ENDES 2017 en R: análisis de salud cardiovascular ideal

Akram Hernández, Horacio Chacón

14 de septiembre de 2018

# Descripción

Este es un ejemplo de extracción, unión y análisis de datos de la **Encuesta Demográfica de Salud Familiar (ENDES)** para el uso particular del cálculo de los indicadores ideales de salud cardiovascular definidos por la **American heart Association (AHA)** [[1](#ref-RN1)].

Para poder calcular el estado cardiovascular de la población peruana, utilizamos estos indicadores, de los cuales pueden encontrarse en la ENDES: *Presión Arterial, Dieta saludable, IMC, Tabaquismo y Diabetes*.

## 1. Librerías

Cargamos las librerías necesarias para el análisis. El uso de estos paquetes permite ejecutar funciones específicas que no vienen por defecto en los comandos base de R.

library(foreign)  
library(tidyverse)  
library(plyr)  
library(survey)

En caso no esten instaladas previamente las librerias, instalarlas con el comando install.package().

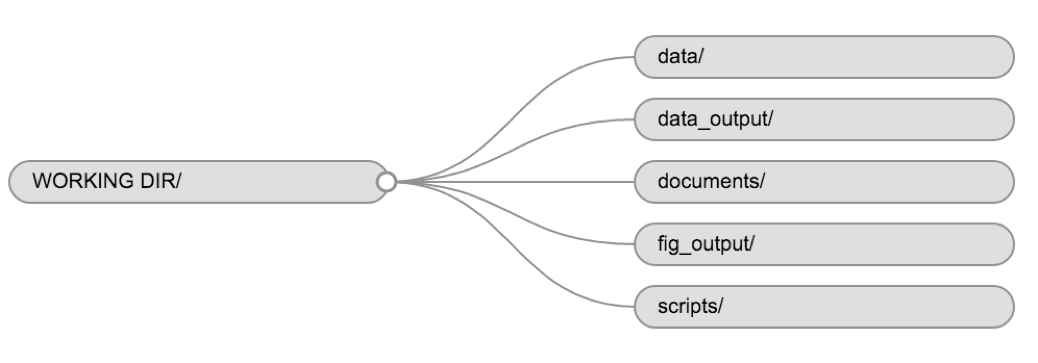
## 2. Lectura de datos

Antes de poder leer los datos debemos descargarlos desde la web de INEI (<http://iinei.inei.gob.pe/microdatos/Consulta_por_Encuesta.asp>). En este caso utilizaremos las bases **CSALUD01, RE223132, REC42, RECH1, RECH23, RECH0**. Estas bases pueden ser encontradas en los módulos **64, 65, 67, 70 y 414**.



*Imagen 1: Página de consulta de las bases de datos de la ENDES (INEI)*

Una vez descargados los archivos en formato **.sav** (SPSS), debemos guardarlos en una carpeta de facil acceso. Para facilidad de consulta, nosotros la guardaremos en la carpeta **/data/** de nuestra raiz de proyecto. Para mantener orden en el proceso usamos la siguiente estructura de carpeta.



*Imagen 2: Estructura de archivos del proyecto*

Una vez preparados los archivos procedemos a la importación de cada una de las bases con el comando **read.spss()** del paquete *Foreign*. Adicionalmente utilizamos el comando **as.tibble()** para convertir las bases al formato Tibble [[2](#ref-wickham2016r)]

CSALUD01 <- as.tibble(read.spss("data/CSALUD01.sav",to.data.frame = T))  
MUJER\_RE223132 <- as.tibble(read.spss("data/RE223132.SAV",to.data.frame = T))  
MUJER\_REC42 <- as.tibble(read.spss("data/REC42.SAV",to.data.frame = T))  
HOGAR\_RECH1 <- as.tibble(read.spss("data/RECH1.SAV",to.data.frame = T))  
VIVIENDA\_RECH23 <- as.tibble(read.spss("data/RECH23.SAV",to.data.frame = T))  
HOGAR\_RECH0 <- as.tibble(read.spss("data/RECH0.SAV",to.data.frame = T))

Podemos, a manera de ejemplo, explorar los datos mediante el comandos **head()**. No recomendamos usar **summary()** ni **view()** por la cantidad de variables.

head(CSALUD01)

