Tipología y ciclo de vida de los datos

Limpieza y validación de los datos

Alumno:
Cantero Lorenzo, Javier (Aula 3)

Índice de contenidos

1. Descripción del dataset

1.1 Introducción

En la literatura moderna se ha identificado la obesidad como un factor de riesgo para el cáncer de mama en mujeres premenopáusicas. El aumento de los niveles de estrógenos en las mujeres derivados de esta condición, evaluado según el biomarcador *serum estradiol*, es el principal factor de riesgo identificado en relación al cáncer de mama.

Con el objetivo de evaluar estas relaciones, se estudió el caso de 211 mujeres americanas en edad premenopáusica, 151 de las cuales eran afroamericanas y 60 caucásicas. Como métricas de la adiposidad se utilizaron BMI y WHR. Un análisis hormonal adicional permite conocer el nivel de estradiol de cada una de estas mujeres. Finalmente se consideran otros factores de riesgo como pueden ser el número de hijos o la edad de la mujer.

1.2 Descripción

El dataset ha sido obtenido en [1] e incluye 211 registros identificados por 11 descriptores:

- **Id:** Identificador.
- **Estrad:** Niveles de *serum estradiol* (a partir de una analítica hormonal).
- **Ethnic:** Etnia (opciones: 'Caucasian', 'African American').
- **Entage:** Edad.
- **NumChild:** Número de hijos.
- **Agefbo:** Edad en la que la persona ha tenido su primer hijo.
- **Anykids:** ¿Tiene hijos? (Opciones: '0' (no), '1' (sí)).
- **Agemenar:** Edad de la menarquia (primera menstruación).
- **BMI:** Medida de la adiposidad general. [2]
- **WHR:** Medida de la adiposidad abdominal.[3]
- **Area:** Hábitat (Opciones: '0' (urbana), '1' (rural)).

2. Integración y selección de los datos

Para la realización de esta práctica seleccionaremos los datos correspondientes a las historias médicas y analíticas hormonales del repositorio original [1]. Originalmente los datos se encuentran en formato *rdata*. Para poder generalizar el lenguaje de trabajo y facilitar su entrega de cara a esta práctica convertiremos este archivo a formato *csv*.

Para ello hacemos doble click sobre el archivo *ESTRADL.DAT.rdata*, lo que permitirá cargar el conjunto de datos en RStudio con el nombre *estradl*. Mediante la siguiente secuencia guardaremos en nuestra máquina una copia en formato csv.

```
write.csv(estradl, file='estradiol.csv')
```

Los datos necesarios para la realización de la práctica se encuentran en un único fichero. Por este motivo no se requerirá de procesos adicionales de integración.

3. Limpieza de los datos

A lo largo de este apartado realizaremos distintos procedimientos de evaluación inicial y preprocesamiento de los datos con el objetivo de asegurar la consistencia y facilitar los análisis posteriores.

Comenzamos cargando los datos asegurando el uso de la coma como separador y el uso del estándar americano como separador decimal.

```
# load csv file
input_filename <- "estradiol.csv"
dataset <- read.csv(input_filename, sep=",", dec=".")
```

Para comprobar que la información se ha cargado correctamente comprobamos el número de registros, el número de atributos, y el nombre de cada uno de los atributos.

```
# number of records
nrow(dataset)
```

```
## [1] 211
```

```
# number of variables
ncol(dataset)
```

```
## [1] 11
```

```
# name of the variables
colnames(dataset)
```

```
##  [1] "X"        "Id"       "Estradl"  "Ethnic"   "Entage"   "Numchild"
##  [7] "Agefbo"   "Anykids"  "Agemenar" "bmi"      "whr"
```

Identificamos la presencia de un atributo no deseado (de nombre 'X'), que más adelante eliminaremos. Realizamos una exploración preliminar de algunos registros del conjunto de datos para asegurar que el tipo y rango de cada atributo ha sido cargado correctamente.

X	Id	Estradl	Ethnic	Entage	Numchild	Agefbo	Anykids	Agemenar	BMI	WHR	Area
157	3440	15.3	African American	13.0	2	23	1	27	35,6201	0,75	1
205	3470	37.6	African American	24	0	0	0	11.0	37,3134	0,81	0
184	5031	65	Af Am	30	4	16	1	15.0	29,1801	0,89	1
188	17	26.83	Caucasian	35	0	0	0	12.5	28,5377	0,83	0
170	39	24.1	Caucasian	19	0	0	0	12.0	29,0085	0,77	1

Una vez comprobada la integridad de la carga de los datos procedemos a eliminar el atributo adicional identificado para que no interfiera en el resto de los procesos de limpieza.

```
dataset$X <- NULL
```

Identificamos algunas inconsistencias de formato entre variables numéricas y categóricas. Para asegurar que el resto de procedimientos son consistentes indicamos explícitamente el tipo de dato de cada atributo y estandarizamos el criterio decimal.

```
# apply American standard of decimal separation
columns_quantitative <- c('Id', 'Entage', 'Numchild', 'Agefbo', 'Agemenar', 'Estradl', 'BMI', 'WHR')
dataset[columns_quantitative] <- sapply(dataset[columns_quantitative], gsub, pattern=',', replacement='.')

# convert to numerical data type
dataset[columns_quantitative] <- sapply(dataset[columns_quantitative], as.character)
dataset[columns_quantitative] <- sapply(dataset[columns_quantitative], as.numeric)
```

3.1 Inconsistencias

Identificamos distintos tipos de inconsistencias que pueden darse en los datos. Para cada tipo de inconsistencia aplicaremos técnicas específicas para corregirla.

- **Errores de formato:** Inconsistencias derivadas de errores de tipografía o formato.
- **Errores de coherencia:** Inconsistencias derivadas de la lógica semántica de la variable.

