TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa UPC-UB

Título: Estimación del riesgo relativo mediante el modelo log-binomial

Autor: Silvia Pérez Fernández

Director: Klaus Langohr

Departamento: Departamento de Estadística e Investigación Operativa

Universidad: Universitat Politècnica de Catalunya

Convocatoria: 2015/2016

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques i Estadística
Màster en Estadística i Investigació Operativa

Estimación del riesgo relativo mediante el modelo log-binomial

Autor: : Silvia Pérez Fernández
Director : Klaus Langohr

Junio, 2016
Barcelona

Resumen

Una de las herramientas más usadas para estimar la magnitud de asociación entre una exposición y una enfermedad es la regresión logística, que utiliza el *odds ratio* (OR) como medida de asociación. Sin embargo, cuando la variable de interés es común en la población de estudio tiende a exagerar el riesgo. El modelo log-binomial, con error binomial y función de enlace logaritmo, es una buena alternativa. Este modelo utiliza la razón de prevalencia (PR) o el riesgo relativo (RR) como medida de asociación, dependiendo del tipo de estudio. Estas medidas son mucho más interpretables que el OR. Sin embargo, no existe demasiada información sobre este modelo. Por ello, en este trabajo hemos ampliado la información existente, definiendo su expresión y explicando los problemas de convergencia. Hemos presentado el método COPY desarrollado por Deddens et al. (2003) en SAS para resolver esos problemas y lo hemos implementado en el software R. Por último, hemos comparado este método con las funciones ya existentes en R mediante bases de datos reales comprobando cuál resulta más efectiva en cada caso.

Palabras clave: log-binomial, convergencia, riesgo relativo, razón de prevalencia, estimador de máxima verosimilitud, método COPY

Abstract

In order to estimate the magnitude of the association between an exposure and a disease, we usually apply the logistic regression model which uses the odds ratio (OR) as the measure of the disease-exposure association. However, when the outcome is common in the study population it tends to exaggerate the risk. The log-binomial model, which assumes binomial errors and uses the logarithm as link function, is a good alternative. This model uses the prevalence ratio (PR) or relative risk (RR) as a measure of association, depending on the type of study. However, there is not much information about this model. Therefore, in this work we have expanded the existing information, defining its expression and explaining convergence problems. We have presented COPY method developed by Deddens et al. (2003) in SAS to solve these problems and we have implemented it in the statistical software package R. Finally, we have compared this method with existing functions in R with real data bases checking which is more effective in each case.

Keywords: log-binomial, convergence, relative risk, prevalence ratio, maximum likelihood estimator, COPY method

Índice general

Índice de tablas

Capítulo 1

Introducción

La epidemiología es aquella parte de la medicina que se dedica a estudiar el desarrollo epidémico y la incidencia de las enfermedades infecciosas en la población. Su significado deriva del griego Epi (sobre), Demos (Pueblo) y Logos (ciencia). La literatura científica reconoce al inglés John Snow como padre de la epidemiología, ya que aportó importantes avances en el conocimiento de la epidemia de cólera que acechaba Londres en aquella época [21].

Su objetivo principal es poder estimar la magnitud o el grado de asociación entre una exposición y una enfermedad. Existen distintas formas de cuantificar esa asociación. Una de las herramientas más utilizadas es el análisis multivariante. Entre los principales métodos de análisis multivariante se encuentra el modelo de regresión logística, que ayuda a describir de forma sencilla como influye la presencia o no de diversos factores en la probabilidad de aparición de un suceso. Debido a su uso en la epidemiología, no es difícil poder implementarlos en cualquiera de los software estadísticos hoy en día disponibles.

Entre las principales carácterísticas del modelo se encuentra que los términos de error siguen una distribución binomial, que utiliza la función de enlace logit y que a la hora de interpretar los parámetros del modelo utiliza el *odds ratio* (OR) como medida de asociación. Pero el modelo de regresión logística presenta una limitación o desventaja importante. Cuando la variable de interés es común en la población, el *odds ratio* (OR) puede verse aumentado, exagerando la asociación de riesgo.

Es por ello que nuestro objetivo en este trabajo es presentar un método alternativo para estimar la asociación entre la enfermedad y la exposición; el modelo de regresión log-binomial. Se trata de un modelo en el que, al igual que en el de la regresión logística, los términos de error siguen una distribución binomial, pero en este caso utiliza el logaritmo como función de enlace y el riesgo relativo (RR) como medida de asociación. La ventaja de utilizar este modelo es que no tiende a exagerar el riesgo y que su interpretación es más sencilla de entender mediante el riesgo relativo (RR).

El principal problema del modelo log-binomial es que apenas se nombra en la literatura, por lo que no se tiene demasiado conocimiento sobre él. Es por ello, que no todos los software estadísticos tienen implementada una función para realizar el ajuste. A este problema se le debe añadir que en ocasiones, cuando se intenta ajustar el modelo en cualquiera de los software disponibles, el proceso de maximización falla al no encontrar el estimador de máxima verosimilitud (MLE), generando así problemas de convergencia.

Nuestro propósito es ampliar la información existente sobre el modelo de regresión log-binomial, definiendo su expresión e interpretación de parámetros, su estimación, la bondad de ajustes del modelo y explicando detalladamente los problemas de convergencia. También presentaremos un nuevo método desarrollado por Deddens et. al (2003) [16] en el software SAS denominado COPY. Nuestra principal tarea ha sido implementar este método en R, comparando los resultados obtenidos y viendo si realmente soluciona los problemas de estimación, mejorando así las funciones ya existentes.

Este trabajo se distribuye de la siguiente manera. Comenzaremos hablando de la terminología y definiendo las distintas medidas epidemiológicas. Explicaremos en qué consisten el modelo de regresión logística y de Poisson y cómo se definen. A continuación, hablaremos del modelo de regresión log-binomial y la implementación del método COPY en R. Y finalmente compararemos el modelo de regresión logística con el log-binomial mediante ejemplos de datos reales, usando las distintas funciones ya existentes de R y el método anteriormente desarrollado COPY. Por último, presentaremos una serie de recomendaciones para futuras investigaciones sobre el modelo log-binomial.

Capítulo 2

Medidas epidemiológicas

En este capítulo, comenzaremos explicando los principales conceptos de la epidemiología. El objetivo principal de la mayoría de epidemiólogos es obtener una estimación válida y precisa del efecto de una causa potencial en la enfermedad de ocurrencia, la cual es usualmente una respuesta binaria. En el anexo A se describen con brevedad los tipos de estudios epidemiológicos. Buena parte de la notación usada a continuación se basa en la empleada en Jewell (2004) [1].

2.1 Prevalencia e incidencia

Tanto la prevalencia como la incidencia representan proporciones de caso de enfermedad en una población determinada que se enferma en un momento específico. Para calcularlas nos interesa cuantificar la exposición de un individuo ante una variedad de factores y su nivel de enfermedad.

2.1.1 Prevalencia

La prevalencia de una enfermedad D es la proporción de individuos afectados por la enfermedad entre la población de interés en un tiempo dado t:

$$P = \frac{X}{N}$$

donde X es el número de casos de enfermedad y N es el tamaño poblacional en el tiempo t. Se trata de una proporción, por tanto, su valor oscila entre el 0 y el 1.

Podemos distinguir dos tipos de prevalencia, la puntual y la periódica. Por un lado, la **prevalencia puntual** se define como la proporción de una población definida en riesgo de enfermedad que se ve afectada por ella en un punto especifico en el tiempo. Por otro lado, hablamos de **prevalencia periódica** cuando nos referimos a la proporción de la población en riesgo afectados en algún punto del intervalo de tiempo.

Dada una muestra de tamaño n, definimos la **estimación de la prevalencia** mediante la siguiente fórmula:

$$\hat{P} = \frac{X_n}{n}$$

donde X_n es el número de casos entre la muestra.

Dada una muestra de observaciones independientes y asumiendo que X_n sigue una distribución binomial con parámetros n y P, el intervalo de confianza puede ser calculado utilizando esta distribución [3]. Para calcularlo podemos usar también la siguiente aproximación mediante la distribución normal:

$$CI(P, 1-\alpha) = \hat{P} \pm z_{1-\alpha/2} \cdot \sqrt{\hat{P}(1-\hat{P})/n}$$

donde $1-\alpha$ es el nivel de confianza y $z_{1-\alpha/2}$ el cuantil de la distribución normal estándar.

Es importante tener en cuenta que la prevalencia puede ser estimada en los estudios transversales pero no en los de cohorte o caso control.

2.1.2 Incidencia acumulada

Definimos la incidencia acumulada de una enfermedad como la proporción de nuevos casos en un periodo de tiempo t entre la población libre de la enfermedad inicialmente:

$$I(t) = \frac{I}{N_0}$$

donde I representa el número de nuevos (casos incidentes) casos durante un periodo de tiempo y N_0 el número de individuos libres de la enfermedad inicialmente.
La incidencia acumulada puede ser estimada en estudios de cohorte pero no en los estudios transversales y caso-control. Su estimador es la proporción de nuevos casos en la cohorte y su intervalo de confianza se puede calcular usando una distribución binomial o su aproximación mediante la distribución normal.

2.1.3 Tasa de incidencia

La tasa de incidencia de una enfermedad D se define como el número de nuevos casos por unidad de persona-tiempo en riesgo. No se trata de una proporción, por tanto no puede ser interpretada como una probabilidad y su unidad es $(tiempo)^{-1}$.

Definimos su estimador mediante la fórmula:

$$\hat{I_r} = \frac{I}{\Delta T}$$

Se puede suponer que el número de casos incidentes durante todo el tiempo total en riesgo, ΔT, sigue una distribución de Poisson con parámetro $I_{r_0} \cdot \Delta T$. Asumiendo esto, se pueden calcular los intervalos de confianza basados en ésta distribución o usar la aproximación mediante la distribución normal. Por tanto, el valor esperado y la varianza de $\hat{I}_r$ son I_{r_0} y $I_{r_0}/\Delta T$. Esto nos conduce a la siguiente aproximación del intervalo de confianza:

$$CI(I_{r_0}, 1-\alpha) = \hat{I}_r \pm z_{1-\alpha/2} \cdot \sqrt{\hat{I}_r/\Delta T}$$

Al igual que en el caso anterior, debemos tener en cuenta que la tasa de incidencia puede ser estimada en estudios de cohorte pero no en estudios transversales ni caso-control.

2.2 Medidas de asociación enfermedad-exposición

En este apartado presentaremos medidas para estudiar la asociación entre un factor de riesgo, o exposición, y la ocurrencia de una enfermedad. Nos referiremos a la enfermedad mediante D y al factor de riesgo mediante E. Por tanto, tendremos cuatro clases de individuos, los que tienen tanto E como D, los que solamente tienen un factor y los individuos que no tienen ninguno. Compararemos las probabilidades de $P(D|E)$ y $P(D|\bar{E})$. Las siguientes secciones nos demuestran distintas alternativas para cuantificar esta comparación.

2.2.1 Riesgo relativo

El riesgo relativo (RR) explica la relación que existe entre el riesgo de enfermedad (D) en la población expuesta (E) y el riesgo en la población no expuesta ($\bar{E}$) y se define mediante la siguiente fórmula:

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

Como podemos observar en la definición el RR no puede ser negativo. Si $RR > 1$, hay una probabilidad superior de enfermedad entre los expuestos, por tanto, la exposición es un posible factor de riesgo para la enfermedad. Si al contrario, $RR < 1$ la exposición es un factor protector para la enfermedad. Y si $RR = 1$, no hay diferencia entre los dos grupos respecto al riesgo de enfermedad, es decir, E y D son independientes.

Cabe destacar que el riesgo relativo solo se puede estimar en los estudios de cohorte, mientras que para los estudios transversales deberemos estimar el riesgo relativo de prevalencia (PRR o PR). En los estudios caso-control no es posible estimar $P(D|E)$ o $P(D|\bar{E})$, por tanto tampoco el riesgo relativo.

Para poder entenderlo mejor presentaremos un ejemplo simple. Para ello utilizaremos la base de datos del paquete de R *"epitools"* [25] que contiene el data frame *wcgs* con datos de un estudio de cohorte. La finalidad de este estudio es analizar la relación entre enfermedades cardiovasculares y una serie de posibles factores de riesgo como son el tipo de patrón de

comportamiento, A o B. Es importante destacar que en el ejemplo no se especifica a que se refieren cada uno de estos tipos de comportamiento. A continuación mostramos la tabla con los datos de exposición, enfermedad y tiempo total bajo riesgo:

	Enfermedad			
Comportamiento	Sí	No	**Total**	ΔT
Tipo A	a=178	b=1411	1589	11370,1
Tipo B	c=79	d=1486	1565	11805,6

Tabla 2.1: Enfermedades cardiovasculares y tiempo de seguimiento según tipo de comportamiento

Tratando los datos como poblacionales calculamos el riesgo relativo mediante la fórmula anteriormente expuesta y obtenemos

$$RR = \frac{P(D|E)}{P(D|\bar{E})} = \frac{178/1589}{79/1565} = 2,22$$

Por tanto, la probabilidad de enfermar es 2,22 veces mayor en caso del tipo A.

2.2.2 Diferencia de riesgos

La diferencia de riesgos es una medida absoluta que se define como la diferencia entre el riesgo de enfermedad entre los individuos expuestos y no expuestos.

$$RD = P(D|E) - P(D|\bar{E})$$

Tiene un rango de -1 a 1. $RD = 0$ si no hay asociación entre la presencia del factor y el evento; $RD > 0$ si la asociación es positiva, por lo que la presencia del factor se asocia a mayor ocurrencia del evento y $RD < 0$ si la asociación es negativa. Esta medida se puede calcular en los estudios transversales y cohorte, incluso cuando no hay ningún evento en alguno de los grupos. Sin embargo, como sucede con el RR, no es posible estimarlo en estudios caso-control.

2.2.3 *Odds ratio*

El *odds ratio* (OR) se define como la relación entre el *odds* de enfermedad en los casos expuestos en comparación con el *odds* en los no expuestos. Por tanto, es una medida de tamaño del efecto y la definimos:

$$OR = \frac{odds(D|E)}{odds(D|\bar{E})} = \frac{P(D|E)/(1 - P(D|E))}{P(D|\bar{E})/(1 - P(D|\bar{E}))} \tag{2.1}$$

Al igual que sucedía con el riesgo relativo, si $OR > 1$, hay mayor riesgo de enfermedad entre los expuestos, si $OR < 1$ hay menor riesgo de enfermedad entre los expuestos y si $OR = 1$ E y D son independientes.

A diferencia del riesgo relativo, el *odds ratio* sí que se puede calcular en estudios caso-control, ya que se ha demostrado que la siguiente fórmula es equivalente al OR definido en la ecuación (2.1):

$$OR = \frac{P(E|D)/(1 - P(E|D))}{P(E|\bar{D})/(1 - P(E|\bar{D}))}$$

Los estudios casos-control a menudo se llevan a cabo cuando la enfermedad de interés es rara o poco común; el RR puede ser por tanto aproximado mediante el OR. Si la enfermedad de interés no es muy común, el *odds ratio* es interpretado en términos de *odds* y no de riesgo. Veamos cómo calcularlo utilizando la tabla del ejemplo anterior,

$$OR = \frac{P(D|E)/(1 - P(D|E))}{P(D|\bar{E})/(1 - P(D|\bar{E}))} = \frac{178 \cdot 1411}{79 \cdot 1486} = 2,37$$

Es decir, el *odds* de enfermar es 2,37 veces mayor en caso del tipo A.

Vemos que los resultados obtenidos al calcular el riesgo relativo y el *odds ratio* son muy similares, aunque el *odds ratio* es un poco más elevado. Si comparamos el RR con el OR, vemos que este último suele tomar un valor un poco más elevado.

