Estadística Española • Volumen 59, número 194 / 2017, pp. 151-165

Detección de parejas con valores potencialmente influyentes en el modelo final de un estudio caso-control emparejado 1:1. Una nueva aproximación básica

Miguel Ángel Castro-Jiménez[1]

María Ximena Meneses[2]

Resumen

Este manuscrito propone una metodología sencilla para la detección de pares potencialmente influyentes en los modelos finales de estudios epidemiológicos de casos y controles emparejados con razón 1:1. La metodología se basa en la resolución de la ecuación de las variables independientes del modelo final para cada pareja (estrato) caso: control. Se realiza una comparación con un método estadístico propuesto por otros autores. Se explica su ejecución manual y en el programa estadístico R.

Palabras clave: Aplicaciones a biología y ciencias médicas, comparación emparejada y múltiple, diagnóstico.

Clasificación AMS: 46N30, 62J15, 62J20

Detecting influential pairs in the final model of a 1-1 matched case-control study: A new basic approach

Abstract

This manuscript proposes a simple methodology for the detection of potentially influential pairs in the final models of 1:1 matched case- control studies. The methodology is based on the solving the equation of the independent variables from

[1] Grupo de Investigación e Innovación en Salud. Centro de Investigación en Salud. Colsubsidio. Bogotá, D.C., Colombia
Subdirección de Vigilancia en Salud Pública. Dirección de Epidemiología, Análisis y Gestión de Políticas de Salud Colectiva. Subsecretaría de Salud Pública. Secretaría Distrital de Salud. Bogotá, D.C., Colombia
Grupo Colombiano de Estudios Alfa en Epidemiología, Salud Poblacional, Estadística Aplicada y Ciencias Aliadas. Magna Science Corporation. Bogotá, D.C., Colombia
[2] Grupo Área Análisis de Datos. Instituto Nacional de Cancerología. Bogotá, D.C., Colombia

the final model for each pair (strata) case: control. We explained each step and it is explained in R.

Keywords: Applications to biology and medical sciences, paired and multiple comparison, diagnostic.

AMS Classification: 62P10, 62J15, 62J20

1. Introducción

La utilidad real de cualquier modelo estadístico que pretenda explicar la aparición de una enfermedad depende, entre otros aspectos, de la implementación de estrategias de control que conduzcan a la minimización de errores durante la recolección y la digitación de la información, y también del cumplimiento de algunas premisas necesarias para la escogencia de las variables del modelo final (Greenland 1989, Rothman, 1981; Stevens, 1984). Después de establecer los factores independientes que componen el mejor modelo final, es necesario determinar si este modelo representa de manera adecuada los datos analizados y detectar aquellas observaciones que se alejan del comportamiento general del grupo y, por tanto, podrían estar influyendo en los resultados obtenidos (Hosmer, Taber y Lemeshow S, 1991; Cook RD, 1977).

La regresión logística condicional (RLC) es un método estadístico alternativo al modelo clásico (no condicional) que puede ser utilizado en el análisis de estudios de casos y controles emparejados (Breslow y Day, 1980; Ekstrom, 2017), siendo éste un tipo de estudio epidemiológico que generalmente tiene una variable dependiente binaria y en el que la selección de un sujeto control depende de su similitud al caso en las características de emparejamiento previamente establecidas (por ejemplo, institución, vecindario, edad y sexo).

Este artículo describe un método estadístico básico que podría ser usado para identificar aquellas parejas caso-control que no se comportan como el resto de parejas del estudio cuando se utiliza como diseño un estudio de casos y controles emparejados (1:1). Se utiliza, para este artículo, información de un estudio dirigido a detectar factores de riesgo para la aparición de leucemia linfoblástica aguda en niños menores de 15 años en Colombia (Base: CACO) (Castro-Jiménez y Orozco-Vargas, 2011). En ese estudio se decidió utilizar controles seleccionados al azar del mismo vecindario del menor caso, quienes, además de estar emparejados por las exposiciones ambientales o socioeconómicas del momento del diagnóstico, también debían cumplir criterios de edad y sexo.