Mientras que los errores de formato trataremos de corregirlos manualmente, los errores de coherencia (entre los que se encuentran los valores extremos) requerirán de métodos más sofisticados. La técnica escogida será sustituir estas inconsistencias por valores nulos. De este modo trataremos las inconsistencias como valores faltantes que serán completados con métodos de imputación. Estos se sumarán a los valores faltantes que se encuentren por defecto en el conjunto de datos.

3.1.1 Identificación de inconsistencias univariables

Comenzamos estudiando la unicidad del atributo *Id*. Al ser un identificador único para cada mujer no deberían observarse valores duplicados (fuente de inconsistencias).

```
# duplicate values
subset(dataset$Id, duplicated(dataset$Id))
```

```
## [1] 2
```

Debido a que encontramos una duplicidad de registro, optamos por eliminarla del conjunto.

```
# remove rows with same id (keeping one)
dataset <- subset(dataset, !duplicated(dataset$Id))
```

Sabemos que el atributo *Ethnic* sólo puede presentar dos valores: 'African American' y 'Caucasian'. Mostramos por pantalla los distintos valores de este atributo en el conjunto de datos.

Var1	Freq
African American	11
Caucsian	11
Af Am	17
African american	22
African American	101
Caucacian	3
Caucasian	43
Caucsian	2

Mapeamos las faltas tipográficas a los posibles valores del atributo.

```
# trim whitespaces
dataset$Ethnic <- trimws(dataset$Ethnic)

# title capitalization standard
dataset$Ethnic <- str_to_title(dataset$Ethnic)

# specific spelling errors
dataset$Ethnic <- str_replace_all(dataset$Ethnic, "Caucsian", "Caucasian")
dataset$Ethnic <- str_replace_all(dataset$Ethnic, "Caucacian", "Caucasian")
dataset$Ethnic <- str_replace_all(dataset$Ethnic, "Af Am", "African American")
```

Comprobamos que los valores son ahora adecuados.

Var1	Freq
African American	151
Caucasian	59

Efectivamente ya sólo contamos con los valores posibles. Vemos como el registro duplicado que hemos eliminado anteriormente correspondía a una mujer de etnia caucásica.

Los atributos *Anykids* y *Area* son variables categóricas en formato *Label encoding*. Para facilitar la interpretabilidad de los datos asignamos valores de texto que describan mejor las categorías.

```
# data type transformation
dataset$Anykids <- as.factor(dataset$Anykids)

# one-hot encoding to label encoding
dataset$Anykids <- factor(dataset$Anykids, levels=c(0, 1), labels=c("No", "Yes"))
```

```
# data type transformation
dataset$Area <- as.factor(dataset$Area)

# one-hot encoding to label encoding
dataset$Area <- factor(dataset$Area, levels=c(0, 1), labels=c("Urban", "Rural"))
```

A continuación procedemos a realizar un análisis de los valores extremos de las distintas variables numéricas. Para ello usaremos el criterio de considerar valor extremo como aquel más alejado de 1.5 veces el rango intercuartil (pese a existir otros criterios similares). Para facilitar el proceso utilizaremos directamente visualizaciones de tipo boxplot, que realizarán automáticamente los cálculos.

Mostramos por pantalla los valores atípicos.

```
boxplot.stats(dataset$Estradl)$out
```

```
## [1]  94.00 332.70  95.57 170.10  97.40  95.50
```

```
boxplot.stats(dataset$Agemenar)$out
```

```
## [1] 31 22 99 25 99 99 23 27 99
```

El caso del atributo *Agefbo* es un caso especial en el que las mujeres sin hijos son indicadas por el valor 0. Para evitar que afecten al análisis de outliers, indicamos estos como valores NA.

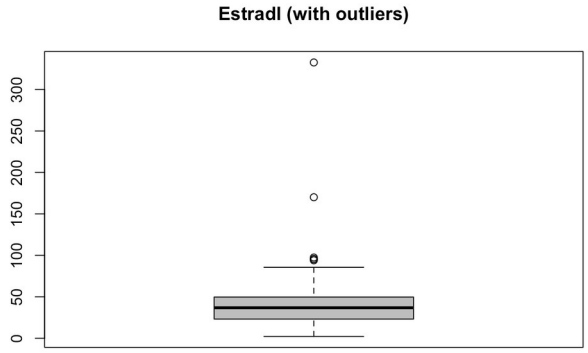
```
tmp_agefbo <- dataset$Agefbo
tmp_agefbo[tmp_agefbo==0] <- NA

boxplot.stats(tmp_agefbo)$out
```

```
## [1] 99 99 99  4 99 99 99
```

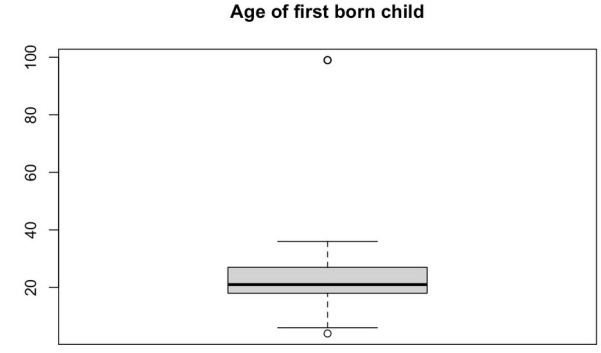
Según la semántica de los atributos (donde valores atípicos podrían ser reales), solo dejaremos de considerar aquellos valores que claramente sean atípicos. Para ello mostramos explícitamente la visualización boxplot para cada variable.

```
boxplot(dataset$Estradl, main='Estradl (with outliers)', col='gray')
```




```
boxplot(dataset$Agemenar, main='Age of menarche')
```


```
boxplot(tmp_agefbo, main='Age of first born child')
```



Eliminamos aquellos valores extremos fácilmente identificables.

```
dataset$Estradl[dataset$Estradl > 100] <- NA
dataset$Agemenar[which(dataset$Agemenar %in% boxplot.stats(dataset$Agemenar)$out)] <- NA

dataset$Agefbo[dataset$Agefbo==0] <- NA
dataset$Agefbo[which(dataset$Agefbo %in% boxplot.stats(dataset$Agefbo)$out)] <- NA
```

El resto de atributos también han sido analizados con técnicas similares. Al no identificar ninguna inconsistencia particular se han omitido estos procedimientos en el informe. Estos análisis pueden encontrarse en el código adjunto.

