Práctica 2: Limpieza y validación de los datos

Teguayco Gutiérrez González

6 de diciembre de 2017

Índice

1. Detalles de la actividad

1.1. Descripción

En esta actividad se elabora un caso práctico, consistente en el tratamiento de un conjunto de datos (en inglés, *dataset*), orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

1.2. Objetivos

Los objetivos que se persiguen mediante el desarrollo de esta actividad práctica son los siguientes:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que permita continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

1.3. Competencias

Así, las competencias del Máster en *Data Science* que se desarrollan son:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

2. Resolución

2.1. Descripción del dataset

El conjunto de datos objeto de análisis se ha obtenido a partir de este enlace en *Kaggle* y está constituido por 26 características (columnas) que presentan 206 coches (filas o registros). Entre los campos de este conjunto de datos, encontramos los siguientes:

- **symboling:** índice de denota el riesgo del vehículo en términos de aseguración. Un valor de +3 indica que el vehículo tiene riesgo, mientras que un valor de -3 indica que es bastante seguro.
- **normalized-losses:** promedio del pago de pérdida de un vehículo asegurado por año.
- **make:** marca (Alfa Romeo, Honda, Chevrolet, etc).
- **fuel-type:** tipo de combustible: gasolina (*gas*) o gasóil (*diesel*).
- **aspiration:** tipo de aspiración del motor: turbo o estándar.
- **num-of-doors**: número de puertas.
- **body-style:** estilo (*hatchback*, sedán, familiar...).
- **drive-wheels**: tracción en las ruedas: trasera (*rwd*), delantera (*fwd*) o en las cuatro ruedas (*4wd*).
- **engine-location**: ubicación del motor: parte delantera (*front*) o trasera (*rear*).
- **wheel-base**: distancia entre los ejes delantero y trasero en centímetros.
- **length**: longitud del vehículo en centímetros.
- **width**: ancho del vehículo en centímetros.
- **height**: altura del vehículo en centímetros.
- **curb-weight**: peso del vehículo sin ocupantes.
- **engine-type**: tipo de motor (*SOHC*, *DOHC*, etc).
- **num-of-cylinders**: número de cilindros.
- **engine-size**: tamaño del motor.
- **fuel-system**: tipo de sistema de combustión.
- **bore**: diámetro del cilindro.
- **stroke**: carrera del pistón (distancia recorrida por el pistón en cada ciclo del motor).
- **compression-ratio**: índice de compresión.
- **horsepower**: potencia del motor en caballos de vapor.

- **peak-rpm**: revoluciones por minuto del motor a máxima potencia.
- **city-mpg**: consumo en ciudad.
- **highway-mpg**: consumo en autopista.
- **price**: precio del coche.

2.2. Importancia y objetivos de los análisis

A partir de este conjunto de datos se plantea la problemática de determinar qué variables influyen más sobre el precio de un automóvil. Además, se podrá proceder a crear modelos de regresión que permitan predecir el precio de un coche en función de sus características y contrastes de hipótesis que ayuden a identificar propiedades interesantes en las muestras que puedan ser inferidas con respecto a la población.

Estos análisis adquieren una gran relevancia en casi cualquier sector relacionado con la automoción. Un ejemplo de ello se puede observar en el servicio de peritaje interno en una compañía de alquier de coches. En este caso, el perito, encargado de realizar informes técnicos que recogen la valoración económica de los coches, podría valerse de los análisis que se plantean en esta actividad para utilizarlos como soporte a la hora de llevar a cabo las tasaciones.

