**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:

Donde son las probabilidades de selección y , el número de individuos en cada grupo de exposición y enfermedad. Y por ello, si , 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:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | 100 | 200 |  | 90% | 90% |
|  | 100 | 400 |  | 90% | 70% |

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

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:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 100·0.9=90 | 200·0.9=180 |
|  | 100·0.9=90 | 400·0.7=280 |

Se puede observar cómo el grupo control () 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:

Como , se confirma que hay sesgo de selección y, más concretamente, al ser , 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:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **Total** |
|  | 100 | 900 | 1000 |
|  | 50 | 950 | 1000 |

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

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:

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | **Total** |
|  | 90 | 720 | 819 |
|  | 40 | 760 | 800 |

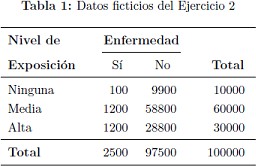
El valor del riesgo relativo en este nuevo contexto seria:

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.**

****

**¿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:

Donde RR es el riesgo relativo y se calcula como:

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

1. Ninguna-Media.

→

1. Ninguna-Alta.

→

1. Media-Alta.

→

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:

→

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 X1 y X2 dos variables categóricas con 2 y 3 categorías, respectivamente, y X3 una variable numérica.**

1. **Plantead el modelo de regresión logística (para una enfermedad D) que incluye las tres variables y además la interacción entre X1 y X2.**

Un modelo de regresión logística tiene la siguiente expresión:

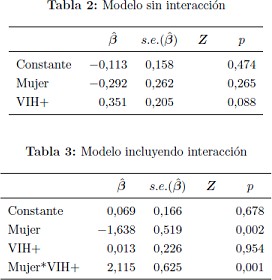
Donde (siendo p=1 presencia de enfermedad y p=0, ausencia), es el intercept, la variable explicativa que se considera relevante para explicar el comportamiento de la variable respuesta, el parámetro asociado a la variable , y m el número de variables total que hay en el modelo.

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 X2.

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:

1. **¿Cuál es el odds ratio asociado a la comparación de X2 = 2 vs. X2 = 3?**

# 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:**

**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.**

1. **Plantead formalmente ambos modelos.**

Codificación de las dos variables:

Modelo sin interacción:

Modelo con interacción:

Donde es la constante y .

1. **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:

vs

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:

Por ejemplo:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **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 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| **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 |

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

Variable :

Variable :

1. **¿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.**

Intervalo de confianza:

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

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 . También tenemos el resto de los componentes necesario, así que substituimos:

Si :

Si :