Si usamos la definición del *odds ratio*:

$$\begin{aligned} OR &= \frac{P(D|E)}{P(D|\bar{E})} \cdot \frac{P(\bar{D}|\bar{E})}{P(\bar{D}|E)} \\ &= RR \cdot \frac{P(\bar{D}|\bar{E})}{P(\bar{D}|E)} \\ &= RR \cdot \frac{1 - P(D|\bar{E})}{1 - P(D|E)} \end{aligned} \tag{2.2}$$

Si el $RR > 1$, entonces tenemos que

$$\begin{aligned} P(D|E) &> P(D|\bar{E}) \\ \Rightarrow 1 - P(D|E) &< 1 - P(D|\bar{E}) \\ \Rightarrow P(\bar{D}|E) &< P(\bar{D}|\bar{E}) \\ \Rightarrow OR &> RR \end{aligned}$$

Del mismo modo se puede comprobar que si $RR < 1$, entonces $OR < RR$. Por tanto, el *odds ratio* se encuentra siempre más lejos del valor 1 que el riesgo relativo, excepto cuando las dos medidas son iguales a 1. Lo veremos más claramente en el gráfico que mostramos a continuación sacado del paper de Zhang y Yu (1998) [11], en el que se muestra la relación entre el riesgo relativo y el *odds ratio* mediante la incidencia entre los no expuestos (Figura 2.1).

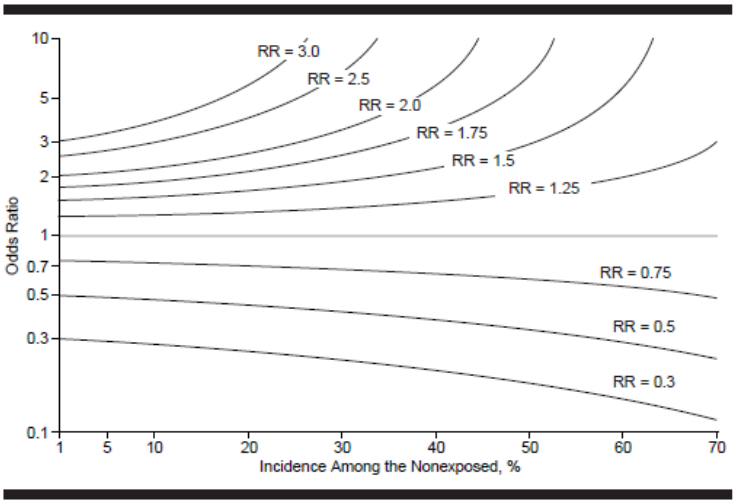

Figura 2.1: Relación entre RR y OR mediante la incidencia entre los no expuestos (Zhang y Yu, 1998)

Para verlo más claramente pondremos un ejemplo simple [2]. Si tenemos que $P(D|\bar{E}) = 0{,}05$ y $P(D|E) = 0{,}10$, entonces el $RR = 2$ y $OR = 2{,}11$, por lo que no existe mucha diferencia. En cambio, si $P(D|E) = 0{,}5$ el $RR = 10$ y $OR = 19$, es decir, el *odds ratio* es casi el doble. Más adelante explicaremos en profundidad la propuesta de Zhang y Yu (1998) [11] para estimar el RR en función del OR.

2.2.4 Razón de las tasas de incidencia

La razón de las tasas de incidencia (IRR) se define como el cociente de dos tasas de incidencia. Para calcularlo, dividimos la tasa de incidencia de la parte expuesta de la población por la tasa de incidencia de la parte no expuesta de la población, obteniendo así una medida relativa (IRR) de los efectos de una exposición dada. Si los sucesos son poco comunes se aproxima al riesgo relativo o al *odds ratio*.

Volviendo al ejemplo anterior, la razón de las tasas de incidencia es

$$IRR = \frac{178/1137,1}{79/11805,6} = 2,34$$

Es decir, la aparición de nuevos casos de enfermedades cardiovasculares es 2,34 veces mayor en el caso del tipo de comportamiento A.

A la hora de relacionar las 3 medidas de asociación explicadas anteriormente Symons et al. (2002) [8] comentan que matemáticamente, en estudios prospectivos, cuando el seguimiento es corto, las tasas del eventos son pequeñas o el riesgo relativo está cercano al 1, el RR, OR y IRR se aproximan entre sí. Por lo tanto, definiríamos la siguiente relación entre estas 3 medidas:

$$1 < RR < IRR < OR$$

2.2.5 Estimación univariada e intervalos de confianza

Odds ratio

Para estimar el *odds ratio*, solamente necesitaremos los términos anteriormente utilizados para su definición, es decir, $P(D|E)$ y $P(D|\bar{E})$. Si consideramos la siguiente tabla

	D	$\bar{D}$
E	a	b
$\bar{E}$	c	d

donde a es el número de individuos con la enfermedad expuestos, b el número de no enfermos expuestos, c el número de enfermos no expuestos y d el número de no enfermos y expuestos. Entonces, la estimación para $P(D|E)$ será simplemente la proporción observada de individuos expuestos que están enfermos, es decir, $a/(a+b)$. Del mismo modo calculamos la estimación de la probabilidad $P(D|\bar{E})$, y lo sustituimos en la fórmula del OR, obteniendo así el estimador de máxima verosimilitud (MLE):

$$\widehat{OR} = \left[\frac{a/(a+b)}{b/(a+b)}\right] / \left[\frac{c/(a+b)}{d/(a+b)}\right] = \frac{ad}{bc}$$

En el caso de los estudios caso-control, usariamos el *odds* de exposición entre los casos y los controles, y obtendríamos exactamente la misma estimación que para los estudios de cohorte.

Como la distribución muestral del $\widehat{OR}$ es asimétrica hacia la derecha para muestras pequeñas, una forma sencilla de tratar con este problema es utilizar una transformación. Dado que la distribución muestral de $\log(\widehat{OR})$ es más simétrica que la de $\widehat{OR}$ y se aproxima mejor mediante una distribución Normal cuando el tamaño de la muestra es grande se aprovecha esta distribución para calcular los intervalos de confianza. Si suponemos dos cohortes independientes (aunque sea lo mismo cogiendo otro tipo de estudio) y definimos $\hat{p}_1 = \hat{P}(D|E)$ y $\hat{p_0} = \hat{P}(D|\bar{E})$ y los sustituimos en la ecuación,

$$\log(\widehat{OR}) = \log \frac{\hat{p}_1}{1-\hat{p}_1} - \log \frac{\hat{p_0}}{1-\hat{p_0}}$$

Siguiendo el método Delta y la serie de Taylor de primer orden, cuando se tiene una función $f(x)$, se puede aproximar de la forma:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0)$$

En nuestro caso, cogeremos $f(x) = \log(x)$, $x = \hat{p}_1$ y $x_0 = p_1$

$$\log(\hat{p}_1) \approx \log(p_1) + \frac{\hat{p}_1 - p_1}{p_1}$$

$$\log \frac{\hat{p}_1}{1-\hat{p}_1} \approx \log \frac{p_1}{1-p_1} + (\hat{p}_1 - p_1) \cdot \frac{1}{p_1(1-p_1)}$$

Por tanto, como a sigue una distribución binomial con parámetros n_1 y p_1,

$$Var\left(\log \frac{\hat{p}_1}{1-\hat{p}_1}\right) \approx Var(\hat{p}_1) \cdot \frac{1}{(p_1(1-p_1))^2} = \frac{1}{n_1 p_1 (1-p_1)}$$

y estimamos la varianza mediante

$$\widehat{Var}\left(\log \frac{\hat{p}_1}{1-\hat{p}_1}\right) \approx \frac{1}{\hat{n}_1 \hat{p}_1 (1-\hat{p}_1)} = \frac{1}{a} + \frac{1}{b}$$

Realizando exactamente los mismo cálculos, obtendremos

$$\widehat{Var}\left(\log \frac{\hat{p_0}}{1-\hat{p_0}}\right) \approx \frac{1}{c} + \frac{1}{d}$$

Y como los expuestos y no expuestos son independientes,

$$\widehat{Var}(\log(\widehat{OR})) \approx \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

Así, los intervalos de confianza de log(OR) y OR son:

$$CI(\log(OR); 1-\alpha) = \log(\widehat{OR}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

$$CI(OR; 1-\alpha) = \widehat{OR} \cdot \exp\left(\pm z_{1-\alpha/2}\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}\right)$$

A veces sucede que la muestra con la que estamos tratando es demasiado pequeña y el estimador $\widehat{OR}$ es algo sesgado. Para ello, Jewell (1986) encontró la solución a este problema:

$$\widehat{OR} = \frac{ad}{(b+1)(c+1)}$$

Algo tan simple como añadir una unidad a los valores del divisor. Para calcular su varianza, simplemente sumamos 1/2 a cada valor del divisor, ya que el sesgo es menor que un punto al usar la transformación logarítmica. Así mismo, conseguiremos calcular también el intervalo de confianza.

$$\widehat{Var}(\log(\widehat{OR})) \approx \frac{1}{(a+1/2)} + \frac{1}{(b+1/2)} + \frac{1}{(c+1/2)} + \frac{1}{(d+1/2)}$$

$$CI(OR; 1-\alpha) = \widehat{OR} \cdot \exp\left(\pm z_{1-\alpha/2}\sqrt{\frac{1}{(a+1/2)} + \frac{1}{(b+1/2)} + \frac{1}{(c+1/2)} + \frac{1}{(d+1/2)}}\right)$$

Riesgo relativo

Al igual que en la sección anterior, utilizaremos los términos $P(D|E)$ y $P(D|\bar{E})$ para definir su estimador. Teniendo en cuenta que la definición es la misma que la expuesta para el *odds ratio*, conseguimos:

$$\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}$$

Es importante tener en cuenta que el riesgo relativo no es directamente estimable en los estudios de caso-control, aunque como ya dijimos en el apartado 2.2.2, si que puede ser aproximado mediante el *odds ratio* si la variable de interés es poco frecuente.

Del mismo modo que para el *odds ratio*, la distribución muestral de $\widehat{RR}$ es asímetrica hacia la derecha, particularmente cuando la muestra es muy pequeña. Es por ello que en este caso también utilizaremos la transformación logarítmica para reducirlo.
Siguiendo las propiedades del logaritmo, tenemos

$$\log \hat{p}_1 \approx \log p_1 + \frac{(\hat{p}_1 - p_1)}{p_1}$$

Teniendo en cuenta que el término $\log p_1$ es constante,

$$Var(\log \hat{p}_1) \approx \frac{Var(\hat{p}_1)}{p_1^2} = \frac{1-p_1}{n_1 p_1}$$

Sustituyendo $\hat{p}_1$ por p_1

$$\widehat{Var}(\log \hat{p}_1) \approx \frac{Var(\hat{p}_1)}{\hat{p}_1^2} = \frac{b}{a(a+b)}$$

Para p_0 seguimos exactamente el mismo procedimiento hasta obtener

$$\widehat{Var}(\log(\widehat{RR})) = \frac{b}{a(a+b)} + \frac{d}{c(c+d)}$$

Su intervalo de confianza es:

$$CI(\log(RR), 1-\alpha) = \log(\widehat{RR}) \pm z_{1-\alpha/2}\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}$$

$$CI(RR, 1-\alpha) = \widehat{RR} \cdot \exp\left(\pm z_{1-\alpha/2}\sqrt{\frac{b}{a(a+b)} + \frac{d}{c(c+d)}}\right)$$

Si nos encontramos ante una muestra pequeña, reajustaremos el riesgo relativo para así evitar la posible división por 0:

$$\widehat{RR} = \frac{a/(a+b)}{(c+1)/(c+d+1)}$$

2.2.6 Modelos de regresión para estimar el OR y el IRR

En este apartado definiremos tanto del modelo de regresión logística, que utiliza el *odds ratio* como medida de asociación, como el modelo de regresión de Poisson, el cuál utiliza la razón de tasas de incidencia como medida de asociación. Explicaremos brevemente cuál es su expresión y cómo podemos estimar los parámetros.

El modelo de Regresión Logística

El estimador de Mantel-Haenszel se puede usar para cuantificar la asociación entre la enfermedad (D) y la exposición (E) de interés en presencia de una variable confusora. Sin embargo, este estimador puede no ser adecuado si existe más de una posible variable confusora, si una de esas posibles variables confusoras es continua o si E es una variable continua. En estas situaciones, la regresión logística puede ser una herramienta adecuada para analizar el nivel de asociación entre la enfermedad y la exposición.

El modelo de regresión logística es probablemente uno de los modelos más comúnmente empleados en la epidemiología a la hora de analizar respuestas binarias. Se asume que los términos de error siguen una distribución binomial, y que usa la función logit como función de enlace. Estos modelos pueden ser aplicados a distintos diseños de estudios como son transversales, cohorte o caso-control, teniendo en cuenta varios factores que explicaremos más adelante.

Siendo Y la variable binaria de interés, diremos que toma el valor 1 si presenta la enfermedad (D) y el valor 0 si el individuo no la presenta ($\bar{D}$). Usaremos el modelo de regresión logística para modelar la probabilidad condicionada $P(Y = 1|\boldsymbol{X} = \boldsymbol{x})$ como una función de $\boldsymbol{x}$, estimando mediante la máxima verosimilitud. $\boldsymbol{X}$ es el vector de las variables $X_1 \dots X_m$.

Para definir este modelo usamos la siguiente expresión:

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m \tag{2.3}$$

Esta expresión es equivalente a

$$\begin{aligned} p &= P(Y = 1|\mathbf{X}) = P(Y = 1|X_1, \ldots, X_m) \\ &= \frac{\exp(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m)}{1 + \exp(\beta_0 + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m)} \end{aligned}$$

Esto implica que X_i es un factor de riesgo (o factor protector) para Y, si $\beta_i > 0$ (o $\beta_i < 0$).

Cuando tenemos variables categóricas que incluir en nuestro modelo usamos la codificación *dummy*. Esta codificación se compone por una serie de números asignados para indicar la pertenencia en distintos grupos. Por ejemplo, si uno de los regresores, X_k, es una variable categórica con s niveles, en el modelo incluiremos $s-1$ variables *dummy*:

$$X_{k_1} = \begin{cases} 1, & \text{si } X_k = 2 \\ 0, & \text{caso contrario} \end{cases} \quad \dots \quad X_{k_{s-1}} = \begin{cases} 1, & \text{si } X_k = s \\ 0, & \text{caso contrario} \end{cases}$$

Se puede coger cualquiera de los niveles de s como categoría de referencia. Sin embargo, si X_k es una variable ordinal, es preferible elegir $X_k = 1$ o $X_k = s$ como nivel de referencia siempre que el número de observaciones no sea demasiado pequeña, ya que de este modo facilitaremos la interpretación del modelo.

Los parámetros de los modelos de regresión logística pueden ser estimados usando el estimador de máxima verosimilitud. Sea $(y_i, \boldsymbol{x_i})$, $i = 1, \dots, n$, una muestra de observaciones independientes, podemos expresar la función de máxima verosimilitud de la siguiente manera:

$$\begin{aligned} L(\alpha, \boldsymbol{\beta}|Y, \mathbf{X}) &= \prod_{i=1}^{n} P(Y = y_i|x_i) f(\boldsymbol{x_i}) \propto \prod_{i=1}^{n} P(Y = y_i|\boldsymbol{x_i}) \\ &= \prod_{i=1}^{n} P(Y = 1|\boldsymbol{x_i})^{\delta_i} P(Y = 0|\boldsymbol{x_i})^{1-\delta_i} \\ &= \prod_{i=1}^{n} \frac{\exp(\alpha + \boldsymbol{\beta'}\boldsymbol{x_i})^{\delta_i}}{1 + \exp(\alpha + \boldsymbol{\beta'}\boldsymbol{x_i})} \end{aligned}$$

donde $\boldsymbol{\beta'} = (\beta_1, \dots, \beta_m)'$ y $\delta_i = 1$ si $Y_i = 1$, y 0 en caso contrario. Es importante tener en cuenta que se basa en la suposición que $f(x)$ no depende de α y $\boldsymbol{\beta'}$; es decir, que no hay sesgo de selección.