## # A tibble: 6 x 263  
## QHCLUSTER QHNUMBER QHHOME QSNUMERO QSINTD QSINTM QSINTY QSEQUIPO QSENTREV  
## <dbl> <dbl> <dbl> <dbl> <dbl> <fct> <fct> <fct> <dbl>  
## 1 1 14 1 1 27 Mayo 2017 AMAZONA~ 8  
## 2 1 25 1 2 27 Mayo 2017 AMAZONA~ 2  
## 3 1 32 1 2 29 Mayo 2017 AMAZONA~ 8  
## 4 1 36 1 2 29 Mayo 2017 AMAZONA~ 2  
## 5 1 45 1 2 27 Mayo 2017 AMAZONA~ 8  
## 6 1 52 1 2 27 Mayo 2017 AMAZONA~ 2  
## # ... with 254 more variables: QSTOTVISIT <dbl>, QSRESULT <fct>,  
## # QSRESULTX <fct>, QSNINOS <dbl>, QSRESINF <fct>, QSORDNIN1 <fct>,  
## # QSRESNIN1 <fct>, QSORDNIN2 <fct>, QSRESNIN2 <fct>, QSORDNIN3 <fct>,  
## # QSRESNIN3 <fct>, QSORDNIN4 <fct>, QSRESNIN4 <fct>, QSORDNIN5 <fct>,  
## # QSRESNIN5 <fct>, QSORDNIN6 <fct>, QSRESNIN6 <fct>, QSORDNIN7 <fct>,  
## # QSRESNIN7 <fct>, QSORDNIN8 <fct>, QSRESNIN8 <fct>, QSORDNIN9 <fct>,  
## # QSRESNIN9 <fct>, QSORDNIND <fct>, QSRESNIND <fct>, CONTROL\_S <fct>,  
## # QS20C <dbl>, QSSEXO <fct>, QS21AH <dbl>, QS21AM <dbl>, QSMEF <fct>,  
## # QSDIA <fct>, QS22M <fct>, QS22A <fct>, QS23 <fct>, QS24 <fct>,  
## # QS25N <fct>, QS25AG <fct>, QS25A <fct>, QS25G <fct>, QS25AA <fct>,  
## # QS25AAX <fct>, QS25BB <fct>, QS25BBX <fct>, QS26 <fct>, QS27 <fct>,  
## # QS27X <fct>, QS28 <fct>, QS29A <fct>, QS29B <fct>, QS100 <fct>,  
## # QS101 <fct>, QS101CX <fct>, QS101X <fct>, QS102 <fct>, QS103U <fct>,  
## # QS103C <dbl>, QS104 <fct>, QS105 <fct>, QS106 <fct>, QS107 <fct>,  
## # QS108 <fct>, QS108CX <fct>, QS108X <fct>, QS109 <fct>, QS110U <fct>,  
## # QS110C <dbl>, QS111 <fct>, QS112 <fct>, QS113 <fct>, QS202 <fct>,  
## # QS203U <fct>, QS203C <dbl>, QS204U <fct>, QS204C <dbl>, QS205U <fct>,  
## # QS205C <dbl>, QS206 <fct>, QS207U <fct>, QS207C <dbl>, QS208 <fct>,  
## # QS209 <fct>, QS210 <fct>, QS211U <fct>, QS211C <dbl>, QS212T <fct>,  
## # QS212A <fct>, QS212AV <fct>, QS212AB <fct>, QS212B <fct>,  
## # QS212BV <fct>, QS212BB <fct>, QS212C <fct>, QS212CV <fct>,  
## # QS212CB <fct>, QS212D <fct>, QS212DV <fct>, QS212DB <fct>,  
## # QS212E <fct>, QS212EV <fct>, ...

## 3. Preparación de las bases

Para poder unir las bases de datos es necesario crear nuevas variables para luego poder hacer una unión (merge) satisfactorio entre las distintas bases. La unión se basa en dos identificadores que son *HHID* y *QSNUMERO*. En las bases que se carezca de una u otra de estas variables se deberá, como se demuestra a continuación, generarlas.

