obtenido y el segundo: en el segundo caso el OR ha sido inferior que cuando no se han considerado las probabilidades de selección.

Ejemplo 2.

En este caso se considera un estudio de cohorte retrospectivo realizado en una empresa en la que algunos empleados trabajan manipulando cierto componente químico y se quiere observar si este hecho conduce o no a una enfermedad concreta. Para ello, se utilizaron los registros de salud de los trabajadores, los datos se muestran en la siguiente tabla de contingencia:

	D	$\overline{D}$	Total
E	100	900	1000
$\overline{E}$	50	950	1000

Se calcula el riesgo relativo de padecer esa enfermedad entre los empleados expuestos y los no expuestos:

$$RR = \frac{100/1000}{50/1000} = 2$$

Ahora se supone que, debido al paso del tiempo, algunos de los registros de salud se perdieron o se descartaron debido a que el trabajador estaba sano, y los datos recogidos son los siguientes:

	D	$\overline{D}$	Total
E	90	720	819
$\overline{E}$	40	760	800

El valor del riesgo relativo en este nuevo contexto seria:

$$RR = \frac{99/819}{40/800} = 2.42$$

Como consecuencia, si se hubiera dado esta situación, se hubiera producido un sesgo de selección y una sobreestimación de la asociación entre el componente químico y la aparición de una determinada enfermedad.

En el primer caso, se concluye que el riesgo de padecer la enfermedad es 2 veces mayor entre los trabajadores expuestos al químico. Pero en el segundo, al haber descartado información de trabajadores sanos, se ha obtenido que el riesgo es 2.42 (0.42 más de lo que debería ser).

EJERCICIO 2

La Tabla 1 muestra datos ficticios de una población sobre la relación entre una enfermedad (D) y una exposición E que tiene tres niveles: ninguna exposición, nivel medio y nivel alto.

Tabla 1: Datos ficticios del Ejercicio 2

Nivel de	Enfermedad		
Exposición	Sí	No	Total
Ninguna	100	9900	10000
Media	1200	58800	60000
Alta	1200	28800	30000
Total	2500	97500	100000

¿Cuál es el riesgo atribuible a la exposición en la población?

El riesgo atribuible en la población es la proporción de casos de enfermedad en la población que son atribuibles a la exposición. Y se calcula como:

$$PAR = \frac{P(D) - P(D|\bar{E})}{P(D)} = P(E|D) \cdot \left(1 - \frac{1}{RR}\right).$$

Donde RR es el riesgo relativo y se calcula como:

$$RR = \frac{P(D|E)}{P(D|\bar{E})}$$

Al existir 3 grupos de exposición, se calcularán 3 riesgos atribuibles:

1. Ninguna-Media.

$$\mathrm{RR}_1 = \frac{1200/60000}{100/10000} = \frac{0.02}{0.01} = 2 \quad \rightarrow \quad PAR_1 = \frac{1200}{2500}\left(1 - \frac{1}{2}\right) = 0.24$$

2. Ninguna-Alta.

$$\mathrm{RR}_2 = \frac{1200/30000}{100/0000} = \frac{0.04}{0.01} = 4 \quad \rightarrow \quad PAR_2 = \frac{1200}{2500}\left(1 - \frac{1}{4}\right) = 0.36$$

3. Media-Alta.

$$\mathrm{RR}_3 = \frac{1200/30000}{1200/60000} = \frac{0.04}{0.02} = 2 \quad \rightarrow \quad PAR_3 = \frac{1200}{2500}\left(1 - \frac{1}{2}\right) = 0.24$$

Si se considera el grupo sin exposición y el de exposición media (primer caso) así como el de exposición media y exposición alta (tercer caso) se concluye que el 24% de los casos de enfermedad se pueden atribuir a la exposición, mientras que al considerar el grupo sin exposición y el de exposición alta, se observa cómo el son atribuibles a la exposición el 36% de los casos.

Si sumamos los dos tipos de exposición obtenemos:

$$\mathrm{RR}_4 = \frac{2400/90000}{100/10000} = \frac{0.026}{0.01} = 2.6 \quad \rightarrow \quad PAR_4 = \frac{2400}{2500}\left(1 - \frac{1}{2.6}\right) = 0.59$$

Por lo que el 59% de los casos de enfermedad podrían ser evitados si se eliminara cualquier tipo de exposición (media o alta).

EJERCICIO 3

Sean X_1 y X_2 dos variables categóricas con 2 y 3 categorías, respectivamente, y X_3 una variable numérica.

a) Plantead el modelo de regresión logística (para una enfermedad D) que incluye las tres variables y además la interacción entre X_1 y X_2.

Un modelo de regresión logística tiene la siguiente expresión:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \cdots + \beta_m X_m.$$

Donde $p \in [0,1]$ (siendo p=1 presencia de enfermedad y p=0, ausencia), α es el intercept, X_i la variable explicativa que se considera relevante para explicar el comportamiento de la variable respuesta, β_i el parámetro asociado a la variable X_i, y m el número de variables total que hay en el modelo.