Para interpretar los parámetros de este modelo se utiliza el *odds ratio* como medida de asociación. Si tenemos una variable dicotómica, el OR asociado con $X_k = 1$ y ajustado para todas las covariables puede ser expresado mediante:

$$OR_{X_k} = \frac{odds(Y = 1|X_1, \dots, X_k = 1, \dots, X_m)}{odds(Y = 1|X_1, \dots, X_k = 0, \dots, X_m)} = \exp(\beta_k)$$

Por tanto,el estimador del OR_{X_k} y el correspondiente intervalo de confianza son:

$$\begin{aligned} \widehat{OR}_{X_k} &= \exp(\hat{\beta}_k) \\ CI(OR_{X_k}, 1-\alpha) &= \exp\left(\hat{\beta}_k \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{Var}(\hat{\beta}_k)}\right) \end{aligned}$$

En el caso de tener una variable continua, como es la edad, el *odds ratio* asociado comparando dos individuos expuestos que difieren en l unidades es

$$OR_{X_k} = \frac{odds(Y = 1|X_1, \dots, X_k = x + l, \dots, X_m)}{odds(Y = 1|X_1, \dots, X_k = x, \dots, X_m)} = \exp(l \cdot \beta_k)$$

Por último, para interpretar la constante del modelo, que nos indica la probabilidad de que $Y = 1$ en caso de que un individuo tenga valor 0 en todas las covariables, simplemente tendremos que calcular

$$p_0 = \frac{\exp(\alpha)}{1 + \exp(\alpha)}$$

Es importante recordar que esta interpretación es válida en los estudios transversales y de cohorte pero no en los estudios caso-control, ya que en este caso es imposible estimar $P(Y = 1|\mathbf{X})$.

En el caso de los estudios caso-control no podemos ajustar el modelo en (2.3), ya que no se puede estimar la probabilidad de enfermar. En cambio, es posible estimar $P(Y = 1|Z = 1, X)$, donde Z es la variable indicadora de si una persona está incluida en el estudio o no y

$$\pi_i = P(Z = 1|Y = i, \mathbf{X}) = P(Z = 1|Y = i), \;\; i \in \{0, 1\}$$

las probabilidades desconocidas para posibles casos y controles para el muestreo. Es decir, π_1 es la probabilidad de formar parte del estudio teniendo la enfermedad (caso) y π_0 la probabilidad de formar parte, no teniéndola (control). Esto es una premisa del modelo, es decir, aquí también se supone que no hay sesgo de selección. Así,

$$\begin{aligned}
P(Y = 1|Z = 1, \mathbf{X}) &= \frac{P(Z = 1|Y = 1, \mathbf{X})P(Y = 1|\mathbf{X})}{\sum_{i\in\{0,1\}} P(Z = 1|Y = i, \mathbf{X})P(Y = i|\mathbf{X})} \\
&= \frac{\pi_1 P(Y = 1, \mathbf{X})}{\pi_1 P(Y = 1, \mathbf{X}) + \pi_0 P(Y = 0, \mathbf{X})} \\
&= \frac{\pi_1 P(Y = 1, \mathbf{X})/P(Y = 0|\mathbf{X})}{\pi_1 P(Y = 1, \mathbf{X})/P(Y = 0|\mathbf{X}) + \pi_0} \\
&= \frac{\pi_1 \ \exp(\alpha + \boldsymbol{\beta'}\mathbf{X})}{\pi_1 \ \exp(\alpha + \boldsymbol{\beta''}\mathbf{X}) + \pi_0} \\
&= \frac{\exp(\alpha * + \boldsymbol{\beta'}\mathbf{X})}{1 + \exp(\alpha^* + \boldsymbol{\beta'}\mathbf{X})}
\end{aligned}$$

donde $\alpha^* = ln(\pi_1/\pi_0) + \alpha$.

Por tanto, el modelo de regresión logística puede ser ajustado para los estudios caso-control estimando sus parámetros mediante el estimador de máxima verosimilitud. La interpretación de las $\boldsymbol{\beta'}s$ será la misma que en los estudios de cohorte, mientras que la interpretación de la constante del modelo dependerá de π_1 y π_0.

Una vez realizado el ajuste del modelo, se aconseja comprobar la bondad de ajuste. Bajo la hipótesis de una especificación correcta de un modelo, el número de eventos predichos se espera que sea similar al número de eventos observados. Para comprobarlo, uno de los métodos más conocidos hoy en día es el denominado test de Hosmer-Lemeshow [9],[10]. Se trata de un test que ordena los sujetos de acuerdo al riesgo predicho para la enfermedad, $\hat{p} = \hat{P}(Y = 1|\mathbf{X})$ (o $\hat{p} = \hat{P}(Y = 1|Z = 1, \boldsymbol{X})$ en el caso de los estudios caso-control), y los divide en grupos de 5 a 10 aproximadamente del mismo tamaño.

Con cada uno de estos g grupos, el número de eventos observados, O_k, donde $k = 1, \ldots, g$ es comparado con el número de eventos esperado E_k :

$$E_k = \sum_{i=1}^{N_k} \hat{p}_i = \sum_{i=1}^{N_k} \hat{P}(Y = 1|\mathbf{X}_i) = \sum_{i=1}^{N_k} \frac{\exp(\hat{\alpha} + \hat{\beta}'\mathbf{X}_i)}{1 + \exp(\hat{\alpha} + \hat{\beta}'\mathbf{X}_i)}$$

donde N_k es el tamaño del grupo k. El test estadístico del test HL sigue asintóticamente una distribución χ^2 y se define como:

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(O_k - E_k)^2}{V_k} \sim^{H_0} \chi^2_{g-2}$$

Bajo H_0 que el modelo es correcto se esperan valores pequeños de χ^2. Hay que tener en cuenta que una importante desventaja de este test es que depende del número de grupos y de cómo los sujetos estén asignados a estos en caso de empate. Además, no es un test estadístico muy robusto [9].

Modelo de regresión Poisson

Se dice que una variable aleatoria Y sigue una distribución de Poisson de parámetro μ si toma valores enteros $y = 0, 1, 2, \ldots$ con probabilidad

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

para $\mu > 0$. Se puede demostrar que la media y la varianza de la distribución son

$$E(Y) = Var(Y) = \mu$$

La regresión de Poisson es un tipo de modelo lineal generalizado en el que la variable respuesta sigue una distribución de Poisson y su función de enlace es el logaritmo. Se usa generalmente para estudios sobre enfermedades donde los pacientes pueden ser seguidos durante distintos periodos de tiempo, tales como estudios de cohortes de resultados poco frecuentes realizados durante varios años con algunos pacientes que se perdieron durante el seguimiento. Se define de la siguiente forma:

$$E(Y) = \mu = \exp(\alpha + \beta'\mathbf{X})$$

o en su forma log-lineal,

$$\log(E(Y)) = \log(\mu) = \alpha + \beta'\mathbf{X}$$

donde $\mathbf{X} = (X_1, \ldots, X_m)'$ es el vector de covariables.

Esta expresión se utiliza cuando el tiempo de exposición o el tamaño poblacional es el mismo para todos los individuos o poblaciones, respectivamente. En caso contrario, si la exposición varía en función del individuo, tendremos que incorporar dicho tiempo o el tamaño poblacional en el modelo. Por tanto, tendríamos la siguiente expresión del modelo:

$$\log(E(Y)) = \log(\mu) = \log(T) + \alpha + \beta'\mathbf{X}$$

donde el término $\log(T)$ es llamado *offset*.

Al igual que los modelos de regresión logística, el modelo de regresión de Poisson se estima mediante la máxima verosimilitud. Este modelo utiliza la razón de tasas de incidencia como medida de asociación. Por tanto, para una variable categórica la razón de tasas asociado a $X_k = 1$ y ajustado para todas las covariables se define como:

$$IRR_{X_k} = \frac{\exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k + \ldots + \beta_m \cdot X_m)}{\exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m)} = \exp(\beta_k)$$

Para una variable continua, como puede ser la edad, podemos calcular la razón de tasas de incidencia de un individuo con l años más que otro individuo dado:

$$IRR_{X_k} = \frac{\exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k = x + l + \ldots + \beta_m \cdot X_m)}{\exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_k \cdot X_k = x + \ldots + \beta_m \cdot X_m)} = \exp(l \cdot \beta_k)$$

Nos gustaría destacar que en una buena parte de la literatura se interpreta $exp(\beta)$ erroneamente como el RR. Ambas medidas están relacionadas pero no son exactamente iguales.

2.2.7 Propuesta de Zhang y Yu (1998) para estimar el RR

Como brevemente hemos comentado en el apartado 2.2.2, el *odds ratio* se encuentra siempre más lejos del valor 1 que el riesgo relativo. Para saber cuánto de lejos está el *odds ratio* del 1, solamente tendremos que fijarnos en las dos probabilidades condicionadas que lo definen. Si el riesgo de enfermedad es bajo ($< 10\,\%$) en el grupo de expuestos y no expuestos, $\dfrac{1 - P(D|\bar{E})}{1 - P(D|E)}$ en (2.2), está cercano al 1 y por tanto el *odds ratio* y el riesgo relativo son aproximadamente iguales. Este es un resultado a tener en cuenta cuando el riesgo relativo no se puede estimar; como por ejemplo en los estudios caso-control. Sin embargo, cuanto más frecuente es el resultado, mas posible es que el *odds ratio* sobrestime el riesgo relativo cuando este es mayor que 1 o subestime cuando es menor que 1.

La regresión logística es una buena herramienta para ajustar los factores de confusión. Pero cuando la variable de interés es común en la población de estudio el *odds ratio* ajustado puede exagerar una asociación de riesgo o el efecto del tratamiento. Es por ello que en 1998 Jun Zhang y Kai F. Yu [11] propusieron un método simple para aproximarse al riesgo relativo mediante el *odds ratio* ajustado y obtener una estimación de un efecto tratamiento que representa mejor el verdadero riesgo relativo.

La fórmula que proponen requiere la proporción de sujetos de control que experimentan el resultado. En concreto, el riesgo relativo se aproxima:

$$RR = \frac{OR}{(1 - P_0) + (P_0 \times OR)}$$

donde P_0 es la incidencia de los resultados en el grupo no expuesto [6].

Aunque es un método fácil de usar, esta fórmula tiene varias limitaciones: no resulta muy fiable en presencia de covariables, por tanto produce intervalos de confianza más reducidos de lo que deberían ser, se puede sobrestimar ligeramente el RR cuando existe confusión, ignora la covarianza entre la incidencia y el *odds ratio* estimado, y por último, no se puede utilizar en un OR ajustado para estimar un riesgo relativo ajustado ya que es incorrecto y produciría una estimación sesgada cuando la confusión está presente [6].

Como vemos existen varias desventajas tanto al hacer uso de la regresión logística como del método propuesto por Zhang y Yu (1998). Por ello, en el siguiente capítulo presentaremos una alternativa para estimar la asociación entre la enfermedad y la exposición, el modelo de regresión log-binomial.

Capítulo 3

Modelo de regresión log-binomial

3.1 Expresión y estimación de los parámetros

El modelo log-binomial, inicialmente presentado por Wachilder en 1986, es un modelo lineal generalizado para una respuesta binaria que asume que los términos del error siguen una distribución binomial y utiliza el logaritmo como función de enlace [12]. Específicamente, si Y es una variable binaria y $\mathbf{X} = X_1, \ldots, X_m$ es el conjunto de variables explicativas, entonces el modelo log-binomial se define:

$$\log P(Y = 1|\mathbf{X}) = \alpha + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_n$$

siendo $\beta_1, \ldots, \beta_m$ los parámetros del modelo. Por tanto, modelamos la probabilidad de la variable respuesta de interés $(Y = 1)$ como:

$$P(Y = 1|\mathbf{X}) = \exp(\boldsymbol{\beta'} \cdot \mathbf{X}) = \exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m)$$

donde $P(Y = 1|\mathbf{X}) \in (0, 1)$ y $\boldsymbol{\beta'} \cdot \mathbf{X} \leq 0$, para que $\exp(\alpha + \beta_1 \cdot X_1 + \ldots + \beta_m \cdot X_m) \in (0, 1)$.

Para estimar los parámetros β del modelo, maximizamos la verosimilitud de los datos observados [12],

$$\begin{aligned} L(\alpha, \boldsymbol{\beta'}|Y, \mathbf{X}) &= \prod_{i=1}^{n} P(Y = y_i|\boldsymbol{x_i}) f(x_i) \propto \prod_{i=1}^{n} P(Y = y_i|x_i) \\ &= \prod_{i=1}^{n} P(Y = 1|\boldsymbol{x_i})^{\delta_i} P(Y = 0|\boldsymbol{x_i})^{1-\delta_i} \\ &= \prod_{i=1}^{n} \exp(\alpha + \boldsymbol{\beta'} \boldsymbol{x_i})^{\delta_i} \cdot (1 - \exp(\alpha + \boldsymbol{\beta'} \boldsymbol{x_i}))^{1-\delta_i} \end{aligned}$$

donde $\boldsymbol{\beta'} = (\beta_1, \ldots, \beta_m)'$ y $\delta_i = 1$ si $Y_i = 1$, y 0 en caso contrario. Al igual que en el modelo logístico, es importante tener en cuenta que se basa en la suposición que $f(x)$ no depende de α y $\boldsymbol{\beta'}$, es decir, que no hay sesgo de selección.

Es un punto importante a destacar que dependiendo del tipo de estudio que tengamos, la medida de asociación será distinta. Si estamos ante un estudio transversal la medida de asociación será la razón de prevalencia (PR), mientras que en los estudios longitudinales y cohorte la medida será el riesgo relativo (RR). Esta es una de las ventajas de este modelo, ya que éstas medidas de asociación resultan mucho más interpretables que el *odds ratio*. Conviene señalar que no es posible usar este modelo para estudios caso-control, dado que como vimos en el capítulo anterior el riesgo relativo no es estimable en este tipo de estudios. Entonces,

$$\begin{aligned} PR/RR &= \frac{P(Y=1|X_1,\ldots,X_k=1,\ldots,X_n)}{P(Y=1|X_1,\ldots,X_k=0,\ldots,X_n)} = \\ &= \frac{\exp(\alpha+\beta_1\cdot X_1+\ldots+\beta_k\cdot X_k+\ldots+\beta_m\cdot X_m)}{\exp(\alpha+\beta_1\cdot X_1+\ldots+\beta_m\cdot X_m)} = \exp(\beta_k) \end{aligned}$$

por cada unidad que incrementa en X_k, ajustado para la otras variables.

Como sucede con el modelo de regresión de Poisson, para una variable continua en la que queremos calcular el riesgo relativo con l unidades de diferencia para un individuo comparado con otro dado:

$$PR/RR = \frac{\exp(\alpha+\beta_1\cdot X_1+\ldots+\beta_k\cdot X_k=x+l+\ldots+\beta_m\cdot X_m)}{\exp(\alpha+\beta_1\cdot X_1+\ldots+\beta_k\cdot X_k=x+\ldots+\beta_m\cdot X_m)} = \exp(l\cdot\beta_k)$$

Al igual que definíamos en los modelos de regresión logística, para interpretar la constante del modelo (la probabilidad de $Y=1$ con todas las covariables igual a 0) simplemente procederemos a calcular:

$$p_0 = P(Y=1|\mathbf{X}=0) = \exp(\alpha)$$

Cuando ya hemos obtenido una solución aceptable para el modelo log-binomial, el siguiente paso es comprobar la bondad de ajuste. Para ello, Blizzard y Hosmer (2004) [14] presentaron varias extensiones de la bondad de ajuste del modelo de regresión logística para el modelo log-binomial. Entre ellas, el nombrado anteriormente test de Hosmer y Lemeshow (HL).