2.3. Limpieza de los datos

Antes de comenzar con la limpieza de los datos, procedemos a realizar la lectura del fichero en formato CSV en el que se encuentran. El resultado devuelto por la llamada a la función `read.csv()` será un objeto `data.frame`:

```
# Lectura de datos
automoviles <- read.csv("Automobile_data.csv", header = TRUE)
head(automoviles[,1:5])
```

```
##   symboling normalized.losses        make fuel.type aspiration
## 1         3                 NA alfa-romero       gas        std
## 2         3                 NA alfa-romero       gas        std
## 3         1                 NA alfa-romero       gas        std
## 4         2                164        audi       gas        std
## 5         2                164        audi       gas        std
## 6         2                 NA        audi       gas        std
```

```
# Tipo de dato asignado a cada campo
sapply(automoviles, function(x) class(x))
```

```
##         symboling normalized.losses              make         fuel.type
##         "integer"         "integer"          "factor"          "factor"
##        aspiration      num.of.doors        body.style      drive.wheels
##          "factor"          "factor"          "factor"          "factor"
##   engine.location        wheel.base            length             width
##          "factor"         "numeric"         "numeric"         "numeric"
##            height       curb.weight       engine.type  num.of.cylinders
##         "numeric"         "integer"          "factor"          "factor"
##       engine.size       fuel.system              bore            stroke
##         "integer"          "factor"         "numeric"         "numeric"
## compression.ratio        horsepower          peak.rpm          city.mpg
##         "numeric"         "integer"         "integer"         "integer"
##       highway.mpg             price
##         "integer"         "integer"
```

Además, observamos cómo los tipos de datos asignados automáticamente por R a las variables se corresponden con el dominio de estas.

Nota. Originalmente, los valores desconocidos eran denotados en el dataset mediante el caracter '?'. Por ello, se ha realizado una sustitución de estos valores por una cadena vacía previa a la lectura para que R marque estos valores desconocidos como NA (del inglés, *Not Available*). Esto simplificará el manejo de los datos en los apartados posteriores.

2.3.1. Selección de los datos de interés

La gran mayoría de los atributos presentes en el conjunto de datos se corresponden con características que reúnen los diversos automóviles recogidos en forma de registros, por lo que será conveniente tenerlos en consideración durante la realización de los análisis. Sin embargo, podemos prescindir de los dos primeros campos (*symboling* y *normalized.losses*) dado que no son atributos técnicos de los coches en sí y, por tanto, nos resultan menos relevantes a la hora de resolver nuestro problema.

```
# Eliminar las dos primeras columnas
automoviles <- automoviles[, -(1:2)]
```

2.3.2. Ceros y elementos vacíos

Comúnmente, se utilizan los ceros como centinela para indicar la ausencia de ciertos valores. Sin embargo, no es el caso de este conjunto de datos puesto que, como se comentó durante el apartado relativo a la lectura, se utilizó el caracter '?' para denotar un valor desconocido. Así, se procede a conocer a continuación qué campos contienen elementos vacíos:

```
# Números de valores desconocidos por campo
sapply(automoviles, function(x) sum(is.na(x)))
```

```
##              make         fuel.type        aspiration      num.of.doors
##                 0                 0                 0                 0
##        body.style      drive.wheels   engine.location        wheel.base
##                 0                 0                 0                 0
##            length             width            height       curb.weight
##                 0                 0                 0                 0
##       engine.type  num.of.cylinders       engine.size       fuel.system
##                 0                 0                 0                 0
##              bore            stroke compression.ratio        horsepower
##                 4                 4                 0                 2
##          peak.rpm          city.mpg       highway.mpg             price
##                 2                 0                 0                 4
```

Llegados a este punto debemos decidir cómo manejar estos registros que contienen valores desconocidos para algún campo. Una opción podría ser eliminar esos registros que incluyen este tipo de valores, pero ello supondría desaprovechar información.

Como alternativa, se empleará un método de imputación de valores basado en la similitud o diferencia entre los registros: la imputación basada en k vecinos más próximos (en inglés, *kNN-imputation*). La elección de esta alternativa se realiza bajo la hipótesis de que nuestros registros guardan cierta relación. No obstante, es mejor trabajar con datos "aproximados" que con los propios elementos vacíos, ya que obtendremos análisis con menor margen de error.

```
# Imputación de valores mediante la función kNN() del paquete VIM
suppressWarnings(suppressMessages(library(VIM)))

automoviles$bore        <- kNN(automoviles)$bore
automoviles$stroke      <- kNN(automoviles)$stroke
automoviles$horsepower  <- kNN(automoviles)$horsepower
automoviles$peak.rpm    <- kNN(automoviles)$peak.rpm
automoviles$price       <- kNN(automoviles)$price

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

