

Test

Joshua Kock

11/10/2018

Introducción:

Para el presente ejercicio utilizaremos datos de mamografía y cancer de seno disponibles libremente en <https://www.kaggle.com/overratedgman/mammographic-mass-data-set>, bajo la licencia de CC BY-NC-SA 4.0. Las pacientes fueron reclutados en el año 2007 midiendo variables relacionadas con la mamografía y posteriormente fueron sometidas a biopsia para definir la malignidad o no de los hallazgos imagenológico. Con los presentes datos realizaremos un modelo lineal generalizado logístico. La metodología a implementar es:

- (1) La exploración inicial de la base de datos y las variables.
- (2) La generación de un Modelo lineal generalizado.
- (3) La simplificación del modelo (si aplica).
- (4) El diagnóstico del modelo.
- (5) Ajuste del modelo (si es necesario)
- (6) Conclusión.

En nuestro ejercicio vamos a establecer a *priori* como mi variable respuesta la presencia o no de malignidad en la biopsia de seno.

Posteriormente, cargamos los datos del archivo llamado **cancer_seno** y le asignamos el nombre **ca_mama** a los datos; de esta forma se genera un objeto con el nombre gases. Dentro de R este objeto tendrá la característica de ser un dataframe.

```
ca_mama <- read.csv(here("data", "cancer_seno.csv"), header = TRUE)
```

Exploración de base de datos.

Una vez cargada la base de datos exploramos su estructura con la función **str** “structure” y observamos que la base de datos contiene **5 variables** y **830 observaciones**. De igual manera procedo a cambiar el nombre de las variables al español con la función **names**.

```
str(ca_mama)

## 'data.frame':    830 obs. of  5 variables:
## $ Age      : int  67 58 28 57 76 42 36 60 54 52 ...
## $ Shape    : int  3 4 1 1 1 2 3 2 1 3 ...
## $ Margin   : int  5 5 1 5 4 1 1 1 1 4 ...
## $ Density  : int  3 3 3 3 3 3 2 2 3 3 ...
## $ Severity: int  1 1 0 1 1 1 0 0 0 0 ...

names(ca_mama) <- c("edad",
                    "forma",
                    "margen",
                    "dens",
                    "sev")
```

Una vez importada la base, hay que identificar si hay variables que faltan (datos perdidos), lo anterior lo hacemos con **is.na**, que hace parte del paquete **{base}**, en conjunto con la función de la familia **apply**. Con el resultado podemos concluir que no tenemos datos perdidos en mi data frame.

```
sapply(ca_mama,function(x) sum(is.na(x)))
```

```
##  edad  forma margen  dens   sev  
##    0     0      0     0     0
```

Hacemos la tabla de las variables que hacen parte de la base de datos donde se explica que tipo tienen en R y el tipo de variable así como sus unidades.