Como ya hemos explicado antes, la idea básica de este método es comparar las frecuencias de los valores observados para modelar las frecuencias estimadas entre grupos, basándose en la ordenación de los valores ajustados. Puede utilizarse cualquier número de grupos, pero el más usado en la práctica es g=10. Como ya comentamos en apartados anteriores, el test de HL se aproxima a una χ^2 con g-2 grados de libertad. Los autores comentan que existe un problema cuando utilizas la regla del "número de grupos menos dos", y es que la suma de los valores observados y los esperados no es la misma. Sin embargo han comprobado que esos valores están extremadamente cerca, aunque no sean exactamente iguales. Mediante varias simulaciones demuestran que si se ajusta el modelo de forma correcta, el test se aproxima adecuadamente a una distribución χ^2 con $g-2$ grados de libertad.

Por consiguiente, comparando el modelo log-binomial con el de regresión logística podríamos decir que entre las igualdades se encuentra que los dos modelos son empleados para el análisis de una variable respuesta dicotómica, ambos modelan la probabilidad de la variable respuesta y asumen que el error sigue una distribución binomial.

Entre las diferencias destacaríamos que la regresión logística utiliza logit como función de enlace mientras que el log-binomial usa el logarítmo, que el primero utiliza el *odds ratio* como medida del efecto mientras que el segundo usa el riesgo relativo o la razón de prevalencia y que es posible usar la regresión logística para estudios caso-control mientras que en la log-binomial no.

Como ya comentábamos en el apartado 2.2.6, a veces el OR puede exagerar el efecto del tratamiento y por ello el modelo de regresión log-binomial es un buen sustituto para estos casos. Sin embargo, aunque sea una buena opción, hoy en día es muy poco común encontrar información sobre ellos principalmente por dos razones: la primera es que el software está muy poco desarrollado y la segunda y más importante, porque dependiendo de la situación en la que nos encontremos surgen problemas de convergencia. A continuación explicaremos más detalladamente por qué suceden estos problemas de estimación.

3.2 Problemas de estimación

El problema principal de este modelo es que en ocasiones el proceso de maximización de la función de verosimilitud falla al intentar encontrar el estimador de máxima verosimilitud, creando problemas de convergencia. Como ya hemos comentado al presentar el modelo de regresión log-binomial, $P(Y = 1) \in (0, 1)$ y $\exp(\alpha + \boldsymbol{\beta}'\boldsymbol{X}) > 0$, por lo que este último puede tomar valores por encima del 1, saliendo del espacio de parámetros restringido. Esta es la razón principal por la que suceden estos problemas de estimación. A veces simplemente son debidos al aporte de unos valores iniciales no adecuados, por ejemplo, valores iniciales que no se encuentran en el espacio de parámetros restringido. Para solucionarlo, el usuario simplemente tendrá que elegir unos valores iniciales nuevos. Sin embargo, otras veces sucede que la solución del máximo verosímil se encuentra en el límite del espacio de parámetros y la derivada de la probabilidad en su máxima no puede ser 0. Por ello, muchos paquetes de software que maximizan la probabilidad encontrando el punto en el que la derivada es igual a 0 (llamado método de Newton) no pueden encontrar la solución [13].

3.3 Software estadístico

En esta sección haremos un breve resumen de los software estadísticos disponibles para el modelo log-binomial y qué soluciones presentan para los problemas de convergencia, tomando como referencia la información aportada por Williamson et al. (2013) [15].

Comenzaremos hablando del software estadístico R, ya que es uno de los más utilizados hoy en día por ser de libre uso. Este paquete utiliza la función *"glm"* que ajusta modelos lineales generalizados, especificando la función de enlace y la distribución que sigue el error junto con las variables que se quieren fijar en el modelo. Como bien nos comentan Williamson et at. este paquete tiene una considerable ventaja, y es que aunque devuelva problemas de convergencia es capaz de asegurarse que los valores ajustados se encuentran dentro del espacio de parámetros. De esta forma garantiza que los valores son siempre negativos. Si los valores positivos se encuentran en el principio de la estimación, entonces el ajuste se detiene y pide al usuario unos valores iniciales mejores.

Por otro lado, R también hace uso de un proceso conocido como "step-halving". Si la iteración trata de situarse fuera del espacio de parámetros la actualización se reduce a la mitad y recalcula el valor ajustado. Repite este proceso una y otra vez mientras siga siendo un valor positivo hasta que el valor ajustado sea negativo. Con esto nos referimos a que el predictor lineal, $\hat{\alpha} + \hat{\boldsymbol{\beta}}'\mathbf{X}$, también ha de ser siempre negativo, con $\alpha < 0$ siempre pero pudiendo haber $\beta > 0$. De esta forma se asegura que aunque no converja los valores siempre se encuentren dentro del espacio de parámetros.

En el software SAS se utiliza la función PROC GENMOD. En este paquete no ocurre lo mismo que en R, ya que no es posible asegurarse de que no itere fuera del espacio de parámetros. Sin embargo, si que es capaz de notificar que no encuentra el MLE y no continua con el proceso de búsqueda.

Por otro lado, al igual que para R, STATA utiliza la función "glm" para implementar este modelo. Con este software no ocurre ninguna de las dos cosas anteriomente nombradas para los otros paquetes, ya que es posible que itere fuera del espacio de parámetros y no es capaz de parar el proceso aunque no encuentre el estimador de máxima verosimilitud. El usuario tiene que ser consciente de ello y ser él el que pare el proceso.

Por último, el software SPSS, mediante la función GENLIN hace uso del proceso "step-halving" en su implementación, al igual que sucede en R, y es capaz de devolver un error alertando al usuario de que existen casos de la base de datos que no son válidos.

3.4 Soluciones para los problemas de convergencia

Debido a que no se tiene demasiado información sobre el modelo de regresión log-binomial, ya que apenas es nombrado en la literatura, no existen demasiados soluciones a este problema a la hora de implementarlos en cualquiera de los software anteriormente nombrados. Es por ello que nuestro objetivo principal en este trabajo es conseguir un método eficaz para poder solucionar el problema de convergencia en el software estadístico R. A continuación nombraremos algunos de los métodos ya desarrollados en los últimos años explicando brevemente en qué consisten.

En 2003 Deddens et al. desarrollaron el denominado método COPY en el software estadístico SAS [16]. Se basa en modificar la base de datos original para conseguir aproximar el estimador de máxima verosimilitud. Es un método sencillo de aplicar que no ha sido desarrollado en otros sotfware estadísticos como Stata debido a que no es capaz de solucionar los problemas de convergencia. Por ello, nuestra propuesta es poder implementarlo en R consiguiendo unos resultados satisfactorios, al igual que ocurre en SAS. En el siguiente punto explicaremos más detalladamente en qué consiste este método y cómo lo hemos aplicado.

Por otro lado, en 2015 Mark W. Donoghoe presentó la función de R *"logbin"* [24] del paquete con el mismo nombre que expone distintos métodos para ajustar modelos lineales generalizados con función de enlace log y datos binomiales usando el algoritmo de esperanza-maximización (EM) con propiedades de convergencia más estables que los métodos estándar.

Este algoritmo es un método iterativo para encontrar el estimador de máxima verosimilitud en modelos estadísticos cuando las ecuaciones no se pueden resolver directamente y donde el modelo depende de variables latentes no observadas. Como él mismo destaca, este algoritmo se acomoda a las limitaciones de los parámetros y es más estable que los mínimos cuadrados iterativamente reponderados. Se define una colección de espacios de parámetros restringidos que cubre el espacio de parámetros completo y se aplica el algoritmo EM dentro de cada parámetro de espacio restringido con el fin de encontrar una colección de máximos restringidos de la función de log-verosimilitud, a partir del cual se puede obtener el máximo global sobre el espacio de parámetros completo.

Para poder comparar estas funciones, en el Capítulo 4 presentaremos distintas bases de datos (tanto con problemas de convergencia como sin ellos) y sus implementaciones, contrastando los resultados obtenidos y viendo cuál de los métodos consideraríamos que devuelve resultados más satisfactorios.

3.5 Método COPY

Para poder solucionar el problema de la convergencia, hemos implementado el método COPY desarrollado por Deddens et al. en 2003 usando el software estadístico R [16]. Este método consiste en realizar $(n-1)$ copias de la base de datos inicial y una copia de esta misma base con los valores de la variable de interés intercambiados, es decir, los 1's se cambian por 0's y los 0's por 1's. Esto hará que se obtenga un estimador de máxima verosimilitud prácticamente igual al que hubiésemos obtenido sin realizar ningún cambio, pero sin estar en el límite del espacio de parámetros. Cuando no existen problemas de convergencia, procedemos a usar la base de datos original, sin realizar ningún cambio. Para tener en cuenta el tamaño muestral, el error estándar estimado se multiplicará por la raíz cuadrado del n elegido. Es importante escoger un n suficientemente grande para que el nuevo estimador de máxima verosimilitud esté lo suficientemente próximo al real, pero no tan grande como para que el software no sea capaz de estimarlo.

Pero unos años más tarde, Lumley et al.(2006) [18] demostraron que este método es equivalente a crear una nueva base de datos que contenga una copia de la base de datos inicial con el peso $w = (n-1)/n$ y una copia de la base de datos original con los valores de la variable principal intercambiados y un peso $w = 1/n$. En este caso, no es necesario reajustar los valores del error estándar. Para implementarlo, se utiliza la regresión log-binomial ponderada maximizando la verosimilitud:

$$L(\alpha, \boldsymbol{\beta'}|Y, \mathbf{X}) \quad = \quad \prod_{i=1}^{n} \exp(\alpha + \boldsymbol{\beta'}\boldsymbol{x_i})^{w\delta_i + (1-w)(1-\delta_i)} \cdot (1 - \exp(\alpha + \boldsymbol{\beta'}\boldsymbol{x_i}))^{w(1-\delta_i)+(1-w)\delta_i}$$

De este modo se obtienen estimadores de máxima verosimilitud aproximados a los que conseguiríamos usando la base de datos original pero sin tener problemas con la convergencia. Nosotros nos basaremos en ésta última idea para crear el código en R, ya que es mucho más eficiente computacionalmente hablando.

3.6 Implementación del método COPY en R

En esta sección explicaremos las características básicas necesarias para la implementación del método COPY (A. Deddens et al. [16]). Comenzaremos diciendo que es necesario introducir unos valores iniciales tanto para el intercepto como para las variables que queramos ajustar en el modelo. Los autores nos comentan que en la práctica usando el software SAS, cuando se utiliza la función PROC GENMOD el valor -4 para el intercepto y 0 para el resto de variables siempre ha funcionado bien. Por lo tanto, estos serán los valores que nosotros introduciremos a la hora de programar la función, ya que dependiendo del caso, R pide introducir unos valores iniciales apropiados. Otro punto a tener en cuenta es la elección del tamaño de n. Los autores realizaron varias simulaciones en SAS, llegando a la conclusión de que en la mayoría de los casos un número de 1000 copias era el idóneo.

Sin embargo, en una revisión del método nos advierten que dependiendo del caso en que nos encontremos el número de copias, o en nuestro caso el peso, puede variar [17]. Nosotros probaremos con valores entre 100 y 10.000, comparando los resultados obtenidos.

Por tanto, teniendo en cuenta estas características, el código de R para implementar el método COPY es el siguiente:

```
copy <- function(data, Y, vars,n, W) {
     if (!is.numeric(data[, Y])) {
         data[, Y] <- as.numeric(data[, Y])-1
     }
    data$W <- (n-1)/n
    data.copy <- data
    data.copy[, Y] <- 1-data.copy[, Y]
    data.copy$W <- 1/n
    data.all <- merge(data, data.copy, all = T)
    formul <- paste(Y, paste(vars, collapse = " + "), sep = "~")
    mod.mat <- model.matrix(as.formula(formul), data)
    glm.copy <- glm(as.formula(formul), family = binomial(log), data.all,
    weights = W, control = list(maxit = 100),
    start = c(-4, rep(0, ncol(mod.mat)-1)))
    return(glm.copy)
}
```

Capítulo 4

Ejemplos con datos reales

Para comprobar todo lo anteriormente expuesto, en este capítulo analizaremos distintas bases de datos. En todos los casos, ajustaremos tanto el modelo de regresión logística como el log-binomial mediante las distintas funciones ya existentes en R, concretamente *"glm"* y *"logbin"*, y la función anteriormente presentada COPY. Por último, procederemos a comparar los resultados obtenidos. A continuación mostramos varios ejemplos entre los que se encuentran un estudio cohorte (base de datos Diet) y un estudio transversal (base de datos Can Ruti).

4.1 Base de datos Diet

El data frame Diet del paquete de R *"Epi"* [26] contiene 337 filas y 14 columnas. Los datos pertenecen a una submuestra de sujetos extraídos de estudios de cohorte de la incidencia de la enfermedad cardíaca coronaria (CHD). Tras el cierre de la Unidad de Medicina Social MRC (Medical Research Council), desde donde se dirigen estos estudios, se encontró que se habían producido 46 casos de CHD en este grupo, lo que permite un estudio fortuito de la relación entre la dieta y la incidencia de las enfermedades del corazón. La variable chd indica la aparición (chd=1) o ausencia (chd=0) de una enfermedad coronaria durante el tiempo de seguimiento (variable y).

Se quiere estudiar la asociación entre la incidencia de enfermedades coronarias y el aporte de energía diario (variable energy.grp) en presencia de otros posibles factores de riesgo y/o 5 variables de interés como pueden ser la altura, el peso, el índice de masa corporal, la profesión o el aporte diario de fibras y grasas.

A continuación presentamos una tabla con la descripción de cada una de las variables (n, %, media, desviación estándar).

	No N=291	Si N=46
Años bajo riesgo	14.6 (4.05)	7.73 (4.79)
Trabajo:		
Chofer	90 (30.9 %)	12 (26.1 %)
Conductor	70 (24.1 %)	14 (30.4 %)
Trabajador banco	131 (45.0 %)	20 (43.5 %)
Energia (KCal por día/100)	28.6 (4.43)	26.5 (3.88)
Altura (cm)	174 (6.23)	170 (6.48)
Peso (kg)	72.8 (10.7)	70.7 (11.0)
Grasa ingerida (gr/día)	12.9 (2.36)	11.8 (2.21)
fibra ingerida (gr/día)	1.75 (0.58)	1.49 (0.40)
Aporte de energía díario		
≤2750 KCals	127 (43.6 %)	28 (60.9 %)
>2750 KCals	164 (56.4 %)	18 (39.1 %)
IMC	24.1 (3.19)	24.5 (3.35)

Tabla 4.1: Tabla descriptiva de la base de datos Diet en función de la variable CHD con n (%) para variables categóricas y media (sd) para numéricas

4.1.1 Modelo de regresión logística

Para ajustar el modelo, procedemos a seguir varias pautas. Primero consideramos las siguientes variables de interés: trabajo, edad, ingesta de fibra, ingesta de grasas, aporte de energía diario e índice de masa corporal ($IMC = peso/altura^2$(en metros)). Después, categorizamos la variable de interés principal del estudio 'Aporte de energía diario' en dos grupos, $>$ 2750 KCals vs. $\leq$ 2750 KCals, siendo esta última la categoría de referencia. Por último, ajustamos distintos modelos siguiendo el criterio de información Akaike, eliminando las variables no significativas hasta conseguir el que hemos considerado como mejor modelo final:

$$logit(p) = ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4$$

En nuestro caso,

$$logit(p) = ln\left(\frac{p}{1-p}\right) = 0{,}012 + 0{,}27 \cdot X_1 - 0{,}211 \cdot X_2 - 1{,}058 \cdot X_3 + 0{,}095 \cdot X_4$$

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{OR}$
Constante	0.012	1.606	0.007	0.994	
Aporte de energía	0.27	0.46	0.586	0.558	1.31
Ingesta de grasas	-0.211	0.107	-1.977	0.048	0.81
Ingesta de fibra	-1.058	0.451	-2.344	0.019	0.35
IMC	0.095	0.054	1.766	0.077	1.1

Tabla 4.2: Modelo de regresión logística para la incidencia de enfermedades coronarias

Podemos decir que la ingesta de grasas y fibra son (posibles) factores protectores para las enfermedades coronarias mientras un aumento del IMC implica un aumento del riesgo para dichas enfermedades (aunque no estadísticamente significativo a un nivel del 95 %).