```
##              make         fuel.type        aspiration      num.of.doors
##                 0                 0                 0                 0
##        body.style      drive.wheels   engine.location        wheel.base
##                 0                 0                 0                 0
##            length             width            height       curb.weight
##                 0                 0                 0                 0
##       engine.type  num.of.cylinders       engine.size       fuel.system
##                 0                 0                 0                 0
##              bore            stroke compression.ratio        horsepower
##                 0                 0                 0                 0
##          peak.rpm          city.mpg       highway.mpg             price
##                 0                 0                 0                 0
```

2.3.3. Valores extremos

Los valores extremos o *outliers* son aquellos que parecen no ser congruentes sin los comparamos con el resto de los datos. Para identificarlos, podemos hacer uso de dos vías: (1) representar un diagrama de caja por cada variable y ver qué valores distan mucho del rango intercuartílico (la caja) o (2) utilizar la función `boxplots.stats()` de R, la cual se emplea a continuación. Así, se mostrarán sólo los valores atípicos para aquellas variables que los contienen:

```
boxplot.stats(automoviles$wheel.base)$out
```

```
## [1] 115.6 115.6 120.9
```

```
boxplot.stats(automoviles$length)$out
```

```
## [1] 141.1
```

```
boxplot.stats(automoviles$width)$out
```

```
## [1] 71.4 71.4 71.4 71.7 71.7 71.7 72.0 72.3
```

```
boxplot.stats(automoviles$engine.size)$out
```

```
##  [1] 209 209 209 258 258 326 234 234 308 304
```

```
boxplot.stats(automoviles$stroke)$out
```

```
##  [1] 3.90 4.17 4.17 2.19 2.19 3.90 3.90 2.07 2.36 2.64 2.64 2.64 2.64 2.64
## [15] 2.64 2.64 2.64 2.64 2.64 2.64
```

```
boxplot.stats(automoviles$compression.ratio)$out
```

```
##  [1]  7.0  7.0 11.5 22.7 22.0 21.5 21.5 21.5 21.5  7.0  7.0  7.0 21.9 21.0
## [15] 21.0 21.0 21.0 21.0  7.0  7.0 22.5 22.5 22.5 23.0 23.0 23.0 23.0 23.0
```

```
boxplot.stats(automoviles$horsepower)$out
```

```
## [1] 262 200 207 207 207 288
```

```
boxplot.stats(automoviles$peak.rpm)$out
```

```
## [1] 6600 6600
```

```
boxplot.stats(automoviles$city.mpg)$out
```

```
## [1] 47 49
```

```
boxplot.stats(automoviles$highway.mpg)$out
```

```
## [1] 53 54 50
```

```
boxplot.stats(automoviles$price)$out
```

```
##  [1] 30760 41315 36880 32250 35550 36000 31600 34184 35056 40960 45400
## [12] 32528 34028 37028
```

No obstante, si revisamos los anteriores datos para varios coches escogido aleatoriamente de esta web, comprobamos que son valores que perfectamente pueden darse (hay coches que exceden los 40.000€ y otros que llegan a los 300CV, entre otros). Es por ello que el manejo de estos valores extremos consistirá en simplemente dejarlos como actualmente están recogidos.

2.3.4. Exportación de los datos preprocesados

Una vez que hemos acometido sobre el conjunto de datos inicial los procedimientos de integración, validación y limpieza anteriores, procedemos a guardar estos en un nuevo fichero denominado `Automobile_data_clean.csv`:

```
# Exportación de los datos limpios en .csv
write.csv(automoviles, "Automobile_data_clean.csv")
```

2.4. Análisis de los datos

2.4.1. Selección de los grupos de datos a analizar

A continuación, se seleccionan los grupos dentro de nuestro conjunto de datos que pueden resultar interesantes para analizar y/o comparar. No obstante, como se verá en el apartado consistente en la realización de pruebas estadísticas, no todos se utilizarán.