3.1.2 Identificación de inconsistencias bivariables

Del mismo modo que hemos trabajado sobre las inconsistencias particulares de cada uno de los atributos, buscaremos la consistencia lógica en la semántica entre variables.

En primer lugar comprobaremos que el atributo *Anykids* realmente indique como valor positivo aquellos registros en los que el atributo *Numchild* tenga un valor no nulo. Los registros que no satisfagan esta condición serán, de la misma manera que hemos hecho hasta ahora, clasificados como valores desconocidos.

```
# save indexes with NA values for Anykids attribute
anykids_na <- is.na(dataset$Anykids)

# create new Anykids attribute using Numchild
consistent_anykids <- as.numeric(dataset$Numchild > 0)

# restablish NA entries
consistent_anykids[is.na(dataset$Anykids)] <- NA

dataset$Anykids <- consistent_anykids
```

Otra inconsistencia bivariable que estudiaremos es la que puede ocurrir entre los atributos *Entage* y *Agemenar*. En efecto, la edad de la mujer debe ser superior a la edad de su menarquia.

```
# temporal consistent attributes
tmp_entage <- pmax(dataset$Entage, dataset$Agemenar)
tmp_agemenar <- pmin(dataset$Entage, dataset$Agemenar)

dataset$Entage <- tmp_entage
dataset$Agemenar <- tmp_agemenar
```

Finalmente consideraremos la relación entre los atributos *Agefbo*, *Agemenar* y *Numchild*. En efecto, la edad en la que se ha tenido el primer hijo ha de ser superior al año de menarquia, y debe corresponder con un valor no nulo en el número de hijos.

```
# temporal consistent attributes
tmp_agefbo <- pmax(dataset$Agefbo, dataset$Agemenar)
tmp_agemenar <- pmin(dataset$Agefbo, dataset$Agemenar)

# agefbo is only valid if numchild > 0
tmp_agefbo[dataset$Numchild <= 0] = 0

dataset$Agefbo <- tmp_agefbo
dataset$Agemenar <- tmp_agemenar
```

3.2 Imputación de valores

En la sección anterior hemos identificado todas las inconsistencias y las hemos sustituido por valores faltantes. Contamos con distintas técnicas para tratar este tipo de valores

Un posible método sería eliminar todos aquellos registros que contaran con valores faltantes. En nuestro caso se considera que no resulta adecuado al contar un dataset de tamaño pequeño.

Otra técnica es la de realizar una imputación de los valores faltantes. Para que estos valores sean estadísticamente consistentes podemos escoger como criterio de relleno la media o mediana del resto de valores del atributo. Otras posibles técnicas más avanzadas sería el uso de modelos más avanzados (como por ejemplo regresiones, o predictores ML).

Un posible inconveniente de esta metodología es la de crear valores sintéticos que puedan perder la correlación entre este atributo y algún otro, como por ejemplo, la etnia.

Seleccionamos como estrategia imputar los valores cuantitativos a partir de un modelo KNN donde consideraremos los tres registros cercanos de la misma etnia según la distancia de Gower.

```
# copy of original dataset before imputation
tmp_original_dataset <- dataset
ids_with_nans <- subset(dataset$Id, !complete.cases(dataset))

# temporal dataframes
tmp_dataset_afam <- subset(dataset, dataset$Ethnic == 'African American')
tmp_dataset_cauc <- subset(dataset, dataset$Ethnic == 'Caucasian')

# imputation with ethnic group neighbours
tmp_dataset_afam <- kNN(tmp_dataset_afam, variable=c('Agemenar', 'Agefbo', 'Entage', 'Estradl', 'Numchild'), k=3,
 imp_var=FALSE)

tmp_dataset_cauc <- kNN(tmp_dataset_cauc, variable=c('Agemenar', 'Agefbo', 'Entage', 'Estradl', 'Numchild'), k=3,
 imp_var=FALSE)

# join temporal datasets
dataset <- rbind(tmp_dataset_afam, tmp_dataset_cauc)
```

Mostramos algunos registros con valores faltantes antes y después de la imputación.

```
# original dataset
original_subdataset <- subset(tmp_original_dataset, tmp_original_dataset$Id %in% ids_with_nans)
kable(original_subdataset[order(original_subdataset$Id),])
```

	Id	Estradl	Ethnic	Entage	Numchild	Agefbo	Anykids	Agemenar	BMI	WHR	Area
7	11	32.47	Caucasian	32	9	NA	Yes	13	20.0959	0.73	Rural
80	27	NA	Caucasian	29	0	0	No	13	19.7029	0.79	Urban
126	48	49.50	Caucasian	NA	0	0	No	NA	27.0953	0.76	Rural
31	61	NA	Caucasian	33	0	0	No	15	20.1646	0.68	Rural
150	210	39.93	African American	NA	9	NA	Yes	NA	33.3769	0.68	Rural
152	228	18.37	African American	NA	0	0	No	NA	27.8785	0.73	Rural
122	229	25.13	African American	22	9	NA	Yes	11	25.4509	0.77	Rural
157	3440	15.30	African American	NA	2	23	Yes	NA	35.6201	0.75	Rural
53	3460	12.90	African American	NA	0	0	No	NA	23.1901	0.73	Rural

```
# imputed dataset
subdataset <- subset(dataset, dataset$Id %in% ids_with_nans)
kable(subdataset[order(subdataset$Id),])
```