Cuando se dispone de variables categóricas, estas deben codificarse con variables *dummy* (binarias, con valores 0 o 1) para poder incluirlas en el modelo de regresión. Si una variable tiene s niveles, se deberán crear s-1 variables *dummys* ya que la primera categoría es la categoría base o de referencia, y se recoge en el intercept. Por este motivo, en este caso se deberán crear dos *dummys* para la variable X_2.

$$X_{21} = \begin{cases} 1 & si\ X_2 = 2 \\ 0 & en\ otro\ caso \end{cases}, \qquad X_{22} = \begin{cases} 1 & si\ X_2 = 3 \\ 0 & en\ otro\ caso \end{cases}$$

Ahora ya se puede escribir el modelo de regresión logística siguiendo la fórmula indicada anteriormente y considerando las interacciones entre los factores y la covariable:

$$\boxed{logit(p) = \alpha + \beta_1 X_1 + \beta_2 X_{21} + \beta_3 X_{22} + \beta_4 X_3 + \beta_5 X_1 X_{21} + \beta_6 X_1 X_{22} + \beta_7 X_1 X_3 + \beta_8 X_3 X_{21} + \beta_9 X_3 X_{22}}$$

b) ¿Cuál es el odds ratio asociado a la comparación de $X_2 = 2$ vs. $X_2 = 3$?

$$OR = \frac{odds(p = 1|X_2 = 2, X_1, X_3)}{odds(p = 1|X_2 = 3, X_1, X_3)} = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_{21} + \beta_4 X_3 + \beta_5 X_1 X_{21} + \beta_7 X_1 X_3 + \beta_8 X_3 X_{21})}{\exp(\alpha + \beta_1 X_1 + \beta_3 X_{22} + \beta_4 X_3 + \beta_6 X_1 X_{22} + \beta_7 X_1 X_3 + \beta_9 X_3 X_{22})} =$$

$$= \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 + \beta_4 X_3 + \beta_5 X_1 + \beta_7 X_1 X_3 + \beta_8 X_3)}{\exp(\alpha + \beta_1 X_1 + \beta_3 + \beta_4 X_3 + \beta_6 X_1 + \beta_7 X_1 X_3 + \beta_9 X_3)} =$$

$$= \begin{cases} \exp(\beta_2 - \beta_3) & si\ X_1 = 0, X_3 = 0 \\ \exp(\beta_2 + \beta_5 - \beta_3 - \beta_6) & si\ X_1 = 1, X_3 = 0 \\ \exp(\beta_2 + \beta_8 - \beta_3 - \beta_9) & si\ X_1 = 0, X_3 = 1 \\ \exp(\beta_2 + \beta_5 + \beta_8 - \beta_3 - \beta_6 - \beta_9) & si\ X_1 = 1, X_3 = 1 \end{cases}$$

EJERCICIO 4

En un estudio epidemiológico transversal con 398 pacientes se quería estudiar si la presencia de una cierta enfermedad de interés, D, difiere entre hombres y mujeres y si está asociada a la infección por el VIH. Para ello se ajustaron dos modelos de regresión logística cuyos parámetros estimados se muestran en las Tablas 2 y 3:

Tabla 2: Modelo sin interacción

	$\hat{\beta}$	$s.e.(\hat{\beta})$	Z	p
Constante	−0,113	0,158		0,474
Mujer	−0,292	0,262		0,265
VIH+	0,351	0,205		0,088

Tabla 3: Modelo incluyendo interacción

	$\hat{\beta}$	$s.e.(\hat{\beta})$	Z	p
Constante	0,069	0,166		0,678
Mujer	−1,638	0,519		0,002
VIH+	0,013	0,226		0,954
Mujer*VIH+	2,115	0,625		0,001

Nota: El valor estimado de la covarianza de los parámetros de la variable infección por el VIH y el término de la interacción en el modelo 2 es 0,028.

a) Plantead formalmente ambos modelos.

Codificación de las dos variables:

$$X_{género} = \begin{cases} 1 & si\ género = Mujer \\ 0 & en\ otro\ caso\ (Hombre) \end{cases}, \qquad X_{VIH} = \begin{cases} 1 & si\ VIH = positivo(+) \\ 0 & en\ otro\ caso\ (VIH = -) \end{cases}$$

Modelo sin interacción:

$$\boxed{logit(p) = \alpha + \beta_1 X_{género} + \beta_2 X_{VIH}}$$

Modelo con interacción:

$$\boxed{logit(p) = \alpha + \beta_1 X_{género} + \beta_2 X_{VIH} + \beta_3 X_{género} X_{VIH}}$$

Donde α es la constante y $p \in [0, 1]$.

b) Rellenad la cuarta columna de ambas tablas, que contiene el estadístico de la prueba de Wald. ¿Para qué hipótesis se utiliza esta prueba?

La prueba de Wald se utiliza para conocer si una variable explicativa en concreto está asociada o no con la variable respuesta. Las hipótesis nula y alternativa son:

$$H_0: \beta_k = 0 \text{ vs } H_1: \beta_k \neq 0$$

Así pues, si el valor del parámetro asociado a la variable predictora k es igual a 0, significa que esta desaparece del modelo porque no es relevante para explicar el comportamiento de la variable respuesta.