Como con otros análisis de regresión, el uso de técnicas estadísticas para evaluar la bondad de ajuste de los modelos obtenidos con regresión logística condicional es una parte complementaria del análisis que debe realizarse antes de publicar los resultados de una investigación. Un análisis complementario a la bondad de ajuste es la identificación de los grupos (estratos) que no se ajustan bien luego de obtener un modelo de datos caso-control emparejado. Una de las metodologías usadas, y punto de comparación para la técnica propuesta en este trabajo, ha sido descrita por Hosmer (Hosmer y Lemeshow, 1989).

2. Desarrollo matemático

Se propone un nuevo método que se basa en los coeficientes calculados del modelo de regresión logística condicional con un emparejamiento 1 caso: 1 control. La base de datos debe reflejar los siguientes supuestos:

- La variable de salida (dependiente) identifica al sujeto caso y control de cada pareja y deben estar codificadas como 1 para caso y 0 para control.
- Debe existir una variable que identifique cual caso y control fueron emparejados, es decir, que debe existir un identificador duplicado.
- Las variables independientes (factores de riesgo o protectores en el caso de enfermedades) pueden estar medidas en cualquier escala.

Luego de obtener el modelo final y cumplir los anteriores supuestos, los pasos a seguir para llevar a cabo el procedimiento son:

1. Partición de la base de datos en dos bases que contengan en la primera base la información de los casos y la segunda base de datos la información de los controles. El orden de la partición no es importante, pero debe crearle un sufijo para sea fácil la identificación del origen de la variable. Por ejemplo, se tiene una variable llamada tabaco, cuando se realice la partición en la base de controles esta misma variable se podría llamar tabaco_co y en la base de datos de casos tabaco_ca.
2. Es necesario realizar el cálculo para cada uno de los registros (sujetos) ingresados del valor predicho por el modelo final, para este cálculo se usa la siguiente formula:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

   La matriz de datos toma la siguiente forma:

$$y = \beta_0 + \beta_p x_{pi}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \beta_0 + \beta_p \begin{bmatrix} 0 & 1 & 0 & \dots & p \\ 1 & 0 & 0 & \dots & p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n & n & n & n & pn \end{bmatrix}$$

3. Dado el valor predicho para cada uno de los registros (sujetos) de las bases de datos, una para los controles y otra para los casos, se procede a anexar los datos de la pareja de manera horizontal, es decir, al valor predicho del caso 1 se añade de manera horizontal el correspondiente a su pareja control. Por esta razón es importante tener los identificadores duplicados y los prefijos para cada una valor predicho que contenga la base de datos.
4. Se calculan las diferencias entre el valor predicho para el caso y para el control, así:

$$Difcaco_i = y_i caso - y_i control$$

En donde $Difcaco_i$ representa el valor de la diferencia caso-control.

5. Ordenar las diferencias de menor a mayor y se decide el punto de corte para seleccionar las parejas potencialmente influyentes, tanto como las mayores negativas como las mayores positivas. Las parejas encontradas serán las potencialmente influyentes observadas dentro del conjunto de datos analizado.

3. Desarrollo computacional

Todo el desarrollo computacional se hará en el programa estadístico libre R, esta programación está compuesta de la siguiente manera:

1) La primera función que aparece es la de borrar el historial del programa R.

2) Aparecen las librerías necesarias:

   - ***Foreign:*** Este paquete funciona para leer y escribir datos almacenados en Epi Info, Minitab, S, SAS, SPSS, Stata, Systat y Weka.
   - ***Epi:*** Este paquete funciona para el análisis demográfico y epidemiológico en el diagrama de Lexis, es decir, registra y da seguimiento a datos tipo cohorte, representa y manipula los datos de varios estadíos. También contiene funciones para el modelamiento de edad-período-cohorte y una función de datos de intervalos censurados y algunas funciones útiles para la tabulación y el trazado, así como algunos conjuntos de datos epidemiológicos.
   - ***library (car):*** Este paquete proporciona muchas funciones para usar el modelo de regresión ajustada, y realizar cálculos adicionales en el modelo o si es posible calcular un modelo diferente, y luego devuelve los valores y gráficos.