	Id	Estradl	Ethnic	Entage	Numchild	Agefbo	Anykids	Agemenar	BMI	WHR	Area
157	11	32.47	Caucasian	32	9	19	Yes	13.0	20.0959	0.73	Rural
190	27	38.97	Caucasian	29	0	0	No	13.0	19.7029	0.79	Urban
203	48	49.50	Caucasian	20	0	0	No	12.0	27.0953	0.76	Rural
181	61	66.67	Caucasian	33	0	0	No	15.0	20.1646	0.68	Rural
95	210	39.93	African American	29	9	23	Yes	11.0	33.3769	0.68	Rural
97	228	18.37	African American	22	0	0	No	12.5	27.8785	0.73	Rural
70	229	25.13	African American	22	9	20	Yes	11.0	25.4509	0.77	Rural
102	3440	15.30	African American	30	2	23	Yes	12.0	35.6201	0.75	Rural
19	3460	12.90	African American	21	0	0	No	13.0	23.1901	0.73	Rural

Una vez preprocesado y limpiado el conjunto de datos, lo guardamos en formato csv para su posteriores análisis.

```
output_filename <- 'estradiol_clean.csv'
write.csv(dataset, file=output_filename, row.names=FALSE)
```

4. Análisis de los datos

4.1 Selección y planificación

En este apartado identificaremos los distintos análisis que queremos llevar a cabo sobre los datos.

```
dataset <- read.csv('estradiol_clean.csv', sep=",", dec=".")
```

El primer procedimiento que realizaremos será comprobar la normalidad y homogeneidad de la varianza de las distintas variables cuantitativas del conjunto de datos. Esto nos permitirá asegurar que se satisfacen los prerequisitos de los distintos métodos estadísticos que aplicaremos más adelante.

Evaluaremos la correlación entre las variables del conjunto de datos para estudiar su impacto directo sobre los niveles de estradiol en las mujeres desde una perspectiva meramente cualitativa. Esto servirá como punto de referencia para los otros dos análisis.

Aplicando un contraste de hipótesis comprobaremos la validez de la investigación médica, que afirma que los niveles de estradiol en mujeres caucásicas no corresponde, de media, con los niveles en mujeres negras.

Finalmente entrenaremos un modelo de regresión lineal que tratará de modelar la relación entre las distintas variables del conjunto de datos y los niveles de estradiol, asignando un peso significativo a cada una y permitiendo realizar predicciones.

Tras estos análisis podremos identificar qué factores tienen una mayor influencia en el nivel de estradiol de las mujeres (y por lo tanto podremos tomar medidas preventivas al ser factor de riesgo).

4.2 Comprobación de la normalidad y homogeneidad de la varianza

La mayoría de análisis estadísticos (entre ellos el contraste de hipótesis y la regresión lineal) asumen la normalidad de los datos. Para comprobar que efectivamente podemos aplicar los distintos métodos propuestos en este apartado sobre los datos realizaremos dos pruebas:

- Test de *Shapiro-Wilk* para la comprobación de la normalidad.
- Test de *Fligner-Killeen* para la comprobación de la homogeneidad de la varianza.

Comenzamos evaluando la normalidad de cada una de las variables cuantitativas. Notar que podríamos haber utilizado otros métodos como podrían ser los *QQ plots* o el test de *Anderson-Darling.*

```
columns_quantitative <- c('Entage', 'Agefbo', 'Agemenar', 'Estradl', 'BMI', 'WHR')

# convert to numerical data type
dataset[columns_quantitative] <- sapply(dataset[columns_quantitative], as.character)
dataset[columns_quantitative] <- sapply(dataset[columns_quantitative], as.numeric)

dataset_quantitative <- dataset[columns_quantitative]
```

```
for (i in 1:ncol(dataset_quantitative)) {
  swt <- shapiro.test(dataset_quantitative[,i])
  print(columns_quantitative[i])
  print(swt)
}
```

```
## [1] "Entage"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.94124, p-value = 2.199e-07
##
## [1] "Agefbo"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.64162, p-value < 2.2e-16
##
## [1] "Agemenar"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.95107, p-value = 1.838e-06
##
## [1] "Estradl"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.98273, p-value = 0.01287
##
## [1] "BMI"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.94501, p-value = 4.844e-07
##
## [1] "WHR"
##
##  Shapiro-Wilk normality test
##
## data:  dataset_quantitative[, i]
## W = 0.93623, p-value = 8.057e-08
```

Un p-value superior al nivel de significancia 0.05 indicará la normalidad de la variable con una confianza superior al 95%, mientras que valores inferiores serán considerados como de no-normalidad.

En este caso comprobamos que ninguno de los atributos del conjunto de datos satisface las condiciones de normalidad. Por lo tanto, previo a la realización de cualquier análisis estadístico, deberemos normalizar las variables.

Notar que el teorema del límite central nos asegurará que para muestras lo suficientemente grandes (unos 30 registros, aproximadamente), la distribución de la media de las muestras tenderá a la de una distribución normal, con lo que pese a los resultados del test podremos seguir aplicando los análisis previamente mencionados.

A continuación aplicaremos el test de *Fligner-Killen* para comprobar la homogeneidad de las varianzas. Es un test no-paramétrico, por lo cual será válido pese a la falta de normalidad de los atributos.

En concreto evaluaremos la homogeneidad de la varianza de los valores de estradiol frente a la etnia de la mujer, ya que será el análisis principal que realizaremos.

```
a <- dataset$Estradl[dataset$Ethnic == 'Caucasian']
b <- dataset$Estradl[dataset$Ethnic == 'African American']
fligner.test(x=list(a,b))
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  list(a, b)
## Fligner-Killeen:med chi-squared = 0.46304, df = 1, p-value = 0.4962
```

Obtenemos un p-valor de 0.5, muy superior al 0.05 establecido de significancia. Por lo tanto podemos afirmar que la varianza es homogénea entre grupos.