El estadístico Z se calcula como:

$$Z = \frac{\hat{\beta}_k}{\sqrt{\widehat{Var}(\hat{\beta}_k)}}$$

Por ejemplo:

$$Z_{Constante} = \frac{\hat{\beta}_{Constante}}{\sqrt{\widehat{Var}(\hat{\beta}_{Constante})}} = -\frac{0.113}{\sqrt{0.158^2}} = -0.7152$$

	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{s.e.}(\widehat{\boldsymbol{\beta}})$	$\boldsymbol{Z}$	$\boldsymbol{p}$
Constante	-0.113	0.158	-0.715	0.474
Mujer	-0.292	0.262	-1.115	0.265
VIH+	0.351	0.205	1.712	0.088

	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{s.e.}(\widehat{\boldsymbol{\beta}})$	$\boldsymbol{Z}$	$\boldsymbol{p}$
Constante	0.069	0.166	0.416	0.678
Mujer	-1.638	0.519	-3.156	0.002
VIH+	0.013	0.226	0.058	0.954
Mujer*VIH+	2.115	0.625	3.384	0.001

c) Estimad los odds ratios asociados a las variables sexo e infección por el VIH en el modelo sin interacción e interpretad los valores.

Variable $X_{género}$:

$$\widehat{OR}_1 = \frac{odds\left(p=1\middle|X_{género}=1, X_{VIH}\right)}{odds\left(p=1\middle|X_{género}=0, X_{VIH}\right)} = \frac{\exp\left(\alpha+\hat{\beta}_1 X_{género}+\hat{\beta}_2 X_{VIH}\right)}{\exp\left(\alpha+\hat{\beta}_2 X_{VIH}\right)} = \exp\left(\hat{\beta}_1\right) = \exp(-0.292)$$

$$\widehat{OR}_1 = \boxed{0.74677}$$

Variable X_{VIH}:

$$\widehat{OR}_2 = \frac{odds\left(p=1\middle|X_{VIH}=1, X_{género}\right)}{odds\left(p=1\middle|X_{VIH}=0, X_{género}\right)} = \frac{\exp\left(\alpha+\hat{\beta}_1 X_{género}+\hat{\beta}_2 X_{VIH}\right)}{\exp\left(\alpha+\hat{\beta}_1 X_{género}\right)} = \exp\left(\hat{\beta}_2\right) = \exp(0.351)$$

$$\widehat{OR}_2 = \boxed{1.4205}$$

d) ¿Cuál es el odds ratio asociado a la infección por el VIH según el modelo con interacción? Calculad también el intervalo de confianza (del 95 %) correspondiente.

$$\widehat{OR} = \frac{odds\left(p=1\middle|X_{VIH}=1, X_{género}\right)}{odds\left(p=1\middle|X_{VIH}=0, X_{género}\right)} = \frac{\exp\left(\alpha+\hat{\beta}_1 X_{género}+\hat{\beta}_2 X_{VIH}+\hat{\beta}_3 X_{género} X_{VIH}\right)}{\exp\left(\alpha+\hat{\beta}_1 X_{género}\right)}$$

$$= \begin{cases} \exp\left(\hat{\beta}_2\right) = \exp(0.013) = \boxed{1.0131} & si\ X_{género}=0 \\ \exp\left(\hat{\beta}_2+\hat{\beta}_3\right) = \exp(2.128) = \boxed{8.3981} & si\ X_{género}=1 \end{cases}$$

Intervalo de confianza:

La fórmula general para un modelo con interacciones es la siguiente:

$$IC_{(OR_k;1-\alpha)} = \exp\left(\hat{\beta}_k \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\widehat{Var}(\hat{\beta}_k)}\right)$$

Como se pide un nivel de confianza del 95%, (1 − α = 0.95), sabemos que α = 0.05, por lo que mirando en la tabla de la distribución normal obtenemos que $z_{1-\alpha} = z_{0.975} = 1{,}96$. También tenemos el resto de los componentes necesario, así que substituimos:

Si $X_{género} = 0$:

$$IC_{(0.95)} = \exp\left(\hat{\beta}_2 \pm z_{0.975} \cdot \sqrt{\widehat{Var}(\hat{\beta}_2)}\right) = \exp(0.013 \pm 1.96 \cdot 0.226 = \boxed{[0.65, 1.58]}$$

Si $X_{género} = 1$:

$$IC_{(0.95)} = \exp\left((\hat{\beta}_2+\hat{\beta}_3) \pm z_{0.975} \cdot \sqrt{\widehat{Var}(\hat{\beta}_2)+\widehat{Var}(\hat{\beta}_3)+2\cdot\widehat{Cov}(\hat{\beta}_2,\hat{\beta}_3)}\right) =$$

$$= \exp(2.128 \pm 1.96 \cdot 0.7055) = \boxed{[2.11, 33.47]}$$