3) Verificar el directorio donde se tiene la base de datos en la cual se desea hacer el análisis.

4) Verifica los archivos que hay dentro del directorio anteriormente utilizado, con esta función rectifique que la base de datos que quiere leer le aparezca en el visor de salidas de R.

5) Lea la base de datos, use el lector dependiendo en que software tenga guardada la base de datos a usar. Dentro de los lectores más usados están:

   - **Excel:** Si sus datos estas guardados en Excel, es más sencillo si los guarda en formato .csv (delimitado por comas) y de esta manera con la siguiente función procede a la lectura, tenga en cuenta que el campo *sep* de la función es donde especifica el tipo de separador con que viene predeterminado el guardado .csv y de igual manera en el campo dec el cual hace referencia a la delimitación de los decimales.

     read.csv("NOMBREDELABASEDEDATOS.csv",sep=";",header=T,dec=",")

   - **STATA:** Si sus datos los tiene guardados en STATA desde la versión 5 a la versión 12, con la siguiente función los puede leer fácilmente.

     read.dta("NOMBREDELABASEDEDATOS.dta")

- **SPSS:** Si sus datos los tiene guardados en SPSS, use el siguiente conjunto de funciones para que le lea la base correctamente.

  read.spss("NOMBREDELABASEDEDATOS.sav") y luego corra *as.data.frame(BD)*.

6) Verifique los nombres de las variables que va a usar en el modelo de regresión logística condicional.

   Recuerde que en general los modelos en R son puestos dentro de la función así:

$$Y \sim X_1 + X_2 + \cdots + X_n$$

7) Dentro de la función *clogistic* cambie el modelo que va a usar, dentro de *strata* ponga el nombre de la variable que contiene la identificación de cada par de caso control dentro del estudio, y dentro de data ponga el nombre que le dio a su base de datos.

   Para ver la salida del modelo únicamente en el visor de salida de R, coloque el nombre que le asignó al modelo y aquí le saldrán los coeficientes, recuerde que las variables categorías las ordena en orden alfabético tomando como referencia la primera de éstas. Si quiere ver esta salida del modelo un poco más detallada utilizar la función *ci.lin(NOMBREDELMODELO)*.

8) Desde este momento empieza una función compuesta la cual tiene por nombre *macstat* y para poderla usar debe tener en cuenta 4 datos importantes, los cuales son: el nombre con que guardó la base de datos con la que realizó el modelo, el nombre que le asignó al modelo que finalmente escogió, el nombre de la variable identificadora del par de caso control y el nombre de la variable donde está especificada dentro de la pareja quién es el caso y quién el control mediante el uso de un código binario, siendo 0 control y caso 1.

   Dentro de esta función se harán los siguientes pasos:

   - Se asignará dentro de la base de datos el coeficiente de cada variable que se haya tenido en cuenta en el modelo, esto dependerá de la categoría que tenga cada registro, si la variable es numérica se multiplicará el coeficiente por el valor que tenga cada registro en dicha variable y en caso de que la variable sea no numérica se deja el valor del coeficiente.
   - Construye una nueva variable la cual sería como la variable estimada puesto que se suman todos los coeficientes por registro.
   - Se crea una base de datos donde se tiene cada control frente a su respectivo caso, y después de esto, se procede a hacer la diferencia de la suma de los coeficientes.
   - Se grafica de la base de datos anteriormente creada la variable diferencia.
   - Guarda el grafico final bajo el nombre "Unlike pairs.pdf" en formato pdf en el directorio donde tiene guardada su base de datos.

9) Luego de correr todo el código correspondiente a la función, cambie en la función los requerimientos de ésta con los que sean acorde a su análisis. Después de correr la función parece que no puede introducir una nueva función, para que termine de ejecutar, de clic encima del grafico que aparece, sobre los puntos de los cuales desea conocer su respectiva identificación y luego de esto de clic derecho sobre el grafico y seleccione la opción parar, seguidamente verifique dentro de la carpeta donde tiene la base de datos que el grafico ha sido guardado en formato pdf.