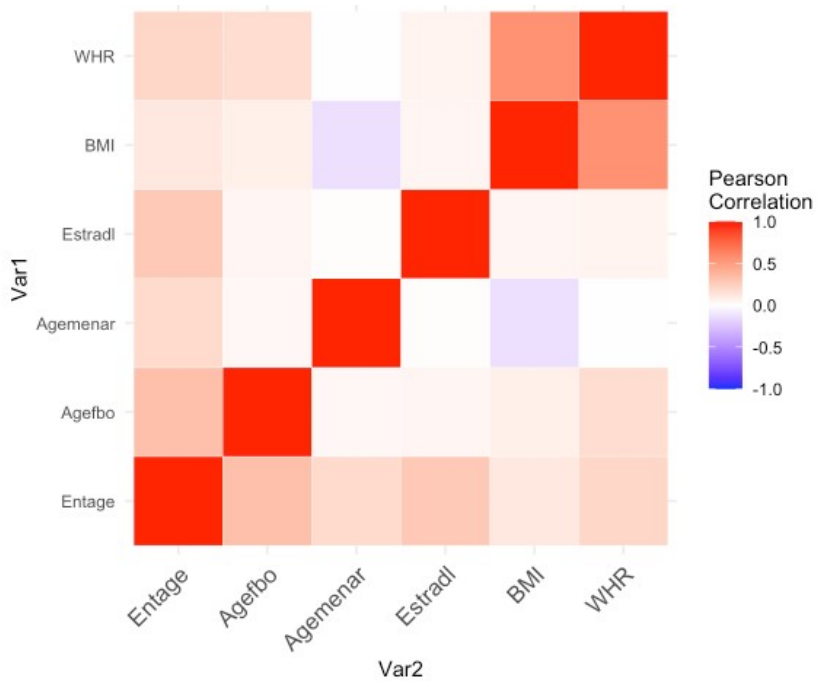
4.3 Análisis de correlaciones

Construimos la matriz de correlaciones para identificar las relaciones cualitativas entre las distintas variables del conjunto de datos. Además de la relación concreta con el nivel de estradiol, de esta manera podremos evaluar también las relaciones a un nivel general.

Utilizaremos como métrica el índice de correlación de Pearson.

```
# Correlation matrix
cormat <- round(cor(dataset_quantitative),2)
melted_cormat <- melt(cormat, na.rm=TRUE)

ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
 geom_tile(color = "white")+
 scale_fill_gradient2(low = "blue", high = "red", mid = "white",
   midpoint = 0, limit = c(-1,1), space = "Lab",
   name="Pearson\nCorrelation") +
  theme_minimal()+
 theme(axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))+
 coord_fixed()
```



Podemos comprobar como en general todas las variables se refuerzan de manera directamente proporcional (el incremento o disminución de una de ellas conduce a la misma variación en otra).

Esta regla presenta una excepción, donde un mayor índice BMI suele estar asociado a una edad de menarquia más temprana.

En relación a los niveles de estradiol podemos comprobar como, pese a no identificar ninguna correlación especialmente alta, la adiposidad, edad, y tardanza de la menarquia están relacionados con mayores niveles de estradiol (constituyendo así un factor de riesgo).

4.4 Contraste de hipótesis

Algunos estudios médicos [4] parecen indicar una influencia directa entre la etnia de las mujeres y los niveles de estradiol en sangre.

En esta sección llevaremos a cabo un contraste estadístico de hipótesis para determinar si la población de mujeres afroamericanas presenta, de media, niveles dispares de estradiol en sangre en relación a las mujeres caucásicas.

En primer lugar definimos como población 1 al conjunto de las mujeres afroamericanas y como población 2 a las mujeres caucásicas. Realizaremos los distintos análisis estadísticos a partir de una muestra de cada una de estas poblaciones (149 en el primero y 56 en el segundo).

```
dataset_afam <- subset(dataset, dataset$Ethnic == 'African American')
dataset_cauc <- subset(dataset, dataset$Ethnic == 'Caucasian')
```

En apartados anteriores hemos comprobado como el nivel de estradiol de estas muestras no obedecen distribuciones normales. Sin embargo, gracias al teorema del límite central podemos asumir, que como ambas muestras cuentan con más de 30 observaciones, la distribución de las medias de las muestras siguen una distribución normal.

4.4.1 Hipótesis nula y alternativa

Definimos la hipótesis nula y la hipótesis alternativa en relación a los parámetros μ_1 y μ_2.

Así, la hipótesis nula será $H_0: \mu_1 - \mu_2 = 0$ y la hipótesis alternativa $H_1: \mu_1 - \mu_2 \neq 0$.

Debido a las características del problema, podemos afirmar que:

- Los parámetros a ser evaluados son medias poblaciones. Así, realizaremos un **contraste de medias**.
- Comparamos parámetros procedentes de dos poblaciones diferentes. Así, realizaremos un **contraste de dos muestras independientes**.
- Debido a que cada una de las muestras cuenta con más de 30 registros, podemos asumir la normalidad en la distribución de las medias muestras por el teorema del límite central. Así, realizaremos un **contraste paramétrico**.
- La hipótesis alternativa indica una **comparación bilateral**.

4.4.2 Cálculos estadísticos

No podemos afirmar que las muestras procedan de una distribución normal. Sin embargo, se cumplen las condiciones del teorema del límite central.