```
# Agrupación por tipo de combustible
automoviles.diesel    <- automoviles[automoviles$fuel.type == "diesel",]
automoviles.gasolina  <- automoviles[automoviles$fuel.type == "gas",]

# Agrupación por tipo de aspiración del motor
automoviles.std     <- automoviles[automoviles$aspiration == "std",]
automoviles.turbo   <- automoviles[automoviles$aspiration == "turbo",]

# Agrupación por estilo de coche
automoviles.convertible <-
    automoviles[automoviles$body.style == "convertible",]
automoviles.hatchback   <-
    automoviles[automoviles$body.style == "hatchback",]
automoviles.sedan       <-
    automoviles[automoviles$body.style == "sedan",]
automoviles.wagon       <-
    automoviles[automoviles$body.style == "wagon",]
automoviles.hardtop     <-
    automoviles[automoviles$body.style == "hardtop",]
```

2.4.2. Comprobación de la normalidad y homogeneidad de la varianza

Para la comprobación de que los valores que toman nuestras variables cuantitativas provienen de una población distribuida normalmente, utilizaremos la prueba de normalidad de *Anderson-Darling*.

Así, se comprueba que para que cada prueba se obtiene un p-valor superior al nivel de significación prefijado $\alpha = 0,05$. Si esto se cumple, entonces se considera que variable en cuestión sigue una distribución normal.

```r
library(nortest)

alpha = 0.05
col.names = colnames(automoviles)

for (i in 1:ncol(automoviles)) {
  if (i == 1) cat("Variables que no siguen una distribución normal:\n")
  if (is.integer(automoviles[,i]) | is.numeric(automoviles[,i])) {
    p_val = ad.test(automoviles[,i])$p.value
    if (p_val < alpha) {
      cat(col.names[i])

      # Format output
      if (i < ncol(automoviles) - 1) cat(", ")
      if (i %% 3 == 0) cat("\n")
    }
  }
}
```

```
## Variables que no siguen una distribución normal:
## wheel.base, length,
## width, height, curb.weight,
## engine.size,
## bore, stroke,
## compression.ratio, horsepower, peak.rpm,
## city.mpg, highway.mpgprice
```

Seguidamente, pasamos a estudiar la homogeneidad de varianzas mediante la aplicación de un **test de Fligner-Killeen**. En este caso, estudiaremos esta homogeneidad en cuanto a los grupos conformados por los vehículos que presentan un motor turbo frente a un motor estándar. En el siguiente test, la hipótesis nula consiste en que ambas varianzas son iguales.

```r
fligner.test(price ~ aspiration, data = automoviles)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  price by aspiration
## Fligner-Killeen:med chi-squared = 1.1626, df = 1, p-value = 0.2809
```

Puesto que obtenemos un p-valor superior a 0,05, aceptamos la hipótesis de que las varianzas de ambas muestras son homogéneas.

2.5. Pruebas estadísticas

2.5.1. ¿Qué variables cuantitativas influyen más en el precio?

En primer lugar, procedemos a realizar un análisis de **correlación** entre las distintas variables para determinar cuáles de ellas ejercen una mayor influencia sobre el precio final del vehículo. Para ello, se utilizará el coeficiente de correlación de Spearman, puesto que hemos visto que tenemos datos que no siguen una distribución normal.

```
corr_matrix <- matrix(nc = 2, nr = 0)
colnames(corr_matrix) <- c("estimate", "p-value")

# Calcular el coeficiente de correlación para cada variable cuantitativa
# con respecto al campo "precio"
for (i in 1:(ncol(automoviles) - 1)) {
  if (is.integer(automoviles[,i]) | is.numeric(automoviles[,i])) {
    spearman_test = cor.test(automoviles[,i],
                             automoviles[,length(automoviles)],
                             method = "spearman")
    corr_coef = spearman_test$estimate
    p_val = spearman_test$p.value

    # Add row to matrix
    pair = matrix(ncol = 2, nrow = 1)
    pair[1][1] = corr_coef
    pair[2][1] = p_val
    corr_matrix <- rbind(corr_matrix, pair)
    rownames(corr_matrix)[nrow(corr_matrix)] <- colnames(automoviles)[i]
  }
}
```