Si desea hacer cambios al gráfico, después de los requerimientos obligatorios para la función puede agregar todos los cambios que le desea hacer al gráfico, como pueden ser nombre, color, tamaño, límites, leyendas, todos los cambios permitidos para un gráfico tipo plot.

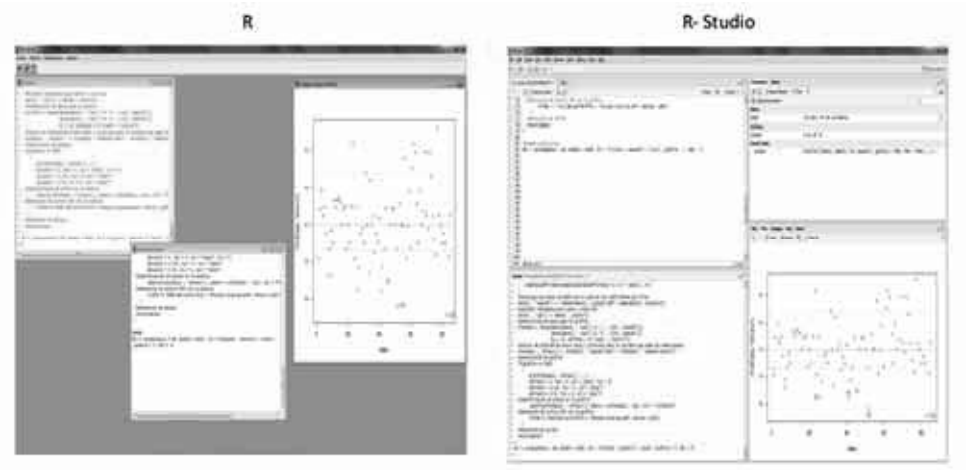
Si desea ver los cambios en la base de datos, abra la base de datos desde R y observe que se crearon más variables donde están los coeficientes asignados dependiendo de los valores de cada registro, también se creó la variable de suma de los coeficientes asignados a cada registro, y al final se crea la variable de código binario, donde 0 es caso y 1 control.

Usted podría no ver la gráfica cambiando la opción grafica = FALSE, pero esto no es deseado dado que es la única manera de observar las diferencias. Usted puede cambiar la opción de que guarde en formato PDF, cambiando en PDF=FALSE; en este caso usted solo podrá observar la gráfica, pero ésta no quedará guardada.

Si usted corre el código bajo la plataforma RStudio cuando finalice de seleccionar las parejas de las que desea conocer su identificación dentro del grafico puede dar ESC y le guardará los cambios, si intenta hacer esto en la consola habitual de R no le guardará los cambios (Grafico 1).

Grafico 1

Comparación de los resultados obtenidos en las dos versiones del programa R



4. Ejecución: ejemplo

CACO: La base de datos para ejemplificar la metodología antes descrita se basa en un estudio de casos y controles emparejados que determina cuáles exposiciones ocurridas antes del nacimiento de un individuo alteran su riesgo de adquirir leucemia linfoblástica aguda antes de los 15 años de vida. Estos datos provienen de un estudio de casos y controles emparejados con una relación 1:1.

Se seguirán los pasos descriptivos en el desarrollo matemático:

1) La base de datos usada para el modelo debe contener los siguientes campos:

   a. **ID:** La variable que contiene el identificador de cada uno de los sujetos incluidos en el estudio tiene el nombre de codencuesta. Pero la variable que tiene la duplicidad de los identificadores tiene el nombre de tripleta.

   b. **CASO-CONTROL:** La variable que define quien es caso y quien es control es caco3.

   c. **EXPOSICIÓN:** La variable que se va a modelar se la misma variable de caso-control.

   d. **VARIABLES EN ESTUDIO:** Las variables que se decidieron modelar son:

      - precowhco24: Variable que define la exposición laboral preconcepción a hidrocarburos: 0 Ni madre, ni padre expuestos; 1 Sólo padre expuesto; 2 Sólo madre expuesta; 3 Ambos expuestos.
      - mptabaq24: Variable que define el tabaquismo activo de los progenitores durante los 24meses anteriores a la concepción: 0 Ninguno fumó; 1 Al menos uno fumó.
      - emn2_35: Variable que define la edad materna durante el nacimiento del menor: 0 Menor de 35 años; 1 35 y más
      - strate2e9: Variable que define el estrato residencial de la madre más frecuente durante el embarazo de interés: 0 Alto; 1 Bajo.