Comenzamos determinando cada una de las medias muestrales.

```
mean(dataset_afam$Estradl)
```

```
## [1] 33.73134
```

```
mean(dataset_cauc$Estradl)
```

```
## [1] 42.99893
```

A continuación determinamos los errores estándar muestrales.

```
sd(dataset_afam$Estradl)
```

```
## [1] 16.51693
```

```
sd(dataset_cauc$Estradl)
```

```
## [1] 18.26891
```

De esta manera, los estadísticos con los que trabajaremos resultan

$$\bar{x}_1 = 33.73134 \qquad s_1 = 16.51693$$
$$\bar{x}_2 = 42.99893 \qquad s_2 = 18.26891$$

Definimos el estadístico de contraste como

$$z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -3.32029$$

El cual es una observación de una variable aleatoria distribuida aproximadamente como una distribución normal N(0,1). Determinamos el p-valor como

```
pnorm(-3.32029)
```

```
## [1] 0.0004496199
```

De esta manera,

$$p = 2P(Z > |z|) = 2P(Z > |-3.32029|) = 2 \cdot 0.0004496199 \approx 0$$

Debido a que buscamos confirmar la hipótesis nula con un 95% de confianza, seleccionamos una significancia de α=0.05.

El intervalo de confianza para la diferencia de medias vendrá dado como

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\, s_{x_1 - x_2}$$

Sustituyendo los valores ya conocidos,

$$-9.26759 \pm z_{\alpha/2} \cdot 2.7912$$

Sabiendo que α/2=0.025, el valor crítico vendrá dado como

```
qnorm(0.025)
```

```
## [1] -1.959964
```

$$z_{\alpha/2} = \pm 1.96$$

Por lo tanto, el intervalo de confianza será (-14.7383, -3.7969).

4.4.3 Interpretación

Dado que el p-valor (0) es inferior a la significancia (0.05), la hipótesis nula es rechazada en favor de la validez de la hipótesis alternativa.

Esta misma conclusión es respaldada mediante el valor crítico al no encontrarse en el intervalo de confianza,

$$-3.32029 \notin (-14.7383, -3.7969)$$

De esta manera queda confirmada la hipótesis alternativa, por la cual los niveles medios de estradiol en las mujeres afroamericanas y caucásicas no son los mismos.

4.5 Modelo de regresión

El último análisis que aplicaremos sobre los datos será la construcción de un modelo de regresión lineal que tendrá el objetivo de predecir el nivel de estradiol de una mujer en función de las variables recogidas anteriormente. De este modo se podrá tener un perfil médico más completo que permita realizar tratamientos preventivos personalizados.

A partir de los resultados del contraste de hipótesis anterior hemos determinado que mujeres de diferente etnia obedecen patrones distintos a la hora de determinar el nivel de estradiol. Por este motivo optamos por, además de construir un modelo global, construir dos modelos independientes, uno para cada etnia. De esta manera trataremos de ajustar mejor las características propias de cada grupo. Seguidamente comprobaremos cual de las dos opciones resulta ser más precisa.

Otras posibles alternativas más avanzadas para la construcción del modelo sería el uso de árboles regresivos de decisión, técnicas de bagging, o incluso recurrir al uso de redes neuronales.

Para evaluar el ajuste del modelo en relación a los datos utilizaremos como métrica el coeficiente de determinación (R^2).

Comenzamos definiendo el modelo más general posible, donde asignaremos un peso a todas las dimensiones posibles.

```
multi.fit <- lm(Estradl~Ethnic+Entage+Numchild+Agefbo+Anykids+Agemenar+BMI+WHR+Area, data=dataset)
summary(multi.fit)
```

```
##
## Call:
## lm(formula = Estradl ~ Ethnic + Entage + Numchild + Agefbo +
##     Anykids + Agemenar + BMI + WHR + Area, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30.91 -12.89  -1.05  10.09  48.52
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.613619  16.773367   0.871 0.384695
## EthnicCaucasian 11.070892   2.790511   3.967 0.000102 ***
## Entage           0.884210   0.235343   3.757 0.000227 ***
## Numchild         1.944093   1.877590   1.035 0.301755
## Agefbo           0.008536   0.412772   0.021 0.983521
## AnykidsYes      -3.012067  10.072712  -0.299 0.765234
## Agemenar        -1.263322   0.887834  -1.423 0.156355
## BMI              0.244746   0.267660   0.914 0.361640
## WHR              4.855293  20.649548   0.235 0.814356
## AreaUrban        3.138135   2.479565   1.266 0.207167
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.34 on 195 degrees of freedom
## Multiple R-squared:  0.163,  Adjusted R-squared:  0.1244
## F-statistic: 4.221 on 9 and 195 DF,  p-value: 5.241e-05
```

El modelo más general posible no parece ajustar correctamente los datos, pues muestra un coeficiente de correlación muy bajo. La inspección de los p-valores parecen indicar que las dos variables más significativas con la etnia y la edad, mientras que el número de hijos no proporciona información de valor. Del mismo modo, el índice WHR y el atributo *Anykids* presentan un error estándar demasiado elevado que pueden perjudicar al modelo.

Probamos de diseñar un modelo lineal más simple que sólo tenga en cuenta la etnia, edad, menarquia y BMI de la mujer.

```
multi.fit <- lm(Estradl~Ethnic+Entage+Agemenar+BMI, data=dataset)
summary(multi.fit)
```

```
##
## Call:
## lm(formula = Estradl ~ Ethnic + Entage + Agemenar + BMI, data = dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.398 -13.905  -2.239  10.406  50.297
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      15.3464    12.7528   1.203     0.23
## EthnicCaucasian  11.1267     2.7027   4.117 5.61e-05 ***
## Entage            0.9402     0.2165   4.343 2.23e-05 ***
## Agemenar         -1.0994     0.8750  -1.256     0.21
## BMI               0.2687     0.2231   1.204     0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.25 on 200 degrees of freedom
## Multiple R-squared:  0.1511, Adjusted R-squared:  0.1341
## F-statistic: 8.899 on 4 and 200 DF,  p-value: 1.24e-06
```

El ajuste del modelo muestra ser muy parecido al obtenido anteriormente a partir de la totalidad de los datos. Así, concluimos que estas cuatro variables llevan la carga semántica del conjunto (pese a no ser suficiente).