4.1.2 Modelo log-binomial

Función *"glm"*

En este caso usamos la función "*glm*" de R de la misma forma que hemos hecho para ajustar el modelo de regresión logística, pero especificando cuál es su función de enlace. Obtenemos el siguiente modelo:

$$\log(p) = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3 + \beta_4 \cdot X_4$$

En nuestro caso,

$$\log(p) = -0{,}620 + 0{,}186 \cdot X_1 - 0{,}18 \cdot X_2 - 0{,}851 \cdot X_3 + 0{,}086 \cdot X_4$$

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-0.620	1.28	-0.485	0.628	
Aporte de energía	0.186	0.385	0.485	0.628	1.204
Ingesta de grasas	-0.18	0.085	-2.123	0.034	0.835
Ingesta de fibra	-0.851	0.367	-2.317	0.021	0.427
IMC	0.086	0.04	2.180	0.029	1.09

Tabla 4.3: Modelo log-binomial (función *"glm"*) para la incidencia de enfermedades coronarias

Lo primera conclusión que sacamos al ejecutarlo, es que con esta base de datos el software consigue encontrar el estimador de máxima verosimilitud, por tanto no tenemos ningún problema con la convergencia. Aún así, procederemos a comparar los resultados tanto con esta función como con las explicadas en los siguientes dos puntos.

Podemos decir que los valores obtenidos al ajustar este modelo son muy parecidos a los obtenidos en el modelo logístico. Observamos que la ingesta de grasas y fibra siguen siendo posibles factores protectores para las enfermedades coronarias. Sin embargo, el IMC es un posible factor de riesgo, esta vez significativo a un nivel del 95 %. El riesgo de enfermedad coronaria es 1.1 veces mayor en el caso de gente con índice de masa corporal fuera de los límites normales.

Vemos que el $\widehat{RR} \approx \widehat{OR}$, ya que tenemos pocos casos (46 entre 337, $\leq 2\,\%$) y también observamos que el $\widehat{RR}$ está siempre más cercano a 1, como ya comentábamos en apartados anteriores.

Función *"logbin"*

Volvemos a ajustar el mismo modelo que en el caso anterior pero esta vez mediante la función *"logbin"* de R. A la hora de ajustar el modelo con la base de datos*Diet*, R nos devuelve varios mensajes de advertencia (*warnings*):

```
Warning messages:
1: nplbin: algorithm did not converge within 10000 iterations
           -- increase 'maxit'.
2: nplbin: fitted probabilities numerically 1 occurred
```

Aún tratándose de avisos, la salida no devuelve los p-valores, el valor z ni la desviación estándar.

$$logit(p) = -0{,}741 - 0{,}006 \cdot X_1 - 0{,}139 \cdot X_2 - 0{,}712 \cdot X_3 + 0{,}065 \cdot X_4$$

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor
Constante	-0.741	NA	NA	NA
Aporte de energía	0.006	NA	NA	NA
Ingesta de grasas	-0.139	NA	NA	NA
Ingesta de fibra	-0.712	NA	NA	NA
BMI	0.065	NA	NA	NA

Tabla 4.4: Modelo log-binomial (función *"logbin"*) para la incidencia de enfermedades coronarias

Por tanto, intentamos solucionarlo ampliando en principio el número de iteraciones y cambiando el valor de tolerancia positiva especificando el interior del espacio de parámetros. Aún probando con distintos valores, no conseguimos aproximarnos a la solución.

Otro punto importante que nos gustaría destacar sobre esta función es que emplea demasiado tiempo en ejecutar el modelo, teniendo en cuenta que no se trata de un modelo con demasiadas variables ni de una base de datos muy grande. Por lo que esto sería un claro inconveniente a la hora de ajustar un modelo con muchas variables o cuando tenemos demasiados datos en nuestra base.

Función "COPY"

Finalmente ejecutamos la función que presentamos en este trabajo para nuestra base de datos Diet y obtenemos la tabla presentada a continuación.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-0.625	1.276	-0.490	0.625	
Aporte de energía	0.185	0.383	0.482	0.63	1.203
Ingesta de grasas	-0.178	0.084	-2.113	0.035	0.837
Ingesta de fibra	-0.843	0.366	-2.305	0.021	0.43
BMI	0.086	0.086	0.039	0.029	1.09

Tabla 4.5: Modelo log-binomial (función COPY) para la incidencia de enfermedades coronarias

Hemos probado tanto con n=100 como con n=1000, y hemos comprobado que para este ejemplo concreto cuando cogemos n=1000 las estimaciones son unas décimas más próximas a las obtenidas mediante la función *"glm"*. Por tanto solamente mostramos los resultados de éste último.

Nos damos cuenta que las estimaciones tanto de la constante como de las variables son prácticamente iguales a las de la función *"glm"*. Con esta base de datos la función COPY obtiene mejor resultados que la función *"logbin"*, dado que aún no siendo necesario su uso, consigue valores muy aproximados a los estimadores de ML de las β. Por tanto, las interpretaciones serían las mismas que en el caso anterior.

Por último, hemos comprobado la bondad de ajuste de todos los modelos mediante el test de Hosmer-Lemeshow, consiguiendo valores por encima de 0.1 en todos los casos salvo para el método COPY, para el que no nos devuelve valores coherentes. Es posible que este test no funcione para este método, sin embargo ni hemos encontrado información en la literatura sobre este tema.

4.2 Tablas del paper Williamson et al. (2013)

Analizaremos también los ejemplos simples que nos proporcionan en el artículo de Williamson et al. (2013) [15] para ver qué ocurre cuando tenemos problemas de convergencia. Al igual que hemos hecho en los apartados anteriores ajustaremos el modelo de regresión logística y el modelo log-binomial mediante las funciones *"glm"*, *"logbin"* y COPY.

4.2.1 Caso 1: máximo en un límite finito

Como ellos mismos explican, si la función de verosimilitud se maximiza en el límite del espacio de parámetros un método iterativo puede tener problemas en encontrarlo mientras que es posible que el algoritmo pise en un espacio ilegal. Un ejemplo de este caso se presenta en la siguiente tabla:

	(X=-1)	(X=0)	(X=1)	
(Y=1) Enfermedad	10	18	5	33
(Y=0) No enfermedad	8	9	0	17
	18	27	5	50

Tabla 4.6: Caso 1: Máximo en un límite finito

Modelo de regresión logística

Estamos ante un modelo con un solo predictor de 3 niveles (X= -1, 0, 1). Creamos la base de datos a partir de la tabla y ajustamos el modelo logístico mediante la función *"glm"*, al igual que en los casos anteriormente vistos. Presentamos los resultados en la Tabla 4.7.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor
Constante	-0.223	0.474	-0.470	0.638
X=0	-0.470	0.626	-0.751	0.453
X=1	-17.343	1769.258	-0.010	0.992

Tabla 4.7: Modelo de regresión logística (función *"glm"*) para la Tabla 1 de Williamson et.al. (2013)

Comprobamos que los valores de los coeficientes no son "normales" cuando X=1, y esto es debido al 0 que tenemos en nuestra tabla, es decir, cuando no hay ningún individuo no enfermo para X=1. Para intentar corregirlo probamos a ajustarlo mediante la función *"logistf"* del paquete de R [27] con el mismo nombre que nos facilita la aplicación del procedimiento del Score de Firth modificado en la regresión logística.

No explicaremos detalladamente como funciona, pero sí diremos que se basa en la idea de Firth (1993) [19] de maximizar la log-verosimilitud penalizada. Después de reajustarlo, vemos que conseguimos valores con más sentido.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	chisq	p	$\widehat{OR}$
Constante	4.481	2.419	0.000	1.000	
X=0	-1.874	2.538	25.902	$< 0{,}001$	0.154
X=1	-3.398	2.629	19.395	$< 0{,}001$	0.033

Tabla 4.8: Modelo de regresión logística (función "logistf") para la Tabla 1 de Williamson et.al. (2013)

Es importante destacar que este método usa un estadístico (χ^2) distinto del de Wald. También comentar que en este ejemplo no interpretaremos los resultados y simplemente veremos que sucede en el cálculo de las estimaciones.

Modelo Log-binomial

Función *"glm"*

Al igual que nos sucedía con la regresión logística, conseguimos coeficientes fuera de lo común, muy elevados y nada coherentes. En este caso no tenemos ningún método que nos sirva para solucionar el problema, por tanto, podríamos decir que no obtenemos resultados satisfactorios con la función *"glm"*.

Función *"logbin"*

Analizamos qué sucede cuando ajustamos mediante la función *"logbin"*, y si realmente soluciona los problemas de convergencia obteniendo estimadores válidos. Destacaremos que para no obtener ningún aviso a la hora de ejecutarlo hemos probado distintas soluciones variando el valor tanto de ϵ como de la tolerancia de los límites. En este modelo también ha sido necesario definir la variable D como numérica. En la tabla mostramos solamente los valores obtenidos con $\epsilon = 5e^{-}2$.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-0.974	0.304	-3.219	$< 0{,}001$	
X=0	-0.232	0.422	-0.549	0.583	0.793
X=1	0.176	0.579	0.304	0.761	1.192

Tabla 4.9: Modelo de regresión log-binomial (función *"logbin"*) para la Tabla 1 de Williamson et.al. (2013)

Fijándonos en los resultados, podemos decir que no hay ningún valor "extraño" y que se podrían dar por válidos. Pero antes de llegar a ninguna conclusión precipitada, veremos que sucede si utilizamos el método COPY.

Función COPY

Por último, empleando el método COPY, obtenemos los resultados de la Tabla 4.10.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p-valor	$\widehat{RR}$
Constante	-0.81	0.263	-3.077	0.002	
X=0	-0.287	0.379	-0.758	0.449	0.751
X=1	-6.097	14.124	-0.432	0.666	0.002

Tabla 4.10: Modelo de regresión log-binomial (función COPY) para la Tabla 1 de Williamson et.al. (2013)

Al igual que con la base de datos Diet, probamos tanto para n=100 como para n=1000. En los dos hemos obtenido las mismas estimaciones. Debido a que no ha sido posible conseguir resultados satisfactorios en los apartados anteriores, no podemos compararlos. Lo que sí podemos destacar es que no nos devuelve ningún error advirtiéndonos de problemas de convergencia. También es importante comentar, que es posible que los resultados se vean afectados por el 0 que tenemos en la tabla, es decir, cuando no hay ningún individuo no enfermo para X=1. Ya hemos visto que para la regresión logística es posible utilizar la función "logistf" cuando nos encontramos ante situaciones como esta, pero no existe ninguna implementación de este método para los modelos log-binomial.

4.2.2 Caso 2: Máximo dentro del espacio de parámetros

Si la solución reside en el interior del espacio de parámetros (es decir, no en el límite) entonces si observamos una convergencia fallida, cuando existe un máximo finito, debe ser considerado como un fracaso del método numérico. Veamos un ejemplo de este caso:

	(X=-1)	(X=0)	(X=1)	
(Y=1) Enfermedad	2	14	2	18
(Y=0) No enfermedad	2	3	17	22
	4	17	19	40

Tabla 4.11: Caso 2: Máximo dentro del espacio de parámetros

Modelo de regresión logística

Mediante la función *"glm"* conseguimos estas estimaciones:

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{OR}$
Constante	$<$ 0,001	1.000	0.000	1.000	
X=0	-1.540	1.185	-1.300	0.194	1.214
X=1	2.140	1.249	1.714	0.087	8.499

Tabla 4.12: Modelo de regresión logística (función *"glm"*) para la Tabla 3 de Williamson et.al. (2013)

Con nuestra versión de R (versión 3.2.5) no nos devuelve ningún error ni mensaje de advertencia. Además, no vemos ningún valor "extraño", por lo que consideramos estas estimaciones correctas. Podemos decir que tanto X=0 y X=1 son factores de riesgo para la enfermedad, siendo para este último el riesgo muy elevado. A continuación veremos que obtenemos al ajustar el modelo de regresión log-binomial con las 3 funciones anteriormente expuestas.

Modelo Log-binomial

Función *"glm"*

Al ajustar el modelo mediante la función *"glm"* R nos pide que introduzcamos valores iniciales, así que utilizamos los mismos valores que cuando empleamos la función COPY, es decir, -4 para el intercepto y 0 para el resto de variables. Por tanto, usando el mismo método que antes y ajustando el modelo, conseguimos:

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{RR}$
Constante	-0.693	0.500	-1.386	0.166	
X=0	-1.041	0.724	-1.438	0.150	0.353
X=1	0.582	0.506	1.150	0.250	1.79

Tabla 4.13: Modelo de regresión log-binomial (función *"glm"*) para la Tabla 3 de Williamson et.al. (2013)

En este caso R sí que consigue sacar unos valores para los estimadores. Si comparamos el riesgo relativo con el *odds ratio* del modelo logístico vemos que el riesgo relativo es bastante menos elevado. Además, en este caso X=0 es un factor protector para la enfermedad. Veamos a continuación que sucede al implementar las otras dos funciones.

Función *"logbin"*

Conseguimos exactamente los mismo valores que con la función anterior, sin ningún problema de convergencia. Nos gustaría destacar que para este ejemplo, al contrario que en los anteriores, no ha sido necesario incluir ningún argumento para que nos devuelva todos los resultados. Si tenemos en cuenta los resultados anteriores como correctos, ésta función también será válida.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p
Constante	-0.693	0.500	-1.386	0.166
X=0	-1.041	0.724	-1.438	0.150
X=1	0.582	0.506	1.150	0.250

Tabla 4.14: Modelo de regresión log-binomial (función *"logbin"*) para la Tabla 3 de Williamson et.al. (2013)

Función COPY

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p
Constante	-0.693	0.500	-1.386	0.166
X=0	-1.038	0.723	-1.435	0.151
X=1	0.581	0.506	1.148	0.251

Tabla 4.15: Modelo de regresión log-binomial (función COPY) para la Tabla 3 de Williamson et.al. (2013)

Finalmente, vemos que esta función también consigue sacar los coeficientes. Por tanto, para este ejemplo en concreto sería posible emplear cualquiera de las tres funciones.

Para estos dos ejemplos (casos 1 y 2) no es posible realizar el ajuste de bondad mediante el test de Hosmer-Lemeshow, ya que es una base de datos con solo tres valores predichos $\hat{p}$ diferentes y por tanto muchos empates en las probabilidades predichas.

4.3 Base de datos Constrict

En este apartado utilizaremos la base de datos Constrict disponible en la documentación de SAS PROC LOGISTIC [23]. Compararemos los resultados conseguidos para el modelo log-binomial con los obtenidos por Deddens et al. (2008) [13], ya que utilizaron este mismo ejemplo con el método COPY en el software estadístico SAS.

Se trata de datos de un experimento para estudiar el efecto de la velocidad y el volumen de entrada de aire en una vasoconstricción refleja transitoria en la piel de los dedos. Se realizaron 39 ensayos bajo diversas combinaciones de velocidad y el volumen de entrada de aire (Finney, 1947) y se tomó como final de cada prueba si se produjo o no la vasoconstricción. De los 39 ensayos, hay un total de 20 casos y 19 controles. Se utiliza como variable de respuesta de interés, restringido=1 o restringido=0 y se modela respecto al logaritmo de la velocidad y volumen de aire.