2) La salida del modelo final (Tabla 1)

Tabla 1

Salida del modelo final de la base CACO

	Estimacion	Error estandar	Z	P	IC (95%)	
Exposición laboral a hidrocarburos 24 meses antes de la concepción						
precowhco24_noexpuesto	0.00					
precowhco24_padreexpuesto	0.51	0.48	1.05	0.29	-0.44	1.46
precowhco24_madreexpuesta	1.85	0.76	2.42	0.02	0.35	3.34
precowhco24_ambosexpuestos	2.60	0.71	3.64	0.00	1.20	4.00
Estrato de las viviendas maternas durante el embarazo (mayor frecuencia)						
strate2e9_alto	0.00					
strate2e9_bajo	1.26	0.54	2.33	0.02	0.20	2.32
Tabaquismo paterno y materno 24 meses antes de la concepción						
mptabaq24_ningunofumo	0.00					
mptabaq24_almenosunofumo	0.97	0.38	2.53	0.01	0.22	1.72
Edad materna durante el embarazo						
emn2_35_menorde35	0.00					
emn2_35_mayorde35	1.32	0.61	2.16	0.03	0.12	2.51

3) Participación de la base de datos en dos bases que contengan en la primera base la información de los casos y la segunda base de datos la información de los controles.

   Base de datos 1: Base de datos de casos tiene 85 registros.

   Base de datos 2: Base de datos de controles tiene 85 registros

4) Pronóstico para cada una de las variables que fueron modeladas como factor de riesgo, se hará uso de los coeficientes estimados en el modelo. El cálculo para cada uno de los registros ingresados del "pronostico", para este caso el modelo quedara :

$$y = 0.51\, precowhco24_{padreexpuesto} + 1.85 precowhco24_{madreexpuesta} + 2.60 precowhco24_{ambosexpuestos} + 1.26\, strate2e9_{bajo} + 0.97 mptabaq24_{almenosunofumo} + 1.32 emn2\_35\_mayorde35$$

   En el anexo 2 se observan las estimaciones.

5) Dado el pronóstico para cada uno de los registros de las bases de datos (base de datos de controles y base de datos de casos). Se realiza el pegue de manera horizontal es decir delante de la información de caso 1 quedara la información del control 1, por esta razón es importante tener los identificadores duplicados y los prefijos para cada una de las variables que contenga la base de datos (Tabla 2).

Tabla 2

Organización de datos de la base CACO que junta en la misma fila información de cada pareja

tripleta_ca	tripleta_co	strate2e9_ca	y_strate2e9_ca	strate2e9_co	y_strate2e9_co	emn2_35_ca	y_emn2_35_ca	emn2_35_ca	y_emn2_35_ca
1	1	Bajo	1.26	Bajo	1.26	35 y mas	1.32	Al menos uno fumo	0.97
2	2	Alto	0.00	Alto	1.26	Menor de 35	0.00	Al menos uno fumo	0.97
3	3	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97
4	4	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97
5	5	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00
6	6	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97
7	7	Bajo	1.26	Alto	0.00	Menor de 35	0.00	Al menos uno fumo	0.97
8	8	Bajo	1.26	Alto	0.00	Menor de 35	0.00	Al menos uno fumo	0.97
9	9	Bajo	1.26	Alto	0.00	Menor de 35	0.00	ninguno fumo	0.00
10	10	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97
11	11	Bajo	1.26	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00