```r
print(corr_matrix)
```

```
##                      estimate      p-value
## wheel.base         0.68622120 7.046367e-30
## length             0.81482985 5.832096e-50
## width              0.81237495 1.940939e-49
## height             0.26025135 1.641690e-04
## curb.weight        0.91195171 1.763348e-80
## engine.size        0.82786575 7.187961e-53
## bore               0.63269667 2.530905e-24
## stroke             0.11747234 9.344938e-02
## compression.ratio -0.19349368 5.439457e-03
## horsepower         0.84855440 4.985180e-58
## peak.rpm          -0.07853931 2.629842e-01
## city.mpg          -0.83299971 4.394043e-54
## highway.mpg       -0.83199551 7.647096e-54
```

Así, identificamos cuáles son las variables más correlacionadas con el precio en función de su proximidad con los valores -1 y +1. Teniendo esto en cuenta, queda patente cómo la variable más relevante en la fijación del precio es el peso en vacío del vehículo (*curb.weight*).

Nota. Para cada coeficiente de correlación se muestra también su p-valor asociado, puesto que éste puede dar información acerca del peso estadístico de la correlación obtenida.

2.5.2. ¿El precio del coche es superior en caso de disponer de un motor turbo?

La segunda prueba estadística que se aplicará consistirá en un contraste de hipótesis sobre dos muestras para determinar si el precio del coche es superior dependiendo del tipo de motor del que se trate (estándar o turbo). Para ello, tendremos dos muestras: la primera de ellas se corresponderá con los precios de los coches con motor estándar y, la segunda, con aquellos que presentan un motor turbo.

Se debe destacar que un test paramétrico como el que a continuación se utiliza necesita que los datos sean normales, si la muestra es de tamaño inferior a 30. Como en nuestro caso, n $>$ 30, el contraste de hipótesis siguiente es válido (aunque podría utilizarse un test no paramétrico como el de Mann-Whitney, que podría resultar ser más eficiente para este caso).

```
automoviles.std.precios   <-
    automoviles[automoviles$aspiration == "std",]$price
automoviles.turbo.precios <-
    automoviles[automoviles$aspiration == "turbo",]$price
```

Así, se plantea el siguiente **contraste de hipótesis de dos muestras sobre la diferencia de medias**, el cual es unilateral atendiendo a la formulación de la hipótesis alternativa:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 < 0$$

donde μ_1 es la media de la población de la que se extrae la primera muestra y μ_2 es la media de la población de la que extrae la segunda. Así, tomaremos $\alpha = 0,05$.

```
t.test(automoviles.std.precios, automoviles.turbo.precios,
          alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  automoviles.std.precios and automoviles.turbo.precios
## t = -3.2358, df = 67.249, p-value = 0.0009417
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -1851.954
## sample estimates:
## mean of x mean of y
##  12465.23  16287.11
```

Puesto que obtenemos un p-valor menor que el valor de significación fijado, rechazamos la hipótesis nula. Por tanto, podemos concluir que, efectivamente, el precio de un coche es superior si éste trae consigo un motor turbo.

2.5.3. Modelo de regresión lineal

Tal y como se planteó en los objetivos de la actividad, resultará de mucho interés poder realizar predicciones sobre el precio de un vehículo dadas sus características. Así, se calculará un modelo de regresión lineal utilizando regresores tanto cuantitativos como cualitativos con el que poder realizar las predicciones de los precios.

Para obtener un modelo de regresión lineal considerablemente eficiente, lo que haremos será obtener varios modelos de regresión utilizando las variables que estén más correladas con respecto al precio, según la tabla obtenido en el apartado 2.1.5. Así, de entre todos los modelos que tengamos, escogeremos el mejor utilizando como criterio aquel que presente un mayor coeficiente de determinación (R^2).