4.3.1 Modelo de regresión logística

Tal y cómo nos describen los autores, transformamos las variables al logaritmo y ajustamos el modelo:

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{OR}$
Constante	-2.875	1.321	-2.177	0.029	
log(Volumen)	5.179	1.865	2.778	0.005	13.323
log(Velocidad)	4.562	1.838	2.482	0.013	9.786

Tabla 4.16: Modelo de regresión logística para base de datos Constrict

El $\widehat{OR}$ es asociado a un incremento de 0.5 en el logaritmo. Vemos que aún así los coeficientes son bastante elevados. Esto nos indica que una mayor velocidad o mayor entrada de aire es posible que aumente el *odds* de la vasoconstricción. Según los datos, el *odds* para "restringido" aumenta en función del volumen y la velocidad, es decir, cuando el volumen aumenta la velocidad disminuye y cuando la velocidad aumenta el volumen disminuye. Es posible que debido a esta correlación inversa entre las variables explicativas obtengamos valores tan elevados.

4.3.2 Modelo log-binomial

Función *"glm"*

Al intentar ajustar el modelo con la función *"glm"* R devuelve un error de convergencia en formato de mensaje de advertencia. Sin embargo, resulta curioso que es capaz de devolver un resultado. Comparando con el de los autores en el software estadístico SAS, no son realmente los valores exactos, pero están muy próximos.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{RR}$
Constante	-1.501	0.385	-3.900	$< 0{,}001$	
log(Volumen)	0.771	0.191	4.028	$< 0{,}001$	2.162
log(Velocidad)	1.286	0.364	3.529	$< 0{,}001$	3.618

Tabla 4.17: Modelo de regresión log-binomial (función *"glm"*) para base de datos Constrict

Función *"logbin"*

En el caso de la función *"logbin"* ocurre lo mismo que en ejemplos anteriores, nos devuelve un mensaje de advertencia que nos sugiere cambiar o la tolerancia de los límites o el valor del ϵ, que indica la tolerancia de convergencia positiva. Un punto importante a destacar es que dependiendo de los valores introducidos devolverá distintas soluciones. Hemos comenzado con $\epsilon = 0{,}1$ y hemos ido disminuyendo el valor. Aún así, los resultados más próximos a lo obtenidos en el apartado anterior son los siguientes:

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{RR}$
Constante	-0.898	0.252	-3.556	$< 0{,}001$	
log(Volumen)	0.399	0.306	1.307	0.191	1.49
log(Velocidad)	0.284	0.301	0.942	0.346	1.328

Tabla 4.18: Modelo de regresión log-binomial (función *"logbin"*) para base de datos Constrict

Estos valores, aun siendo los más próximos obtenidos, siguen siendo distintos a los conseguidos con la función *"glm"*. Por tanto, no consideramos esta función una buena opción para esta base de datos en concreto.

Función COPY

Por último, al aplicar el método COPY conseguimos exactamente los mismos resultados que los autores usando el software estadístico SAS. En este caso también hemos decidido quedarnos con n=1000.

	$\hat{\beta}$	s.e.($\hat{\beta}$)	Z	p	$\widehat{RR}$
Constante	-1.514	0.377	-4.014	$< 0{,}001$	
log(Volumen)	0.771	0.192	4.014	$< 0{,}001$	2.162
log(Velocidad)	1.313	0.329	3.987	$< 0{,}001$	3.717

Tabla 4.19: Modelo de regresión log-binomial (función COPY) para base de datos Constrict

Al igual que hemos hecho en los ejemplos anteriores, utilizaremos la prueba HL para comprobar el ajuste de bondad de los modelos. Como sucedía con la base de datos "Diet", todos los p-valores obtenidos son superiores a 0.05 indicando un ajuste satisfactorio, excepto en el caso del modelo COPY, para el que obtenemos resultados incorrectos.

4.4 Base de datos Can Ruti

EL objetivo principal de este estudio transversal realizado en el Hospital Germans Trias i Pujol en Badalona entre 1994 y 2004 [20], es analizar las transaminasas hepáticas en monoinfectados por el virus de la hepatitis C (VHC) y pacientes coinfectados por VHC/VIH para evaluar el efecto de la infección por VIH en la elevación de las enzimas hepáticas.

Para ello, 429 pacientes con antecedentes de consumo de drogas intravenosas (IDU) fueron admitidos a tratamiento por abuso de sustancias. Al inicio del estudio los pacientes depositaron muestras de sangre para el análisis de enzimas hepáticas, y se les tomaron datos sobre el colesterol, índice de masa corporal, consumo de alcohol, células $CD4^+$ y $CD8^+$ e infecciones VIH y VHC. En este caso, reduciremos el estudio analizando solamente los factores asociados con la elevación de los valores AST (aspartato aminotransferasa) y ALT (alanina transaminasa) en función de VIH^+ y VIH^-. Se crea una variable dicotómica que indica AST elevado cuando tiene valores > 35U/l o no elevado en caso contrario y otra para ALT elevado cuando es > 45U/l o no elevado para caso contrario. Tendremos un total de 175 pacientes en el grupo de VHC^+ & VIH^- y 215 en el de VHC^+ & VIH^+.

A continuación mostramos una tabla descriptiva de las variables empleadas para realizar el análisis.

	VHC$^+$ & VIH$^-$ N=175	VHC$^+$ & VIH$^+$ N=215
Sexo: Hombre	146 (83.4 %)	171 (79.5 %)
Edad	30.1 (6.12)	31.4 (5.62)
ALT elevado (> 45 U/l):		
Sí	100 (57.5 %)	104 (48.6 %)
AST elevado (> 35 U/l):		
Sí	80 (46.0 %)	117 (54.4 %)
células $CD4^+$	1236 (493)	447 (321)
células $CD8^+$	958 (406)	1039 (541)
IMC	20.7 (6.83)	18.4 (8.23)
Colesterol	163 (34.4)	147 (33.6)
Consumo de alcohol:		
Sí	38 (21.7 %)	65 (30.2 %)

Tabla 4.20: Tabla descriptiva de la base de datos can ruti en función de la variable VIH con n (%) para variables categóricas y media (sd) para numéricas

Al igual que los autores, utilizaremos el logaritmo (base 10) de el número de células $CD4^+$ y $CD8^+$, debido a su distribución asimétrica por la derecha.

4.4.1 Modelo de regresión logística

Como ya hemos comentado antes, realizamos el análisis mediante la función *"glm"*, dividiendo los datos en dos grupos, VHC$^+$ & VIH$^-$ y VHC$^+$ & VIH$^+$ y para los dos tipos de enzimas hepáticas, ALT y AST. En nuestro caso, decidimos coger como variables de interés sexo, alcohol y BMI, donde mujer y consumo de alcohol son las categorías de referencia. Como se trata de datos transversales, hablaremos de prevalencia y de riesgo. Obtenemos los siguientes resultados:

		VHC$^+$ & VIH$^-$			VHC$^+$ & VIH$^+$		
		$\widehat{\beta}$	p-valor	$\widehat{OR}$	$\widehat{\beta}$	p-valor	$\widehat{OR}$
ALT	Hombre	1.503	0.002	4.495	0.179	0.624	1.196
	Consumo de alcohol	0.402	0.342	1.495	0.352	0.262	1.422
	IMC	0.021	0.423	1.021	0.019	0.293	1.020
AST	Hombre	1.544	0.009	4.69	-0.696	0.062	0.499
	Consumo de alcohol	0.582	0.165	1.79	0.313	0.326	1.368
	IMC	0.007	0.809	1.01	0.024	0.196	1.024

Tabla 4.21: Modelo de regresión logística (función *"glm"*) para base de datos Can Ruti

Vemos que para ALT, en el caso de pacientes monoinfectados por VHC, los hombres tienen un riesgo bastante más elevado que las mujeres. Lo mismo sucede con el consumo de alcohol y el índice de masa corporal (IMC), ya que los pacientes que consumen y tienen un IMC más elevado tienen un riesgo más alto que los que no.

Para pacientes coinfectados las conclusiones son las mismas para las tres variables, aunque en este caso, el sexo no es tan elevado, aún siendo un factor de riesgo para la enfermedad. Sin embargo, hay que tener en cuenta que si usamos un $\alpha = 0{,}5$ como nivel de significación, en caso de $p > 0{,}05$ solo podemos decir que a nivel de la muestra vemos una asociación, pero no podemos decir que así es en la población.

Para los pacientes monoinfectados por VHC, en función de los valores AST, las tres variables son factores de riesgo para la enfermedad. Sin embargo, para pacientes coinfectados, los hombres tienen un menor riesgo que las mujeres para AST elevado (reducción en el riesgo de 50 %). Los indidivuos que consumen alcohol y tienen un IMC más alto, tienen un riesgo más elevado (37 % para los que consumen alcohol y 3 % para los de mayor IMC).

Por último comprobamos la bondad de ajuste mediante el test de Hosmer-Lemeshow. En todos los modelos ajustados obtenemos p-valores por encima de 0.1, indicando que son modelos con un ajuste satisfactorio.

4.4.2 Modelo log-binomial

Ajustaremos el modelo log-binomial para exactamente las mismas variables y veremos si obtenemos conclusiones similares o realmente en algunos de los casos se está exagerando la asociación de riesgo. Para ello, como hemos hecho en ejemplos anteriores, ajustamos mediante las tres funciones de R, viendo así si existen problemas de convergencia o no, y si fuese así, comprobando si las funciones *"logbin"* y COPY pueden solucionarlos. Es importante tener en cuenta que en este caso la medida de asociación será la tasa de prevalencia, debido a que se trata de un estudio transversal.

Función *"glm"*

Al ajustar el modelo log-binomial con esta función, nos devuelve que el algoritmo no converge para el caso de elevación ALT con VIH^- y para AST tanto con VIH^- como con VIH^+. Sin embargo, tal y como nos sucedía en el ejemplo Constrict, es capaz de devolver los resultados que presentamos a continuación:

		VHC$^+$ & VIH$^-$			VHC$^+$ & VIH$^+$		
		$\widehat{\beta}$	p-valor	$\widehat{PR}$	$\widehat{\beta}$	p-valor	$\widehat{PR}$
ALT	Hombre	0.837	0.012	2.309	0.167	0.41	1.182
	Consumo de alcohol	0.083	0.500	1.087	0.225	0.091	1.252
	IMC	0.013	0.254	1.013	0.013	0.184	1.013
AST	Hombre	1.129	0.016	3.093	-0.250	0.066	0.778
	Consumo de alcohol	0.175	0.278	1.191	0.091	0.463	1.095
	IMC	0.008	0.560	1.008	0.009	0.283	1.009

Tabla 4.22: Modelo de regresión log-binomial (función *"glm"*) para base de datos Can Ruti

Lo primero que habría que comprobar es que los resultados son correctos. En tal caso, vemos que las conclusiones son las mismas que para el modelo de regresión logística, salvo que la razón de prevalencia (PR) es bastante menos elevado para la variable sexo. Para el resto de variables, la razón de prevalencia es un poco más pequeña, sin llegar a existir diferencias destacables. Sabiendo que $1 < PR < OR$, estos resultados nos hacen pensar que a pesar de los problemas de convergencia los estimadores obtenidos son válidos.

Por último, asumiendo que las estimaciones son correctas, realizamos el test HL y comprobamos que todos los modelos se ajustan satisfactoriamente.

Función *"logbin"*

La función *"logbin"* nos sigue resultando bastante inestable. Aumentamos el número de iteraciones a 100000, y vamos probando valores $\epsilon \in \{0{,}1; 0{,}0001\}$. De esta manera, conseguimos que el algoritmo converja.

		VHC$^+$ & VIH$^+$			VHC$^+$ & VIH$^+$		
		$\widehat{\beta}$	p-valor	$\widehat{PR}$	$\widehat{\beta}$	p-valor	$\widehat{PR}$
ALT	Hombre	0.843	0.008	2.323	0.228	0.262	1.256
	Consumo de alcohol	0.086	0.526	1.09	0.149	0.296	1.160
	IMC	0.010	0.391	1.01	0.009	0.380	1.009
AST	Hombre	1.270	0.007	3.561	-0.191	0.178	0.826
	Consumo de alcohol	0.152	0.376	1.164	0.073	0.572	1.076
	IMC	0.015	0.309	1.015	0.004	0.582	1.004

Tabla 4.23: Modelo de regresión log-binomial (función *"logbin"*) para base de datos Can Ruti

Vemos que estos resultados difieren algo de los anteriores, aunque seguimos teniendo las mismas variables como posibles factores de riesgo para la enfermedad. Como en este caso el algoritmo converge y los resultados no difieren tanto a los obtenidos en la Tabla 4.22 se pueden considerar resultados válidos.

Por último comprobamos la bondad de ajuste con estos resultados y verificamos un ajuste de los modelos satisfactorio.

Función COPY

Por último, usamos el método COPY implementado en R y nos encontramos con que el método no consigue resolver la convergencia para el caso de elevación de AST con $n = 1000$. Debido a que los autores nos avisan que es posible que en ocasiones haya que cambiar el tamaño del valor n o los valores iniciales, probamos a cambiar la n [17]. Debemos recordar que este valor no tiene que ser ni demasiado grande ni demasiado pequeño como para que el nuevo estimador estén tan cerca del límite que el software no sea capaz de estimarlo.

Cambiando el valor de n a 100 en el caso de VIH^+ y a 10000 para VIH^- conseguimos que el algoritmo converja. En la Tabla 4.4.2 presentamos todos los resultados que el software R devuelve.

		VHC^+ & VIH^-			VHC^- & VIH^+		
		$\widehat{\beta}$	p-valor	$\widehat{PR}$	$\widehat{\beta}$	p-valor	$\widehat{PR}$
ALT	Hombre	0.835	0.012	2.305	0.166	0.412	1.181
	Consumo de alcohol	0.082	0.502	1.085	0.224	0.092	1.251
	IMC	0.013	0.255	1.013	0.013	0.186	1.013
AST	Hombre	1.125	0.016	3.08	-0.239	0.076	0.787
	Consumo de alcohol	0.174	0.280	1.19	0.085	0.488	1.089
	IMC	0.008	0.562	1.008	0.008	0.311	1.008

Tabla 4.24: Modelo de regresión log-binomial (función COPY) para base de datos Can Ruti

Los resultados apenas varían unas décimas de los conseguidos con la función *"glm"*. Sin embargo, seguimos sin saber si realmente son los resultados correctos que deberíamos obtener.

Por lo tanto, en este último ejemplo también hemos conseguido resultados satisfactorios. Por un lado hemos visto que la prevalencia no es tan exagerado como nos muestra el modelo logístico. Y por otro lado, comprobamos que para esta base de datos también funciona el método COPY.

4.5 Base de datos Can Ruti con interacciones

En esta sección comprobaremos como reaccionan estas 3 funciones ante un modelo con interacción. Es importante destacar que este análisis no está incluido por los autores, ya que no formaba parte del objetivo del estudio ni resultaba de interés clínico. Sin embargo, nosotros lo realizaremos para ver como reaccionan las 3 funciones presentadas cuando nos encontramos con modelos más complejos que incluyen interacciones. En este caso ajustaremos un modelo con las variables de la sección anterior incluyendo la interacción sexo y VIH. Por lo tanto, utilizaremos la base de datos completa sin dividir en pacientes con VIH^+ y VIH^-.