Los resultados parecen indicar que no existen relaciones lineales significativas entre los distintos atributos del conjunto de datos y el nivel de estradiol. Por lo tanto, el modelo de regresión lineal parece no ser el análisis más adecuado sobre estos datos.

Otros factores no considerados como podrían ser la alimentación, la actividad física, o la toma de drogas podrían ser mucho más significativos en determinar el nivel de estradiol.

Tratamos de utilizar el modelo obtenido para comprobar si, pese a no ser un buen ajuste, permite obtener una idea cualitativa de los niveles de estradiol, cogiendo para ello dos registros aleatorios.

```
kable(sample_n(dataset, 2))
```

Id	Estradl	Ethnic	Entage	Numchild	Agefbo	Anykids	Agemenar	BMI	WHR	Area
48	49.5	Caucasian	25	0	0	No	11	27.0953	0.76	Rural
3340	22.6	African American	34	1	20	Yes	11	20.4261	0.74	Rural

Para el **registro 1**, el **modelo 1** da un valor de estradiol de 44.21 y el **modelo 2** de 45.16

Para el **registro 2**, el **modelo 1** da un valor de estradiol de 38.47 y el **modelo 2** de 40.68

En efecto, comprobamos que aunque los modelos han sido de utilidad para predecir el registro 1, la propia varianza del modelo hace que proponga una solución sobreestimada para el registro 2.

A pesar de ello, este modelo puede ser de utilidad para obtener una idea cualitativa aproximada del valor del estradiol en una mujer con las características dadas.

5. Representación de los resultados

A lo largo de la práctica se han mostrado distintas gráficas y tablas con el objetivo de complementar la fase de exploración previa al preprocesado y los distintos análisis llevado a cabo.

En esta sección se plantea complementar el material ya especificado con aquellas visualizaciones omitidas en apartados anteriores, y que pueden ayudar a conocer mejor los datos.

5.1 Métricas robustas vs no-robustas

Determinamos la tendencia central y la dispersión para las distintas variables cuantitativas. Para ello podemos utilizar métricas no-robustas (sensibles a la presencia de valores extremos) o métricas robustas.

La tendencia central se determinará de forma no-robusta por la media, y de forma robusta por la mediana. La dispersión se determinara de forma no-robusta por la desviación estándar, y de forma robusta por desviación sobre la mediana.

```
kable(sapply(dataset_quantitative, mean))
```

	x
Entage	26.048780
Agefbo	6.443902
Agemenar	12.348781
Estradl	36.262976
BMI	25.911943
WHR	0.759561

```
kable(sapply(dataset_quantitative, sd))
```

	x
Entage	5.4121831
Agefbo	10.2207509
Agemenar	1.3608896
Estradl	17.4653735
BMI	5.3993507
WHR	0.0690503

```
kable(sapply(dataset_quantitative, median))
```

	x
Entage	26.0000
Agefbo	0.0000
Agemenar	12.0000
Estradl	35.6000
BMI	24.7151
WHR	0.7400

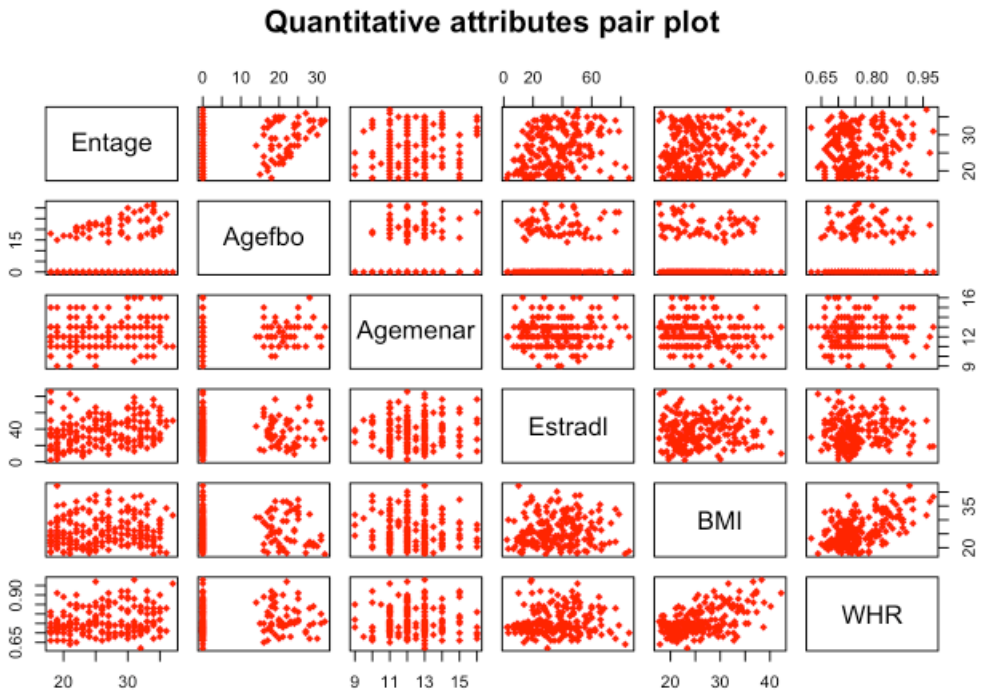
```
kable(sapply(dataset_quantitative, mad))
```

	x
Entage	7.413000
Agefbo	0.000000
Agemenar	1.482600
Estradl	18.680760
BMI	5.371460
WHR	0.059304

Destacar como los valores del atributo Agefbo parecen carecer de sentido. Sin embargo, estos valores representan mayoritariamente registros de mujeres sin hijos (representado pro el valor 0). Debido a su importancia semántica decidimos tenerlos en cuenta, teniendo en mente que podrían ser temporalmente eliminados para obtener medias centrales y dispersiones de ser requeridas.