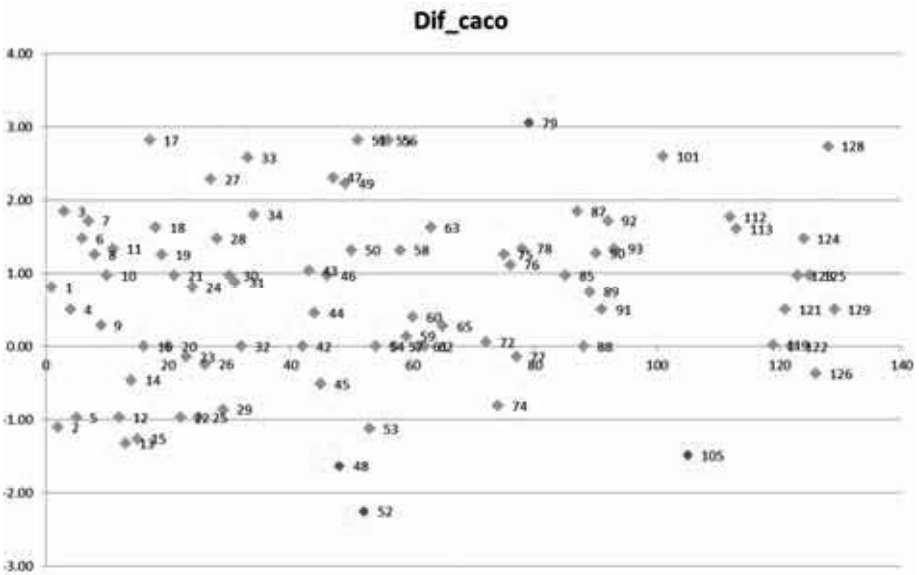
6) Se deben realizar diferencias entre el pronóstico para el caso y el pronóstico para el control. Se calcula:

$$Difcaco_i = y_i caso - y_i control$$

7) Ordenar las diferencias de menor a mayor y se decide un punto de corte para seleccionar las parejas atípicas. Tanto como las mayores negativas como las mayores positivas. Estas parejas encontradas serán las parejas atípicas observadas dentro del conjunto de datos (Gráfico 2).

Gráfico 2

Identificación de las parejas con valores potencialmente influyentes para el modelo final de la base CACO



En las diferencias se observan que las que tiene menor valor son las que más se alejan tanto del origen como del resto de las parejas

5. Discusión

El análisis con este método resulta en la selección de las mismas parejas con valores extremos que fueron identificadas usando las gráficas descritas por Hosmer-Lemeshow (análisis no mostrado). Dos de estas tres parejas tuvieron la mayor influencia en los β: el par 52, que correspondió a un menor control que estuvo expuesto a hidrocarburos (de ambos progenitores) y a tabaquismo durante su etapa preconcepcional, su madre había residido en un estrato bajo durante el embarazo y era menor de 35 años al nacimiento; mientras que el caso no tuvo exposición a hidrocarburos, ni tabaquismo, pero su madre residió en estrato bajo durante la gestación y tenía 35 o más años al nacimiento; y el par 48, compuesto por un caso que sólo estuvo expuesto a tabaquismo activo de los progenitores antes de su concepción, pero que su control tenía exposición a hidrocarburos en la misma etapa.

Dos utilidades que tiene esta prueba son: a. luego de realizar una identificación rápida de las parejas caso-control aparentemente extrañas se puede ir a los datos originales o a la misma fuente de datos (en este estudio, los progenitores de los sujetos) para verificar la información digitada y b. se puede analizar la necesidad de eliminar o no a estas parejas dependiendo de su influencia en el valor de los riesgos obtenidos en el modelo final (comparando el cambio en los riesgos o en sus coeficientes estando incluidas y excluidas).

Sin embargo, la decisión final depende del conocimiento y de la experiencia propia de los investigadores.