Para este caso también tendremos las variables elevación de AST y ALT como variable respuesta. Destacaremos que para la función *"glm"* de la regresión log-binomial, seguimos teniendo problemas de convergencia. Aún así conseguimos que devuelva resultados. Otro aspecto importante a destacar es que en la función *"logbin"* no es posible emplear iteraciones. Es por ello que los autores recomiendan incluir las interacciones mediante el cálculo de un nuevo término factor que tenga niveles que corresponden a todas las combinaciones posibles de los niveles del factor. Sin embargo, no hemos incluido en la tabla los resultados obtenidos mediante este método, ya que aún calculando la interacción aparte, no conseguimos que sean satisfactorios al igual que nos sucedía en todos los ejemplos anteriores. Es importante destacar también que para el método COPY en ambos caso (ALT y AST) con n=1000 hemos conseguido que el algoritmo converja. A continuación presentamos la tabla con los resultados:

		glm (logist.)		**glm (log-bin.)**		**COPY**	
		$\hat{\beta}$	$\hat{OR}$	$\hat{\beta}$	$\hat{PR}$	$\hat{\beta}$	$\hat{PR}$
ALT	Hombre	1.666	5.291	0.889	2.432	0.887	2.428
	HIV$^+$	0.902	2.466	0.553	1.738	0.551	1.35
	Consumo de alcohol	0.345	1.412	0.191	1.21	0.190	1.21
	IMC	0.019	1.020	0.011	1.01	0.011	1.01
	Hombre * VIH$^+$	-1.406	-	-0.678	-	-0.677	
AST	Hombre	1.703	5.493	1.190	3.287	1.186	3.274
	HIV$^+$	2.227	9.75	1.399	4.051	1.394	4.03
	Consumo de alcohol	0.372	1.450	0.163	1.177	0.163	1.177
	IMC	0.018	1.018	0.006	1.006	0.006	1.006
	Hombre * VIH$^+$	-2.283	-	-1.399	-	-1.394	-

Tabla 4.25: Modelo de regresión logístico y log-binomial para base de datos Can Ruti con interacción

Podemos decir que la prevalencia sigue sin ser tan elevada ni tan excesivamente pequeña como indica el modelo de regresión logística. En el caso de la interacción, tanto para elevación AST como ALT los hombres con VIH$^+$ tienen una menor prevalencia que las mujeres con VIH$^-$. Para comparar por ejemplo la razón de prevalencia en hombres con VIH$^+$ tendremos el siguiente modelo de regresión log-binomial:

$$\log P(Y = 1|\mathbf{X}) = \alpha + \beta_1 \cdot X_1 + +\beta_2 \cdot X_2 + \beta_3 \cdot X_1 \cdot X_2$$

donde X_1 es 0 para mujeres y 1 para hombres y X_2 es 0 para VIH$^-$ y 1 para VIH$^+$. Por tanto, la razón de prevalencia asociada a la infección por VIH si ambas personas son hombres para elevación ALT es:

$$\widehat{PR} = exp(1{,}666 - 1{,}406) = 1{,}297$$

Por último, comentar que seguimos obteniendo resultados muy similares para la función *"glm"* y el método COPY.

4.6 Resumen

En este último apartado resumiremos brevemente los resultados obtenidos de los ejemplos anteriormente expuestos. El objetivo de este trabajo era por un lado comparar el modelo de regresión logística con el modelo de regresión log-binomial, y por otro lado, comparar las funciones de R (tanto las ya existentes como el método implementado por nosotros COPY) para poder ajustar el modelo de regresión log-binomial. Por ello, dividimos este resumen en dos secciones, teniendo en cuenta estos dos objetivos principales.

4.6.1 Comparación modelo de regresión logística y log-binomial

Para poder realizar la comparación, lo primero que nos gustaría destacar es que se trata de dos modelos distintos, en el que cada uno tiene sus características, por lo que no son comparables en todos los aspectos.

Comenzaremos diciendo que el modelo de regresión logística sigue siendo una buena alternativa cuando no tenemos una variable de interés común en la población de estudio, ya que está muy desarrollado y es muy sencillo de implementar en cualquiera de los software estadísticos disponibles hoy en día. Sin embargo, cuando esto sucede hemos visto que el modelo de regresión log-binomial es una buena alternativa. En alguno de los casos, hemos obtenidos OR muy elevados, que nos indican un riesgo exageradamente alto. Pero cuando hemos ajustado mediante regresión log-binomial, comprobamos que ese riesgo se ha visto exagerado, y que en realidad no es tan elevado como el otro modelo nos muestra. Es por ello, que la regresión log-binomial es una buena alternativa cuando nos encontramos ante este problema siempre que sea un estudio de cohorte o transversal.

4.6.2 Comparación funciones de R

En esta sección nos centraremos en la comparación de los resultados obtenidos de las funciones de R para ajustar el modelo log-binomial. Mostramos los resultados en la tabla que presentamos a continuación. Para construirla hemos tenido en cuenta tanto si el algoritmo ha sido capaz de converger cuando existen problemas como si los resultados obtenidos son satisfactorios. Es decir, si una de las funciones no converge pero aún así ha conseguido devolver un resultados satisfactorio, como sucede en algunos ejemplos con la función *"glm"*, el símbolo en la tabla será ↯. Si por el contrario no es capaz ni de converger ni de obtener resultados correctos, utilizaremos el símbolo x. Y si los resultados son satisfactorios en ambos casos, el símbolo empleado será ✓.

	Funciones		
	glm	logbin	COPY
Diet	✓	↯	✓
Tabla 1 (Williamson et al. (2013)	x	↯	↯
Tabla 2 (Williamson et al. (2013)	↯	✓	✓
Constrict	↯	↯	✓
Can Ruti	↯	↯	✓

Tabla 4.26: Tabla resumen de funciones R para el modelo log-binomial

Vemos que para la función *"glm"* los resultados son satisfactorios cuando no tenemos problemas de convergencia. Sin embargo, cuando sí que existen estos problemas es capaz de devolver resultados, que en nuestros casos ha sido muy próximo a los obtenidos con la función COPY. Esto no significa que la función *"glm"* devuelva resultados satisfactorios siempre que no converja. Es posible que al ajustar modelos más complejos, con interacciones y con más variables explicativas ya no obtengamos resultados correctos. Pensamos que una de las razones por las que R es capaz de devolver un *output* es porque utiliza el método *step-halving*, explicado en el apartado 3.3. Como ya comentábamos, si la iteración trata de situarse fuera del espacio de parámetros la actualización se reduce a la mitad y recalcula el valor ajustado. Así se asegura que aunque no converja, los valores siempre se encuentren dentro del espacio de parámetros.

En el caso de la función *"logbin"*, no hemos conseguido resultados completamente satisfactorios. Esta función sí que consigue que R no devuelva errores de convergencia, sin embargo, en muchas ocasiones no es capaz de devolver resultados completos, ya que solamente obtenemos los valores de los estimadores. Aún cambiando los argumentos de la función, no conseguimos resultados que se aproximen a los reales. Es posible que al tratarse de una función relativamente reciente, todavía se encuentre en proceso de prueba y le queden varios aspectos que mejorar.

Por último, para la función COPY hemos obtenido resultados satisfactorios en todos los casos menos en el de la Tabla 1 de Williamson et al. (2013). Recordemos que para este ejemplo con ninguna de las 3 funciones obtenemos resultados satisfactorios, seguramente debidos al 0 de la tabla. Por lo tanto, al menos para estas bases de datos, podemos decir que el método ha funcionado en un 80 % de los casos. Es por ello que recomendamos utilizar la función *"glm"* cuando no tengamos problemas de convergencia o COPY en caso contrario.

Capítulo 5

Conclusiones y discusión

En este trabajo hemos presentado el modelo de regresión log-binomial como alternativa al modelo de regresión logística. Hemos podido comprobar que este modelo es de gran utilidad cuando queremos estimar la razón de prevalencias o el riesgo relativo en un estudio con resultados comunes en la población. Como ya comentabamos al inicio, hoy en día apenas encontramos información sobre la regresión log-binomial. Es por ello que su implementación no está lo suficientemente desarrollada en ninguno de los softwares estadísticos. El objetivo principal de este trabajo ha sido por un lado dar a conocer estos modelos reuniendo toda la información posible sobre ellos y por otro implementarlos en el software R comentando los problemas con los que nos podemos encontrar.

A la hora de implementarlos puede suceder que el software no consiga encontrar el estimador de máxima verosimilitud (MLE) creando problemas de convergencia. Estos problemas suceden cuando el MLE se encuentra en el límite del espacio de parámetros restringido. Para estos casos hemos presentado por un lado el método desarrollado por Deddens et al. (2003) [16] en el software estadístico SAS, el cuál consiste simplemente en modificar la base de datos mediante n copias para obtener un MLE muy próximo al real, consiguiendo que el algoritmo converja. Y por otro lado la función *"logbin"* recientemente implementada en R por Mark W. Donoghoe (2015) [24]. Debido a que el primer método no estaba implementado en R otra de las propuestas de este trabajo ha sido implementar el método COPY en este software consiguiendo resultados satisfactorios. Para todos los ejemplos expuestos logramos que el algoritmo converja. Un detalle a tener en cuenta es que en ocasiones es posible que el número de copias varíe dependiendo de la base de datos que tengamos, por lo que es recomendable probar con distintos valores de n para conseguir el ajuste más adecuado. También puede haber casos en los que los valores iniciales no sean los adecuados y el usuario tenga que cambiarlos.

Hemos expuesto varios ejemplos comparando tanto el modelo de regresión logística como el modelo de regresión log-binomial mediante las funciones *"glm"*, *"logbin"* y COPY.

Destacamos que cuando no tenemos problemas de convergencia no es necesario emplear ninguna de estas dos últimas. Sin embargo, cuando sí los tenemos, hemos comprobado que obtenemos resultados más satisfactorios para el método COPY. La función implementada por M. W. Donoghoe produce resultados bastante inestables y es necesario añadir distintos argumentos para que sea capaz de devolver resultados. No hemos conseguido saber cuáles son los valores de la tolerancia de los límites ni los valores de la tolerancia de convergencia positiva que se han de escoger para ajustar el modelo correctamente. Además, no permite emplear interacciones directamente en su formulación; es necesario definirlas con anterioridad. Es posible que debido a su reciente implementación se encuentre todavía en periodo de prueba, por lo que a lo largo del tiempo consiga mejorar sus resultados.

Para realizar la bondad de ajuste de los modelos, hemos utilizado el test de Hosmer-Lemeshow comprobando si el p-valor es superior a 0.05. Tanto para el modelo de regresión logística como para el log-binomial es posible emplear este método, aunque en ocasiones no resulte demasiado fiable o no sirva cuando tenemos pocos datos y muchos empates en las probabilidades predichas.

Sin embargo, queda todavía mucho que investigar acerca del modelo log-binomial. Una de las principales dudas que nos quedan después de realizar este trabajo, es por qué no es posible utilizar el test de Hosmer-Lemeshow cuando empleamos el método COPY. Convendría encontrar un método que nos asegurase que el modelo que estamos ajustando es correcto. Por otro lado, hemos visto que cuando tenemos una base de datos con el número total de casos igual a 0 para alguno de los grupos, es posible utilizar la función "logistf" en los modelos de regresión logística para conseguir que nos devuelva estimaciones estables. Sin embargo, no hemos investigado si es posible implementar esta misma función para el modelo log-binomial, mediante el método Firth.

Por último, hemos observado que el software estadístico R, aún teniendo problemas de convergencia, es capaz de devolver unos resultados que en varios de los ejemplos se considerarían correctos. Podría ser que debido a que R utiliza el método denominado *step-halving*, que recordemos reduce a la mitad el número de pasos y recalcula el valor ajustado cuando la iteración se sitúa fuera del espacio de parámetros restringido, asegurándose que los valores siempre se encuentren dentro del espacio. Sin embargo, no significa que siempre vayamos a obtener resultados correctos, ya que recordemos que en este trabajo se revisan ejemplos simples, sin realizar simulaciones, por lo que no son resultados concluyentes que aseguren que los métodos utilizados resulte satisfactorio en todos los casos para este software.

Creemos que todavía queda mucha nueva información que aportar sobre el modelo de regresión log-binomial. Sin embargo, cuando se tenga tanto conocimiento sobre el como lo tenemos hoy en día sobre el modelo de regresión logística, ofrecerá una buena solución para obtener estimaciones del riesgo relativo en un estudio de cohorte y la razón de prevalencia en un estudio transversal.

Bibliografía

[1] Apuntes de la asignatura de Epidemiología del Máster en Estadística e Investigación Operativa 2015, basados en buena parte en Jewell, 2004.

[2] N. P. Jewell, *Statistics for epidemiology* Boca Raton: Chapman & Hall/CRC, cop. 2004.

[3] C. J. Clopper, E. S. Pearson. *The use of confidence or ducial limits illustrated in the case of the binomial.* Biometrika, 1934.

[4] K. J. Rothman, *Epidemiology : an Introduction.* Oxford University Press, 2002.

[5] K. J. Rothman, S. Greenland, and T. L. Lash, *Modern epidemiology.* Philadelphia: Lippincott Williams & Wilkins, cop. 2008.

[6] L. A. McNutt, C. Wu, X. Xue, and J. P. Hafner, *Estimating the Relative Risk in Cohort Studies and Clinical Trials of Common Outcomes.* American Journal of Epidemiology, 2003.

[7] A. J. D. Barros and V. N. Hirakata, *Alternatives for logistic regression in cross-sectional studies: an empirical comparison of models that directly estimate the prevalence ratio.* BMC Medical Research Methodology, 2003.

[8] M.J. Symons, D.T. Moore, *Hazard rate ratio and prospective epidemiological studies.* Journal of Clinical Epidemiology 55 893–899, 2002.

[9] D. W. Hosmer, T. Hosmer, S. Le Cessie, S. Lemeshow, *A comparison of goodness-of-fit tests for the logistic regression model.* Statistics in medicine, 1997.

[10] D. W. Hosmer, S. Lemeshow, *Applied Logistic Regression.* Second Edition. Wiley, New York, 2000.

[11] J. Zhang, K. Yu *What's relative risk? A method of correcting the odds ratio in cohort studies of common outcomes.* Journal of American Medical Association, 1998.

[12] A. Savu, Q. Liu, Y. Yasui, *Estimation of relative risk and prevalence ratio.* Statistics in Medicine, 2010.

[13] J. A. Deddens, M. R. Petersen, *Approaches for estimating prevalence ratios*. BMJ, 2008.

[14] L. Blizzard, D.W. Hosmer, *Parameter Estimation and Goodness-of-fit in Log Binomial Regression*. Biometrical Journal, 2006.

[15] T Williamson, M. Eliasziw and G. H. Fick, *Log-binomial models: exploring failed convergence*. Emerging Themes in Epidemiology, 2013.

[16] J. A. Deddens, M. R. Petersen, X. Lei, *Estimation of prevalence ratios when PROC GENMOD does not converge*. Statistics and Data Analysis, 2003.

[17] M. R. Petersen and J. A. Deddens, *A revised SAS macro for maximum likelihood estimation of prevalence ratios using the COPY method*. Occupational and Environmental Medicine, 2009.

[18] T. Lumley, R. Kornmal, S. Ma, *Relative risk regression in medical research: models, contrasts, estimators, and algorithms*. UW Biostatistics Working Paper, 2006.

[19] D. Firth, *Bias reduction of maximum likelihood estimates*. Biometrika, 1993.

[20] K. Langohr, A. Sanvisens, D. Fuster, J.Tor, I. Serra, C. Rey-Joly, I. Rivas, R. Muga *Liver Enzyme Alterations in HCV-Monoinfected and HCV/HIV-Coinfected Patients*. The Open AIDS Journal, 2008.

[21] *John Snow, la epidemia de cólera y el nacimiento de la epidemiología moderna.* http://www.ph.ucla.edu/epi/snow/revchilenainfectol24(4)_331_4_2007.pdf

[22] *Diseño de estudios epidemiológicos* http://www.scielosp.org/pdf/spm/v42n2/2383.pdf

[23] *The LOGISTIC Procedure* http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm#statug_logistic_sect064.htm

[24] M. W. Donoghoe (2015). *logbin: Relative Risk Regression Using the Log-Binomial Model. R package version 1.2.* http://CRAN.R-project.org/package=logbin

[25] T. J. Aragon Developer (2012). *epitools: Epidemiology Tools. R package version 0.5-7.* http://CRAN.R-project.org/package=epitools

[26] B. Carstensen, M. Plummer, E. Laara, M. Hills (2016). *Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.0.* http://CRAN.R-project.org/package=Epi

[27] G. Heinze, M. Ploner, D. Dunkler and H. Southworth (2013). *logistf: Firth's bias reduced logistic regression. R package version 1.21.* http://CRAN.R-project.org/package=logistf

Apéndice A

Diseño de estudios epidemiológicos

En epidemiología podemos encontrar distintos tipos de estudios dependiendo de la asignación de la exposición o variable de interés, el número de mediciones que se realiza en el individuo, es decir, si solamente tenemos una medida o varias tomadas a lo largo del tiempo, los criterios utilizados a la hora de seleccionar la población de estudio y la unidad de análisis donde se mide el evento en estudio. Podemos clasificar los estudios en transversales, cohorte, caso-control, aleatorizados y ecológicos. Sin embargo, en este trabajo solamente definiremos brevemente los más utilizados en epidemiología, es decir, los primeros tres estudios nombrados.

