Trabajo 3: AJUSTE DE MODELOS LINEALES

Nuria Rodríguez Barroso, Juan Luis Suárez Díaz.

26 de mayo de 2017

Clasificación.

Comprensión del problema a resolver.

Para el problema de clasificación hemos elegido la base de datos de *South African Heart Disease*, que almacena una muestra recapituladora de hombres en alto riesgo de cardiopatía en la región de Western Cape, en Sudáfrica. Hay básicamente dos casos de CHD (0 ó 1). Algunos de los hombres que tienen CHD positivo, se han sometido a un tratamiento de reducción de la presión sanguínea, siendo los datos que aparecen en la muestra posteriores a estos tratamientos. Estos datos datan de 1983.

Nuestra base de datos consta de 462 muestras donde cada una de ellas cuenta con 10 atributos, uno de ellos la variable de respuesta. Las diferentes atributos a tratar son:

- **sbp:** presión arterial sistólica. Toma valores entre 101 y 218, siendo la media 138.3.
- **tobacco:** tabaco acumulativo (en kg). Toma como valor mínimo 0 y máximo 31.2. En este caso la media es de 3.6356.
- **ldl:** Lipoproteína de baja densidad (colesterol). Toma valores entre 0.98 y 15.330, siendo la media 4.74.
- **adiposidad:** adiposi Tomando valores entre 6.74 y 42.49, siendo la media de los valores 25.41.
- **famhist:** historial familiar de cardiopatías. Toma como valores {Present, Absent}, habiendo del primer tipo 270 y del segundo 192.
- **typea:** Personalidad Tipo-A (mide el grado de estrés en el día a día). Toma valores entre 13 y 78, siendo la media 53.1.
- **obesity:** Obesidad. Toma valores entre 14.7 y 46.58, siendo la media 26.04.
- **alcohol:** Actual consumición de alcohol. Toma valores entre 0 y 147.19, encontrándose el valor medio en 17.04.
- **age:** Edad de los hombres al comienzo de las pruebas. Se encuentra entre 15 y 64 años, siendo la edad media 42.82.
- **chd:** Variable de respuesta, indica si se tiene o no alguna cardiopatía. Toma como valores {0,1}, siendo la media 0.3463.

```
##       sbp           tobacco             ldl           adiposity
##  Min.   :101.0   Min.   : 0.0000   Min.   : 0.980   Min.   : 6.74
##  1st Qu.:124.0   1st Qu.: 0.0525   1st Qu.: 3.283   1st Qu.:19.77
##  Median :134.0   Median : 2.0000   Median : 4.340   Median :26.11
##  Mean   :138.3   Mean   : 3.6356   Mean   : 4.740   Mean   :25.41
##  3rd Qu.:148.0   3rd Qu.: 5.5000   3rd Qu.: 5.790   3rd Qu.:31.23
##  Max.   :218.0   Max.   :31.2000   Max.   :15.330   Max.   :42.49
##     famhist        typea          obesity         alcohol
##  Absent :270   Min.   :13.0   Min.   :14.70   Min.   :  0.00
##  Present:192   1st Qu.:47.0   1st Qu.:22.98   1st Qu.:  0.51
##                Median :53.0   Median :25.80   Median :  7.51
##                Mean   :53.1   Mean   :26.04   Mean   : 17.04
##                3rd Qu.:60.0   3rd Qu.:28.50   3rd Qu.: 23.89
##                Max.   :78.0   Max.   :46.58   Max.   :147.19
##       age              chd
##  Min.   :15.00   Min.   :0.0000
##  1st Qu.:31.00   1st Qu.:0.0000
```

```
##  Median :45.00   Median :0.0000
##  Mean   :42.82   Mean   :0.3463
##  3rd Qu.:55.00   3rd Qu.:1.0000
##  Max.   :64.00   Max.   :1.0000
```

Preprocesado de datos.

Como podemos observar en el breve estudio de los diferentes atributos que vamos a trabajar, cada uno toma valores en una franja muy diferente, lo que hace primar unos atributos sobre otros en los métodos basados en distancias. Además, algunos atributos presentan una gran asimetría, lo cual también es conveniente evitar. Por estos motivos y otros más que estudiaremos a continuación, tenemos que preprocesar los datos.

Modificación de los atributos cualitativos.

Tenemos que convertir los atributos cualitativos en atributos numéricos para que las funciones que usemos más adelante puedan trabajar con ellos. En nuestro problema concreto, solo contamos con un atributo cualitativo: *famhist*, que toma los valores {Present, Absent}. Convertimos este atributo en el atributo *present_famhist*, que tomará el valor 1, cuando *famhist* tomaba el valor Present y 0 en el otro caso. Así, el atributo *present_famhist* tomará valores en {0,1}.