5.2 Relaciones bivariables

En el apartado 4.5 hemos comprobado que no era posible construir un modelo lineal para tratar de predecir el nivel de estradiol en función del resto de variables. Para tener una mayor intuición sobre las interrelaciones bivariables del conjunto de los datos mostramos un *pair plot*.

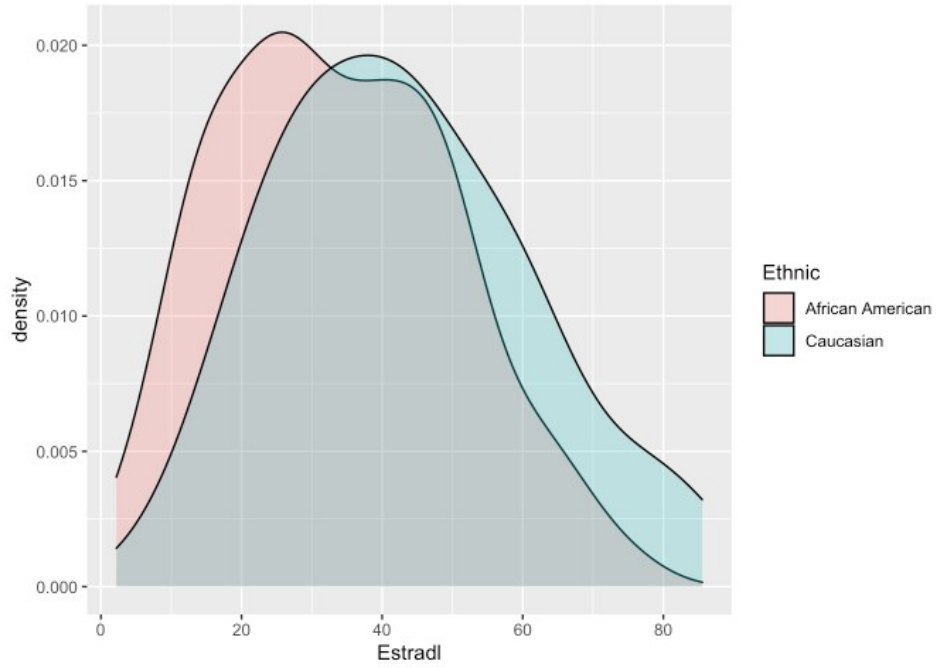


Efectivamente comprobamos la ausencia de relaciones lineales entre las distintas variables, lo que explica el mal ajuste del modelo diseñado anteriormente. Destaca únicamente la relación entre el índice BMI y el índice WHR, que al ser los dos medidas de la adiposidad tiene sentido que compartan una relación lineal.

5.2 Distribución de *serum estradiol* según etnia

En el apartado 4.4 hemos encontrado mediante un contraste de hipótesis que el valor medio de los niveles de estradiol depende en gran medida de la etnia de la mujer. Se muestra a continuación una representación comparativa de los valores de estradiol para cada una de las etnias presentes en el conjunto de datos con el objetivo de poder visualizar cualitativamente este hecho.

```
comb_r <- rbind(dataset_afam, dataset_cauc)
ggplot(comb_r, aes(Estradl, fill = Ethnic)) + geom_density(alpha = 0.25)
```



Efectivamente la distribución de la concentración de estradiol posee medias diferentes en función de la etnia de la mujer, siendo superior para la etnia caucásica.

6. Conclusiones

A lo largo de esta práctica se ha explorado el ciclo de vida de los datos, desde la fase de planificación del proyecto de ciencia de datos pasando por la selección de datos, limpieza y análisis.

En la fase de limpieza se han identificado distintos tipos de inconsistencias en los datos, ya sea desde el punto de vista de formato o desde la coherencia lógica. Debido a las decisiones tomadas, estos valores se han tratado como valores faltantes. Tras identificar los valores extremos para algunos atributos, se ha procedido a la imputación de valores sintéticos mediante técnicas que permitieran mantener la coherencia estadística.

Finalmente en la fase de análisis se han conducido distintos tipos de análisis sobre los datos que han permitido estudiar la influencia de distintas dimensiones sobre la presencia de estradiol en sangre, siendo la etnia una de las características fundamentales.

Los resultados obtenidos en el conjunto de la práctica podrían ser un inicio para el diseño de tratamientos preventivos personalizados con el objetivo de reducir los factores de riesgo principales.

7. Código

El código desarrollado para la realización de esta práctica se encuentra adjunto en este repositorio.

8. Referencias

[1] "Fundamentals of Biostatistics", 8th Edition; CengageBrain. ESTRADL DAT dataset.

[2] BMI calculator. National Heart, Lung, and Blood Institute.

[3] Waste-hip-ratio. Wikipedia.

[4] Sanchez SS, Tachachartvanich P, Stanczyk FZ, Gomez SL, John EM, Smith MT, Fejerman L. Estrogenic activity, race/ethnicity, and Indigenous American ancestry among San Francisco Bay Area women. PLoS One. 2019 Mar 25;14(3):e0213809. doi: 10.1371/journal.pone.0213809. PMID: 30908519; PMCID: PMC6433244.

9. Responsabilidades

Esta práctica ha sido realizada por completo de manera individual por **Javier Cantero Lorenzo** con la autorización del profesor Jose Moreira Sanchez y la profesora responsable a día 04 de Diciembre de 2020.

Contribuciones	Firma
Investigación previa	JCL
Redacción de las respuestas	JCL
Desarrollo del código	JCL