```
# Regresores cuantitativos con mayor coeficiente
# de correlación con respecto al precio
tamaño = automoviles$length
ancho = automoviles$width
peso.vacio = automoviles$curb.weight
tamaño.motor = automoviles$engine.size
caballos = automoviles$horsepower
consumo.ciudad = automoviles$city.mpg
consumo.autopista = automoviles$highway.mpg
max.rpm = automoviles$peak.rpm

# Regresores cualitativos
marca = automoviles$make
combustible = automoviles$fuel.type
estilo = automoviles$body.style

# Variable a predecir
precio = automoviles$price

# Generación de varios modelos
modelo1 <- lm(precio ~ marca + combustible + estilo + caballos +
              max.rpm + consumo.autopista, data = automoviles)
modelo2 <- lm(precio ~ marca + estilo + max.rpm +
              tamaño + ancho, data = automoviles)
modelo3 <- lm(precio ~ combustible + tamaño.motor + peso.vacio +
              caballos + consumo.autopista, data = automoviles)
modelo4 <- lm(precio ~ marca + tamaño.motor + peso.vacio +
              caballos + consumo.ciudad, data = automoviles)
modelo5 <- lm(precio ~ estilo + peso.vacio + max.rpm +
              caballos + consumo.autopista, data = automoviles)
```

Para los anteriores modelos de regresión lineal múltiple obtenidos, podemos utilizar el coeficiente de determinación para medir la bondad de los ajustes y quedarnos con aquel modelo que mejor coeficiente presente.

```
# Tabla con los coeficientes de determinación de cada modelo
tabla.coeficientes <- matrix(c(1, summary(modelo1)$r.squared,
                               2, summary(modelo2)$r.squared,
                               3, summary(modelo3)$r.squared,
                               4, summary(modelo4)$r.squared,
                               5, summary(modelo5)$r.squared),
                               ncol = 2, byrow = TRUE)

colnames(tabla.coeficientes) <- c("Modelo", "R^2")
tabla.coeficientes
```

```
##      Modelo       R^2
## [1,]      1 0.8856685
## [2,]      2 0.8631248
## [3,]      3 0.7899618
## [4,]      4 0.8948572
## [5,]      5 0.7788741
```

En este caso, tenemos que el cuarto modelo es el más conveniente dado que tiene un mayor coeficiente de determinación. Ahora, empleando este modelo, podemos proceder a realizar predicciones de precios de vehículos como la siguiente:

```
newdata <- data.frame(
    marca = "audi",
    tamaño.motor = 122,
    peso.vacio = 1989,
    caballos = 120,
    consumo.ciudad = 20
)

# Predecir el precio
predict(modelo4, newdata)
```

```
##        1
## 14884.47
```

2.6. Conclusiones

Como se ha visto, se han realizado tres tipos de pruebas estadísticas sobre un conjunto de datos que se correspondía con diferentes variables relativas a vehículos con motivo de cumplir en la medida de lo posible con el objetivo que se planteaba al comienzo. Para cada una de ellas, hemos podido ver cuáles son los resultados que arrojan (entre otros, mediante **tablas**) y qué conocimientos pueden extraerse a partir de ellas.

Así, el análisis de correlación y el contraste de hipótesis nos ha permitido conocer cuáles de estas variables ejercen una mayor influencia sobre el precio final del coche, mientras que el modelo de regresión lineal obtenido resulta de utilidad a la hora de realizar predicciones para esta variable dadas unas características concretas.

Previamente, se han sometido los datos a un preprocesamiento para manejar los casos de ceros o elementos vacíos y valores extremos (*outliers*). Para el caso del primero, se ha hecho uso de un método de imputación de valores de tal forma que no tengamos que eliminar registros del conjunto de datos inicial y que la ausencia de valores no implique llegar a resultados poco certeros en los análisis. Para el caso del segundo, el cual constituye un punto delicado a tratar, se ha optado por incluir los valores extremos en los análisis dado que parecen no resultar del todo atípicos si los comparamos con los valores que toman las correspondientes variables para coches que existen en el mercado actual.

3. Recursos

1. Dalgaard, P. (2008). *Introductory statistics with R*. Springer Science & Business Media.

2. Vegas, E. (2017). *Preprocesamiento de datos*. Material UOC.

3. Gibergans, J. (2017). *Regresión lineal múltiple*. Material UOC.

4. Rovira, C. (2008). *Contraste de hipótesis*. Material UOC.

5. *Test for homogeneity of variances - Lavene's test and the Fligner Killeen test* (2016) [en línea]. bioSt@TS. [Consulta: 26 de diciembre de 2017] https://biostats.w.uib.no/test-for-homogeneity-of-variances-levenes-test/