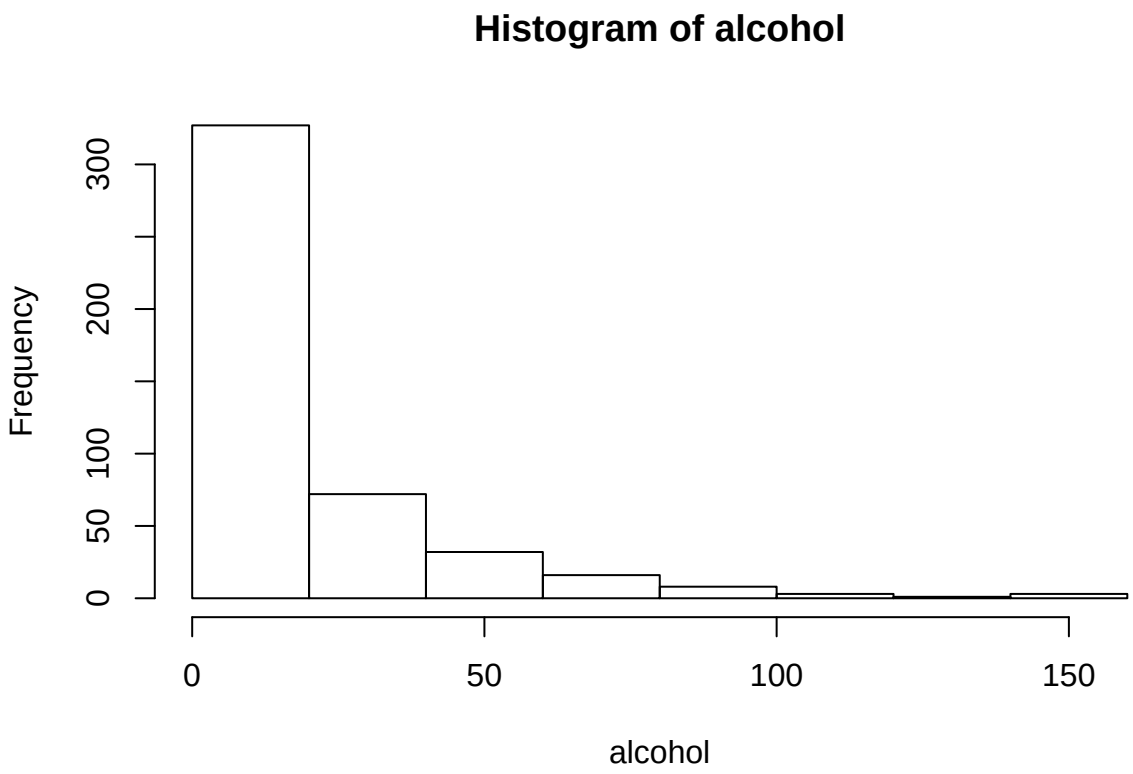
Para el resto de pasos, utilizaremos una función llamada *preProcess()*, la cual terminará con el preprocesado de los datos. Esta función realizará los siguientes cambios:

Tratamiento de la asimetría con BoxCox.

Como ya hemos comentado anteriormente, hay varios atributos que presentan una alta asimetría, lo cual podría hacer que los métodos de predicción que apliquemos a continuación obtengan resultados peores. Para solucionar esto, utilizaremos un método llamado BoxCox. Este método se basa en la transformación potencial según un valor (λ) para aumentar la correlación entre las variables. Para elegir la mejor potencia (mejor λ), se busca entre los λ que proporcionen un menor error residual. Aunque en la práctica, esto se realizará de manera automátcica con la función *preProcess()* vamos a ver cómo funciona en el caso de un atributo. Para que se vea mejor el funcionamiento, vamos a elegir el atributo que presente una mayor asimetría. Para ello, ordenamos los atributos en función de su asimetría:

```
##         alcohol         tobacco             ldl             sbp
##       2.2977031       2.0657278       1.3045896       1.1729355
##         obesity             chd             age           typea
##       0.8993498       0.6438924       0.3792590       0.3441914
## present_famhist       adiposity
##       0.3414684       0.2132541
```

Si dibujamos el histograma correspondiente al atributo con mayor asimetría, *alcohol* corroboramos que los datos se encuentran muy concentrados en los primeros valores que este atributo toma.