6. Anexos

6.1 Anexo 1: Código para R del ejemplo CACO

```
m(list=ls())
##Cargue las librerías necesarias para el análisis
library(foreign)
library(Epi)
library(car)
##Defina el directorio donde esta su base de datos
setwd("C:/Users/MMeneses/Dropbox/INC sin BM/Proyecto MAC/GLOW")
##Verifique que la base de datos este dentro de directorio
dir()
##Lea la base de datos
BD=read.csv("GLOW11M.csv",header=T)
head(BD)
##Verifique los nombres de las variables que va a incluir en el modelo de ser
##necesario anotelos en uns script diferente
##Modelo logistico condicional
Mod1=clogistic(FRACTURE~WEIGHT+BMI+PRIORFRAC+MOMFRAC+ARMASS
IST,strata=PAIR,data=BD)
##Observe la salida del modelo
Mod1
ci.lin(Mod1)


##########################################FUNCION
macstat###############################################
#########################SOLO CORRA ESTE CÓDIGO NO HAGA NINGÚN
CAMBIO#########################
macstat<- function(datos,modelo,id,casoctrl,grafica=TRUE,PDF=TRUE,...){
```

```
#Asigna los coeficientes para las variables categoricas calculados por el modelo
x.modf <- names(modelo$xlevels)
for(kk in x.modf)
{
  for(ll in 1:length(levels(datos[ , kk])))
  {
    if(ll==1)
    {
      datos[datos[, kk] == levels(datos[ , kk])[ll], paste('Coef', kk, sep = '')] <-  0
#unique(datos[ , kk])[ll]
    }
    if(ll>1)
    {
      datos[datos[, kk] == levels(datos[ , kk])[ll], paste('Coef', kk, sep = '')] <-
        modelo$coefficients[names(modelo$coefficients) == paste(kk, levels(datos[
, kk])[ll], sep = '')]
    }
  }
}
#Asigna los coeficientes para las variables numericas calculados por el modelo
x.modn <- attr(modelo$terms, "term.labels")[!attr(modelo$terms, "term.labels") %in%
x.modf]
for(rr in x.modn)
{
  datos[, paste('Coef', rr, sep = '')] <-
modelo$coefficients[names(modelo$coefficients) == rr] * datos[ , rr]
}
#Construye una nueva variable con la suma de los coeficientes por filas
datos[ , 'sumcoef'] <- rowSums(datos[ , c(grep("Coef", names(datos), value=T))])
#Variable indicadora para casos y controles
datos[ , 'caco'] <- datos[ , casoctrl]
```

```
   #Construccion de datos para el grafico
   PlotData <- merge(datos[datos[ , 'caco'] == '1' , c(id, 'sumcoef')],
                                 datos[datos[ , 'caco'] == '0' , c(id, 'sumcoef')],
                                 by = id, suffixes = c(".caso", ".control"))
   #Calculo de diferencias entre casos y controles para la variable que suma los
coeficientes
   PlotData[ , 'difcaco'] <- PlotData[ , "sumcoef.caso"] - PlotData[ , "sumcoef.control"]
   #Construcci?n de grafica
   if(grafica == TRUE)
   {
      plot(PlotData[, 'difcaco'], ...)
      abline(h = 0, lwd = 3, col = "grey", lty = 3)
      abline(h = -1, lty = 3, col = "grey")
      abline(h = -2, lty = 3, col = "grey")
      abline(h = -3, lty = 3, col = "grey")
      #Identificaci?n de puntos en la grafica
      identify(PlotData[ , 'difcaco'], labels = c(PlotData[ , id]), col = "violetred")
      #Generaci?n de archivo PDF con la grafica
      if(PDF == TRUE){dev.print(file = "Parejas atipicas.pdf", device = pdf)}
   }
   #Generacion de salida
   return(datos)
}
##########################################USE LA
FUNCION###############################################
BD_final<-macstat(datos = BD, modelo = Mod1, id = 'PAIR', casoctrl = 'FRACTURE',
grafica = TRUE, PDF = TRUE
                  ,xlab="Parejas",ylab="Diferencia",main="Parejas atipicas")
```

