Dataset Regresión: Abalone

Ricardo Ignacio Shepstone Aramburu

Indice

Análisis exploratorio de datos

Definición del problema

El abulón (oreja del mar) es un molusco cuya concha es larga, plana y generalmente de forma ovalada. Para determinar su edad, normalmente se debe de cortar la concha, pulir, teñir con un colorante y examinar bajo un microscopio, para contar el número de anillos que se van formando conforme la cocha crece. Puesto que ciertos anillos son difíciles de contar, se determina que sumar 1.5 al número de anillos contados es una buena aproximación de la edad del individuo.

Este método para determinar la edad de un abulón es complejo y tedioso, por lo que es de especial interés intentar determinar la edad tomando otro tipo de medidas. En este dataset se dispondrá de ciertas medidas físicas, como dimensiones y pesos, así como el número de anillos de un conjunto de individuos; con el fin de intentar modelizar una regresión para predecir la edad (número de anillos) a partir del sexo y las medidas físicas.

En cuanto a la dependencia de las variables, tenemos que tener en cuenta que en la mayoría de las especies, un individuo crece en tamaño y aumenta en peso a lo largo de su vida, hasta llegar a cierto límite. Estas variables a su vez son dependientes entre ellas, ya que un individuo de mayor tamaño tendrá un peso mayor. Por otro lado, la edad no está directamente relacionada con el sexo, pero debido al dimorfismo sexual que existe en la mayoría de especies, el tamaño de un individuo puede verse afectado por su sexo en mayor o menor grado según la especie. En este caso particular, la variable de sexo tiene un tercer valor “Infant” que puede proporcionar cierta información sobre la edad del individuo, por lo que esta variable también tiene cierto grado de dependencia con el resto.

En base a esta información se procede a plantear ciertas cuestiones e **hipótesis**:

- El tamaño y el peso aumentan con la edad.
- Si es así, ¿Que relacción hay con respecto al número de anillos?
- ¿Alcanzan estas especies un límite de tamaño y peso?
- ¿Cómo es la relación entre las variables de dimensión y las de peso?
- La variable sexo influye sobre las medidas físicas del individuo.
- ¿Hay dimorfismo sexual?¿En qué grado?
- ¿Qué relacción hay entre la edad y que un individuo sea joven (“infant”) o adulto?

Preparación de los datos

Descripción de los datos

A partir del dataset de abalone se puede construir un data frame que consta de 4177 observaciones y 9 variables. Tanto en el archivo de texto proporcionado con los datos, como en la descripción del dataset que aparece en el repositorio de UCI (https://archive.ics.uci.edu/ml/datasets/abalone), se puede obtener la siguiente lista con información sobre las variables:

Nombre	Tipo de dato	Unidades	Descripción
Sex	Categórico nominal		Macho (M), hembra (F) e Infante (I).
Length	Cuantitativo continuo	mm	La medida más larga de la concha.
Diameter	Cuantitativo continuo	mm	Medida perpendicular a la longitud.
Height	Cuantitativo continuo	mm	Altura con la vianda.
Whole weight	Cuantitativo continuo	gramos	Peso del abulón.
Shucked weight	Cuantitativo continuo	gramos	Peso de la vianda.
Viscera weight	Cuantitativo continuo	gramos	Peso de vísceras.
Shell weight	Cuantitativo continuo	gramos	Peso de la concha.
Rings	Cuantitativo discreto		Número de anillos.

Como se ha mencionado anteriormente, la variable de salida será el número de anillos del abulón ("Rings"), mientras que las variables de entrada serán el resto de variables.

Procesamiento de los datos

El primer paso sería incluir los paquetes que utilizaremos y cargar el dataset con el que se va a trabajar.

```
require(tidyverse)
require(readr)
require(moments)
require(car)
require(corrplot)
require(fastDummies)


# Cargamos dataset
abalone.raw <- read.csv("Input/abalone/abalone.dat", comment.char="@", header=FALSE)
# Realizamos una copia con la que se trabajará
abalone <- abalone.raw
```

Resumen de los datos

Descomposición de atributos complicados

Busqueda de datos redundantes

Transformación de datos

Conclusiones

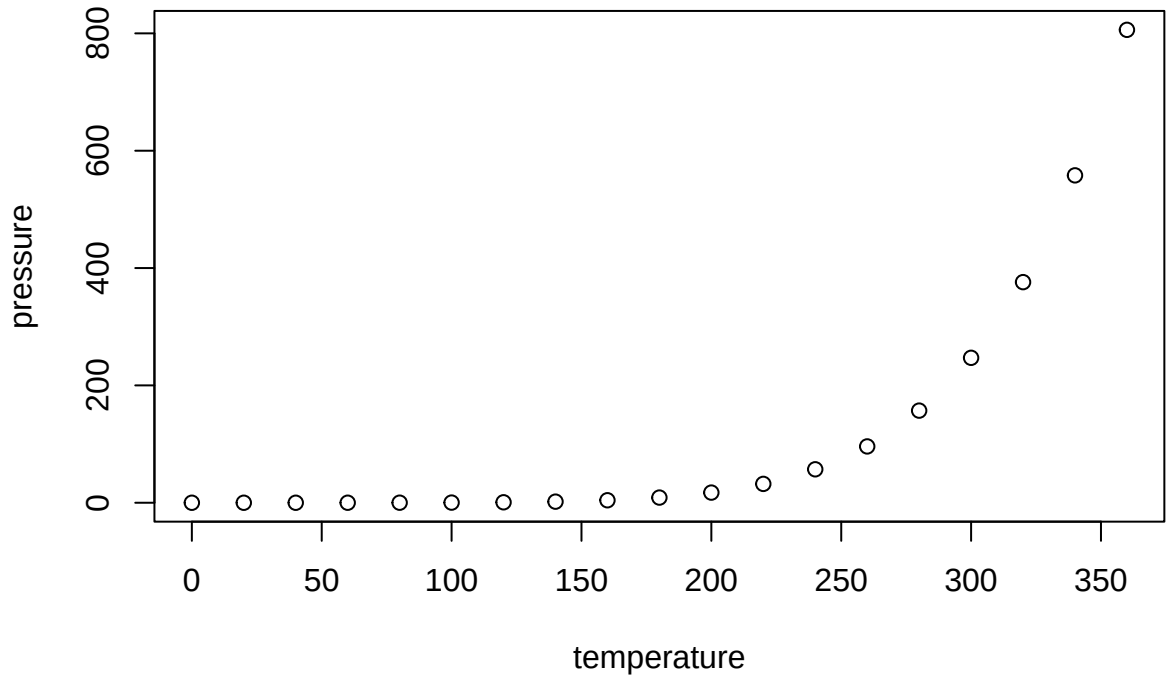
```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
```

```
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.