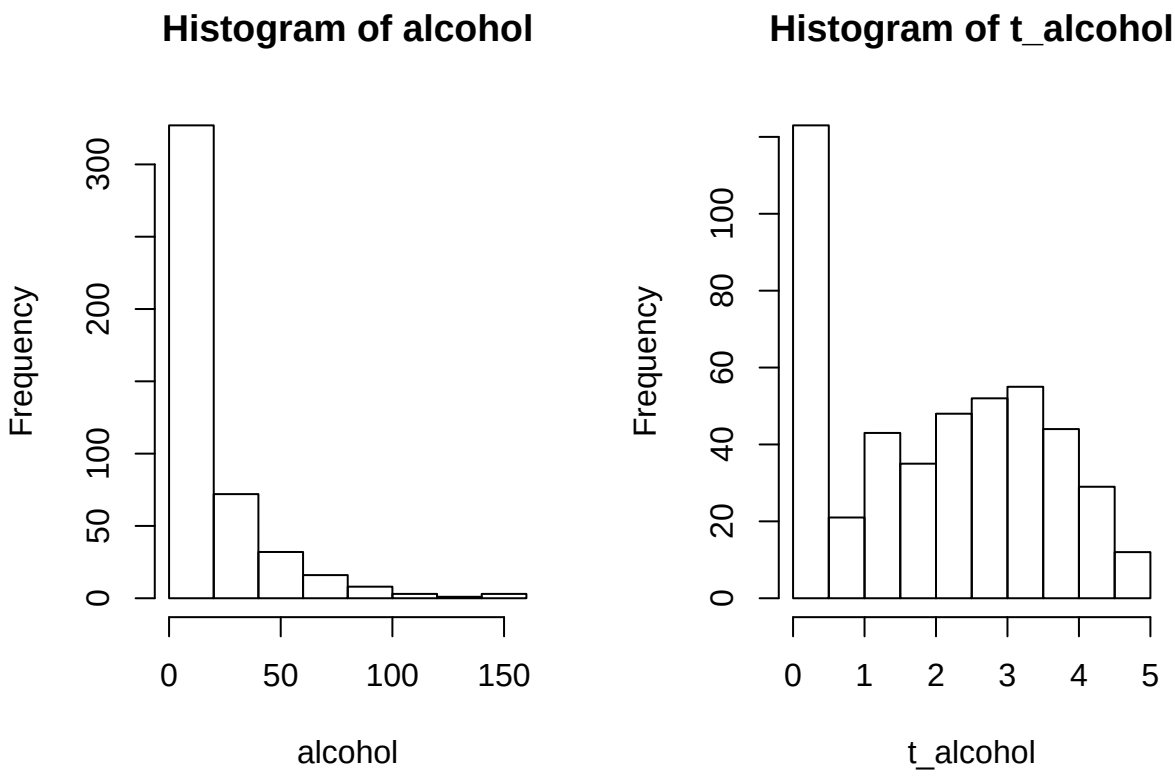
Vamos a aplicar ahora el método *BoxCox* y veremos cómo mejora la simetría del atributo.

```
## Box-Cox Transformation
##
## 462 data points used to estimate Lambda
##
## Input data summary:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.51    7.51   17.04   23.89  147.20
##
## Lambda could not be estimated; no transformation is applied
```

Como observamos, no se encuentra un λ válido para realizar la transformación, y esto se debe a que entre los valores que toma el atributo se encuentra el 0, punto en el cual no está definida la función logaritmo. Para solucionar esto, realizamos una translación de los datos de la forma: datos = min(datos) + 1 + datos, para así conseguir que el mínimo de los datos se desplace a 1. Tras realizar esta transformación obtenemos:

```
## [1] "La asimetría del alcohol transformado es:"
```

```
## [1] 0.02949713
```



Para aplicar esta transformación a todos los atributos con el λ correspondiente, pasaremos como parámetro al método *preProcess* que realice el método *BoxCox* y todo esto se hará de forma automática.

Eliminación de atributos con PCA.

El algoritmo PCA (Principal Components Analysis) es un filtro no supervisado que es de gran utilidad cuando disponemos de una base de datos con un gran número de atributos, entre los que algunos pueden ser redundantes o irrelevantes. Como nuestra base de datos solo consta de 10 atributos, no es necesaria la aplicación de este método.

Centrar y escalar.

No me queda muy claro como funciona pues en internet encuentro cosas raras <- JUANLU ME FIO MASD E TI.

También es conveniente centrar y escalar las variables para que no prioricen unas sobre otras, dado que para métodos que utilizaremos más adelante (regsubsets?? mirar esto) es necesario. El método *preProcess* se encargará de centrar las variables en 0 y de escalar las variables para tener varianza unitaria. Para ello, bastaría con pasar como argumento *scale* y *center* cuando llamemos al método *preProcess*.

Llamada al método PreProcess.

Una vez entendidas las modificaciones que vamos a realizar a los datos, utilizaremos el método preProcess que se encarga de realizar todas estas modificaciones sobre el conjunto de datos pasado como argumento. Como ya hemos comentado, no vamos a aplicar el método *PCA*, luego la llamada quedaría de la siguiente forma:

```
ObjetoTrans = preProcess(sahd[,names(sahd)!="chd"],method = c("BoxCox","center","scale"))
sahdTrans <- predict(ObjetoTrans,sahd)
```

Los conjuntos de validación, training y test usados.

Regresión.

Para el problema de regresión hemos elegido la base de datos de *Los Ángeles Ozone*, la cual se centra en medir el nivel de concentración de ozono en la atmósfera. Para ello, se realizaban 8 mediciones hechas diariamente en Los Ángeles durante el año 1976. Aunque la idea era obtener el nivel de ozono para todos los días del año, algunos datos se han perdido así que no contiene todos los días del año (en concreto contiene 330 días). La base de datos consta de 10 atributos que son los siguientes:

- **ozone**: Es la variable de respuesta, mide la elevación máxima del ozono.
- **vh**: Vandenberg 500 mb Height
- **wind**: Velocidad del viento, medida en mph.
- **humidity:** Tanto por ciento de humedad.
- **temp:** Temperatura
- **ibh:**
- **dpg:** Gradiente de presión de Daggot.
- **ibt:**
- **vis:** La visibilidad medida en millas.
- **day:** Día del año en el que se realizó la medición.