6.2 Anexo 2

Base de datos 1: Base de datos de casos tiene 85 registros

codencuesta_ca	tipocaco_ca	tripleta_ca	strate2e9_ca	y_strate2e9_ca	emn2_35_ca	y_emn2_35_ca	emn2_35_ca	y_emn2_35_ca	precowhco24_ca	y_precowhco24_ca	y_pronostico
1	1	1	Bajo	1.26	35 y mas	1.32	Al menos uno fumo	0.97	ninguno expuesto	0.00	3.55
4	1	2	Alto	0.00	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	1.48
7	1	3	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	madre expuesta	1.85	4.08
10	1	4	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
13	1	5	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00	ninguno expuesto	0.00	1.26
16	1	6	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
19	1	7	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
22	1	8	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
25	1	9	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	1.77
28	1	10	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
31	1	11	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00	madre expuesta	1.85	3.11
34	1	12	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	1.77
37	1	13	Bajo	1.26	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
40	1	14	Bajo	1.26	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	1.77

Base de datos 2: Base de datos de controles tiene 85 registros

codencuesta_co	tipocaco_co	tripleta_co	strate2e9_co	y_strate2e9_co	caco3_co	emn2_35_co	y_emn2_35_c	emn2_35_co	y_emn2_35_co	precowhco24_co	y_precowhco24_co	y_pronostico
3	3	1	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
6	3	2	Alto	1.26	0	35 y mas	1.32	ninguno fumo	0.00	ninguno expuesto	0.00	2.58
9	3	3	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
12	3	4	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
15	3	5	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
18	3	6	Bajo	1.26	0	Menor de 35	0.00	ninguno fumo	0.00	ninguno expuesto	0.00	1.26
21	3	7	Alto	0.00	0	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	0.51
24	3	8	Alto	0.00	0	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	1.48
27	3	9	Alto	0.00	0	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	1.48
30	3	10	Bajo	1.26	0	Menor de 35	0.00	ninguno fumo	0.00	ninguno expuesto	0.00	1.26
33	3	11	Bajo	1.26	0	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	1.77
36	3	12	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	padre expuesto	0.51	2.74
39	3	13	Bajo	1.26	0	35 y mas	1.32	Al menos uno fumo	0.97	padre expuesto	0.51	4.06
42	3	14	Bajo	1.26	0	Menor de 35	0.00	Al menos uno fumo	0.97	ninguno expuesto	0.00	2.23
45	3	15	Bajo	1.26	0	Menor de 35	0.00	ninguno fumo	0.00	padre expuesto	0.51	1.77

Referencias

BRESLOW, N.E. Y DAY, N.E. (1982) «Statistical Methods in Cancer Research. Volume I - The analysis of case-control studies». IARC Scientific Publications No. 32, Lyon, 247-279.

CASTRO-JIMÉNEZ, M. Á. Y OROZCO-VARGAS, L. C. (2011) «Parental Exposure to Carcinogens and Risk for Childhood Acute Lymphoblastic Leukemia, Colombia, 2000-2005». *Preventing Chronic Disease*, 8(5), A106. http://www.cdc.gov/pcd/issues/2011/sep/10_0201.htm.

COOK, R.D. (1977) «Detection of influential observations in lineal regression». *Technometrics*, 19, 1, 15-18.

EKSTROM, C. (2017). R Primer, Second Edition. New York: Chapman and Hall/CRC.

GREENLAND, S. (1989) «Modelling and Variable Selection in Epidemiologic Analysis». *Am J Public Health*, 79, 3, 340-349.

HOSMER, D.W. Y LEMESHOW, S. (1989) «Applied Logistic Regression». A Wiley-Interscience Publication. New York.

HOSMER, D.W., TABER, S. Y LEMESHOW, S. (1991) «The importance of assessing the fit of Logistic Regression Models: a Case Study». *Am J Public Health,* 81, 12, 1630-1635.

ROTHMAN, K.J. (1986) «Epidemiología Moderna». Ediciones Díaz de Santos, 317-346

STEVENS, J.P. (1984) «Outliers and Influential Data Points in Regression Analysis. Psychological Bulletin; 95(2): 334-344.