En **MUJER\_RE223132** creamos *HHID* y *QSNUMERO*. Para obtenerlos debemos entender que *HHID* es la subcadena de texto sin los 3 últimos caracteres de CASEID. Por otro lado, *QSNUMERO* es la subcadena restante (2 últimos caracteres) de la variable *CASEID*. Para extraer subcadenas de texto utilizamos el comando **str\_sub()**:

MUJER\_RE223132$HHID <- str\_sub(MUJER\_RE223132$CASEID,1,(str\_length(MUJER\_RE223132$CASEID)-3))  
MUJER\_RE223132$QSNUMERO <- str\_sub(MUJER\_RE223132$CASEID,-2,-1)  
MUJER\_RE223132$QSNUMERO <- as.numeric(MUJER\_RE223132$QSNUMERO)

De la misma forma, en **MUJER\_REC42** creamos *HHID* y *QSNUMERO* para poder facilitar la unión de las bases posteriormente:

MUJER\_REC42$HHID <- str\_sub(MUJER\_REC42$CASEID,1,(str\_length(MUJER\_REC42$CASEID)-3))  
MUJER\_REC42$QSNUMERO <- str\_sub(MUJER\_REC42$CASEID,-2,-1)  
MUJER\_REC42$QSNUMERO <- as.numeric(MUJER\_REC42$QSNUMERO)

En RECH1, la variable *HVIDX* es equivalente a *QSNUMERO* (el identificador de fila o individuo). Por eso, la renombramos con el siguiente comando:

HOGAR\_RECH1$QSNUMERO <- HOGAR\_RECH1$HVIDX

## 4. Unión de Datos

Cada una de estas bases contiene campos necesarios para obtener la base final que nos permitirá calcular los indicadores de salud ideal cardiovascular. Para poder referenciar entre una base y otra existen identificadores (mencionados previamente) en las diferentes bases. Estos identificadores, según el nivel de desagregación de cada base permiter unirlas relacionalmente.

Esta unión será realizada secuencialmente, creando bases intermedias, hasta conseguir una base de datos que englobe todas las variables requeridas.

Empezamos juntando las bases HOGAR\_RECH1 y CSALUD01 en la base intermediaria **BASE1**. El comando utilizado es **join()** tomando como parametro muy importante el ‘*Left Join*’ [[3](#ref-wiki:join)]. Las variables enlace para la unión o join son tanto *HHID* como *QSNUMERO*. Hacemos lo mismo con el resto de de bases hasta tener una **BASE5**.

BASE1 <- join(HOGAR\_RECH1, CSALUD01, by = c("HHID","QSNUMERO"),type="left")  
BASE2 <- join(MUJER\_RE223132, MUJER\_REC42, by = c("HHID", "QSNUMERO"),type='left')  
BASE3 <- join(HOGAR\_RECH0, VIVIENDA\_RECH23, by = "HHID",type='left')  
BASE4 <- join(BASE1,BASE2,by = c("HHID","QSNUMERO"),type='left')  
BASE5 <- join(BASE3,BASE4,by = 'HHID', type = 'left')

## 5. Aplicación de los filtros

Podemos corroborar cuantas observaciones tiene nuestra **BASE5**.

nrow(BASE5)

## [1] 143094

Ahora es momento de aplicar los filtros necesario para quedarnos solamente con los sujetos necesarios para el análisis. En primer lugar descartamos a todas las gestantes (**V213**) y luego filtramos a solamente los informantes cuya información haya sido reportada como completa (**QSRESINF**).

BASE\_FINAL <- filter(BASE5,V213!="Yes" | is.na(V213))  
BASE\_FINAL <- filter(BASE\_FINAL,QSRESINF == "Iniciar entrevista")

La población final que tiene nuestra **BASE\_FINAL** es de:

nrow(BASE\_FINAL)

## [1] 32514

Borramos los objetos creados que ya no utilizaremos para el análisis.

rm(CSALUD01,BASE1,BASE2,BASE3,BASE4,BASE5,HOGAR\_RECH1,HOGAR\_RECH0,VIVIENDA\_RECH23,MUJER\_REC42,MUJER\_RE223132)

## 6. Creación de Variables para Salud Cardiovascular

# Bibliografía

[1] American Heart Association. 2017 statistical fact sheet: Cardiovascular health. 2017.

[2] Wickham H, Grolemund G. R for data science: Import, tidy, transform, visualize, and model data. “ O’Reilly Media, Inc.” 2016.

[3] Wikipedia. Join — wikipedia, la enciclopedia libre 2018.