A.1 Estudios de cohorte

Estos estudios siguen un criterio de selección basado en la exposición de interés, es decir, se toman las muestras entre los individuos expuestos y no expuestos libres de la enfermedad. Se trata de estudios longitudinales, donde se usan medidas del individuo tomadas en distintos intervalos de tiempo.

Entre las ventajas de estos estudios se encuentra que se puede comparar los riesgos $P(D|E)$ y $P(D|\hat{E})$ y se puede estimar la tasa de incidencia, permite estudiar cuando la exposición no es común y es posible analizar varias enfermedades a la vez.

Entre las desventajas encontramos que son estudios duraderos y por ello de alto coste, existe la posibilidad de pérdida de seguimiento, requiere muestras muy grandes en caso de enfermedad rara y la exposición puede cambiar a lo largo del tiempo. Otra de las limitaciones es que la asignación de la exposición no es de manera aleatoria, por lo que no existe la posibilidad de controlar las posibles diferencias entre los grupos de expuestos y no expuestos en relación con otros factores asociados con la ocurrencia del evento.

A.2 Estudios transversales

En estos estudios se toma la muestra de la población de interés. Al contrario que en los estudios de cohorte, se toma una sola medida del individuo en un tiempo determinado y se determina la presencia o ausencia tanto de la exposición y como de la enfermedad.

Las ventajas más destacables son que podemos estimar tanto $P(D)$, como $P(E)$ y $P(D \cap E)$, que son unos estudios cortos y de bajo coste ya que no se ha de hacer un seguimiento, y que pueden servir para generar hipótesis. Sin embargo, tienen varias limitaciones como que no es posible saber si la enfermedad precede a la exposición o sucede del revés, por lo que no podemos establecer ninguna relación causal. Otra desventaja es que no podemos estimar la incidencia y por último que el investigador no controla la proporción de individuos bajo exposición o enfermedad.

A.3 Estudios caso-control

Son estudios que siguen un criterio de selección basado en la enfermedad de interés, es decir, las muestras son tomadas entre los individuos afectados y libres de la enfermedad, respectivamente. Se trata de estudios retrospectivos, es decir, las medidas se toman desde un tiempo cero hacia el pasado. Estos estudios son muy apropiados cuando tenemos enfermedades raras. Es posible estudiar varios factores de riesgo y tienen un bajo coste. Entre las limitaciones se encuentra que no es posible estimar $P(D|E)$, que no se sabe realmente si E precede a D, por lo que es difícil determinar exactamente la exposición, y es posible que se cree un sesgo de selección.

Apéndice B

Código de R

```
#########################################################
########## IMPLEMENTACIÓN DEL MÉTODO COPY ############
##########  Y EJEMPLOS CON DATOS REALES   ############
#########################################################

#-----------------------------
#________Paquetes____________
#-----------------------------

install.packages('Epi')
library(Epi)
#_________________________
install.packages('logbin')
library(logbin)
#_________________________
install.packages('logistf')
library(logistf)
#_________________________
install.packages('epitools')
library(epitools)
#_________________________
install.packages('Hmisc')
library(Hmisc)
#_________________________
install.packages('compareGroups')
library(compareGroups)
#_________________________
install.packages("vcdExtra")
library(vcdExtra)

########################################
##--------Base de datos Diet----------##
########################################

?diet
data(diet)
summary(diet)
head(diet)
```

```
#cálculo del IMC
diet$height<- (diet$height)/100
diet$IMC<- (diet$weight)/(diet$height)^2

#Descripción de la base de datos
export2latex(createTable(compareGroups(chd~., data=diet)))

#Regresión logística con "glm"
diet.logit <- glm(chd~energy.grp+fat+fibre+IMC, diet,
  family=binomial)
summary(diet.logit)
AIC(diet.logit)

#Regresión log-binomial con "glm"
diet.logbinom <- glm(chd~energy.grp+fat+fibre+IMC, diet,
 family=binomial(log))
summary(diet.logbinom)
AIC(diet.logbinom)

##Regresión log-binomial con "logbin"
diet.logbin <- logbin(chd~energy.grp+fat+fibre+IMC, data=diet,
  maxit=100000, bound.tol=1e-10)
summary(diet.logbin)

#Regresión log-binomial con "COPY"

##100 copias
diet.copy100 <- copy(diet, 'chd', vars = c('energy.grp', 'fat',
'fibre', 'IMC'), 100)
summary(diet.copy100)

##1000 copias
diet.copy1000 <- copy(diet, 'chd', vars = c('energy.grp', 'fat',
         'fibre', 'IMC'), 1000)
summary(diet.copy1000)


############################################################
##--------Tablas paper Williamson et.al (2003)----------##
############################################################

#----------------------
#______tabla 1_________
#----------------------

(tabla1 <- matrix(c(10, 8, 18, 9, 5, 0), nr=2,
                  dimnames=list(D=c('Disease', 'No disease'),
                  E=c('X=-1', 'X=0', 'X=1'))))

tabla1 <- expand.table(tabla1)
summary(tabla1)

#Regresión logística con "glm"
```

```
tabla1.logit <- glm(D~E, tabla1, family=binomial)
summary(tabla1.logit)

#Regresión logística con "logistf"
tabla1.logistf <- logistf(D~E, tabla1, family=binomial)
summary(tabla1.logistf)

#Regresión log-binomial "glm"
tabla1.logbinom <- glm(D~E, tabla1, family=binomial(log))
summary(tabla1.logbinom)

#Regresión log-binomial "logbin"
tabla1.logbin <- logbin(as.numeric(D)-1~E, data=tabla1, epsilon=5e-04)
summary(tabla1.logbin)

#Regresión log-binomial con "COPY"

##100 copias
tabla1.copy100 <- copy(tabla1, 'D', vars = 'E', 100)
summary(tabla1.copy100)

##1000 copias
tabla1.copy1000 <- copy(tabla1, 'D', vars = 'E', 1000)
summary(tabla1.copy1000)

#----------------------
#______tabla 3_________
#----------------------

(tabla3 <- matrix(c(2, 2, 14, 3, 2, 17), nr=2,
                  dimnames=list(D=c('Disease', 'No disease'),
                  E=c('X=-1', 'X=0', 'X=1'))))
tabla3 <- expand.table(tabla3)#tabla
summary(tabla3)

#Regresión logística con "glm"
tabla3.logit <- glm(D~E, tabla3, family=binomial)
summary(tabla3.logit)

#Regresión log-binomial con "glm"
tabla3.logbinom <- glm(D~E, tabla3, family=binomial(log),
                       start=c(-4,rep(1e-4, 2)))
summary(tabla3.logbinom)

#Regresión log-binomial con "logbin"
tabla3.logbin <- logbin(as.numeric(D)-1~E, data=tabla3)
summary(tabla3.logbin)

#Regresión log-binomial con "COPY"

##100 copias
tabla3.copy100 <- copy(tabla3, 'D', vars = 'E', 100)
summary(tabla3.copy100)
```

```
##1000 copias
tabla3.copy1000 <- copy(tabla3, 'D', vars = 'E', 1000)
summary(tabla3.copy1000)


###############################################
##--------Base de datos Constrict----------##
###############################################

constrict <- read.table("constrict.txt", header=T)

#Transformación logarítmica
constrict <- transform(constrict, logV=log(Volume), logR=log(Rate))

#Regresión logística con "glm"
cons.logit <- glm(Cons~log(Volume)+ log(Rate), data=constrict,
                  family=binomial)
summary(cons.logit)

#Regresión log-binomial con "glm"
cons.logbinom <- glm(Cons~log(Volume)+ log(Rate), data=constrict,
                     family=binomial(log), start=c(-4,rep(0, 2)))
summary(cons.logbinom)

#Regresión log-binomial con "logbin"
cons.logbin <- logbin(Cons~log(Volume)+ log(Rate), data=constrict,
                      epsilon=5e-5)
summary(cons.logbin)

#Regresión log-binomial con "COPY"

#100 copias
copy(constrict, "Cons", vars= c("logV", "logR"), 100)

#1000 copias
copy(constrict, "Cons", vars= c("logV", "logR"), 1000)


###############################################
##--------Base de datos Can Ruti-----------##
###############################################

install.packages("compareGroups")
library(compareGroups)
library(Hmisc)
library(Epi)

Label(canrut)

#Descripción de la base de datos
export2latex(createTable(compareGroups(hiv~., data=canrut)))

canruti1 <- subset(canrut, hiv=="HCV+ & HIV+")
canruti2 <- subset(canrut, hiv=="HCV+ & HIV-")
```

```
#regresión logística

#------ALT---------

#HIV+
canruti.logit1 <- glm(altHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti1, family=binomial)
summary(canruti.logit1)
print(ci.lin(canruti.logit1, Exp=T), digits=3)

#HIV-
canruti.logit2 <- glm(altHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti2, family=binomial)
summary(canruti.logit2)
print(ci.lin(canruti.logit2, Exp=T), digits=3)


#--------AST -----------------

#HIV+
canruti.logit3 <- glm(astHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti1, family=binomial)
summary(canruti.logit3)
print(ci.lin(canruti.logit3, Exp=T), digits=3)


#HIV-
canruti.logit4 <- glm(astHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti2, family=binomial)
summary(canruti.logit4)
print(ci.lin(canruti.logit4, Exp=T), digits=3)

#-----HL TEST---------

HosmerLemeshow(canruti.logit1)
HosmerLemeshow(canruti.logit2)
HosmerLemeshow(canruti.logit3)
HosmerLemeshow(canruti.logit4)


#regresión log-binomial con "glm"

#------ALT---------

#HIV+
canruti.logbinom1 <- glm(altHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti1, family=binomial(log), start=c(-4,rep(1e-4, 6)))
summary(canruti.logbinom1)

#HIV-
canruti.logbinom2 <- glm(altHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti2, family=binomial(log), start=c(-4,rep(1e-4, 6)))
summary(canruti.logbinom2)
```

```
#--------AST -----------------

#HIV+
canruti.logbinom3 <- glm(astHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti1, family=binomial(log), start=c(-4,rep(1e-4, 6)))
summary(canruti.logbinom3)

#HIV-
canruti.logbinom4 <- glm(astHigh~sex+logcd8+alcol+bmi+colest+logcd4,
canruti2, family=binomial(log), start=c(-4,rep(1e-4, 6)))
summary(canruti.logbinom4)

#-------HL TEST-----------
HosmerLemeshow(canruti.logbinom1)
HosmerLemeshow(canruti.logbinom2)
HosmerLemeshow(canruti.logbinom3)
HosmerLemeshow(canruti.logbinom4)

#regresión log-binomial con "log-bin"

canruti.logbin1 <- logbin(as.numeric(altHigh)-1~sex+logcd8+alcol+bmi
+colest+logcd4, data=canruti1, maxit=100000, epsilon=1e-2)
summary(canruti.logbin1)

canruti.logbin2 <- logbin(as.numeric(altHigh)-1~sex+logcd8+alcol+bmi
+colest+logcd4, data=canruti2, maxit=100000, epsilon=1e-1)
summary(canruti.logbin2)

canruti.logbin3 <- logbin(as.numeric(astHigh)-1~sex+logcd8+alcol+bmi
+colest+logcd4, data=canruti1, maxit=100000, epsilon=1e-4)
summary(canruti.logbin3)

canruti.logbin4 <- logbin(as.numeric(astHigh)-1~sex+logcd8+alcol+bmi
+colest+logcd4, data=canruti2, maxit=100000, epsilon=1e-2)
summary(canruti.logbin4)

#regresión log-binomial con "COPY"

canruti.copy1 <- copy(canruti1, 'altHigh',
vars = c('sex', 'logcd8', 'alcol', 'bmi','colest', 'logcd4'), 1000)
summary(canruti.copy1)

canruti.copy2 <- copy(canruti2, 'altHigh',
vars = c('sex', 'logcd8', 'alcol', 'bmi', 'colest', 'logcd4'), 1000)
summary(canruti.copy2)

canruti.copy3 <- copy(canruti1, 'astHigh',
vars = c('sex', 'logcd8', 'alcol','bmi', 'colest', 'logcd4'), 1000)
summary(canruti.copy3)

canruti.copy4 <- copy(canruti2, 'astHigh',
vars = c('sex', 'logcd8', 'alcol', 'bmi', 'colest', 'logcd4'), 1000)
summary(canruti.copy4)
```

```
canruti.copy33 <- copy(canruti1, 'astHigh',
vars = c('sex', 'logcd8', 'alcol','bmi', 'colest', 'logcd4'), 100)
summary(canruti.copy33)

canruti.copy44 <- copy(canruti2, 'astHigh',
vars = c('sex', 'logcd8', 'alcol', 'bmi', 'colest', 'logcd4'), 10000)
summary(canruti.copy44)

#####################################
## CAN RUTI CON INTERACCIONES #######
#####################################

#regresión logistica
canruti.logit1.int <- glm(astHigh~sex+hiv+logcd8+alcol+bmi+colest+logcd4
+sex*hiv, canrut, family=binomial)
summary(canruti.logit1.int)
print(ci.lin(canruti.logit1.int, Exp=T), digits=3)

canruti.logit2.int <- glm(altHigh~sex+hiv+logcd8+alcol+bmi+colest+logcd4
+sex*hiv, canrut, family=binomial)
summary(canruti.logit2.int)
print(ci.lin(canruti.logit2.int, Exp=T), digits=3)

#regresión log-binomial ``glm''
canruti.logbinom1.int <- glm(astHigh~sex+hiv+logcd8+alcol+bmi+colest+logcd4
+hiv*sex, canrut, family=binomial(log), start=c(-4,rep(1e-4, 8)))
summary(canruti.logbinom1.int)

canruti.logbinom2.int <- glm(altHigh~sex+hiv+logcd8+alcol+bmi+colest+logcd4
+hiv*sex, canrut,family=binomial(log), start=c(-4,rep(1e-4, 8)))
summary(canruti.logbinom2.int)

#regresión log-binomial ``logbin''
hivsex <-as.numeric(canrut$hiv)-1+as.numeric(canrut$sex)+1
canruti.logbin1.int <- logbin(as.numeric(altHigh)-1~sex+hiv+logcd8+alcol+bmi
+colest+logcd4+hivsex, data=canrut, maxit=100000, epsilon=1e-1)
summary(canruti.logbin1.int)

canruti.logbin2.int <- logbin(as.numeric(astHigh)-1~sex+hiv+logcd8+alcol+bmi
+colest+logcd4+hivsex, data=canrut, maxit=100000, epsilon=1e-2)
summary(canruti.logbin2.int)

#regresión log-binomial COPY
canruti.copy1.int <- copy(canrut, 'astHigh', vars = c('sex','hiv',
 'logcd8', 'alcol','bmi','colest', 'logcd4', 'hiv*sex'), 1000)
summary(canruti.copy1.int)

canruti.copy2.int <- copy(canrut, 'altHigh', vars = c('sex','hiv',
 'logcd8', 'alcol','bmi','colest', 'logcd4', 'hiv*sex'), 1000)
summary(canruti.copy2.int)
```