

Introducción a la ciencia de datos

Ricardo Ignacio Shepstone Aramburu

Indice

Dataset Regresión: Abalone	1
Definición del problema	1
Preparación de los datos	1
Including Plots	2

Dataset Regresión: Abalone

Definición del problema

El abulón (oreja del mar) es un molusco cuya concha es larga, plana y 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, 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.

hipótesis

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	Categorico 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 abalón.
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.

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

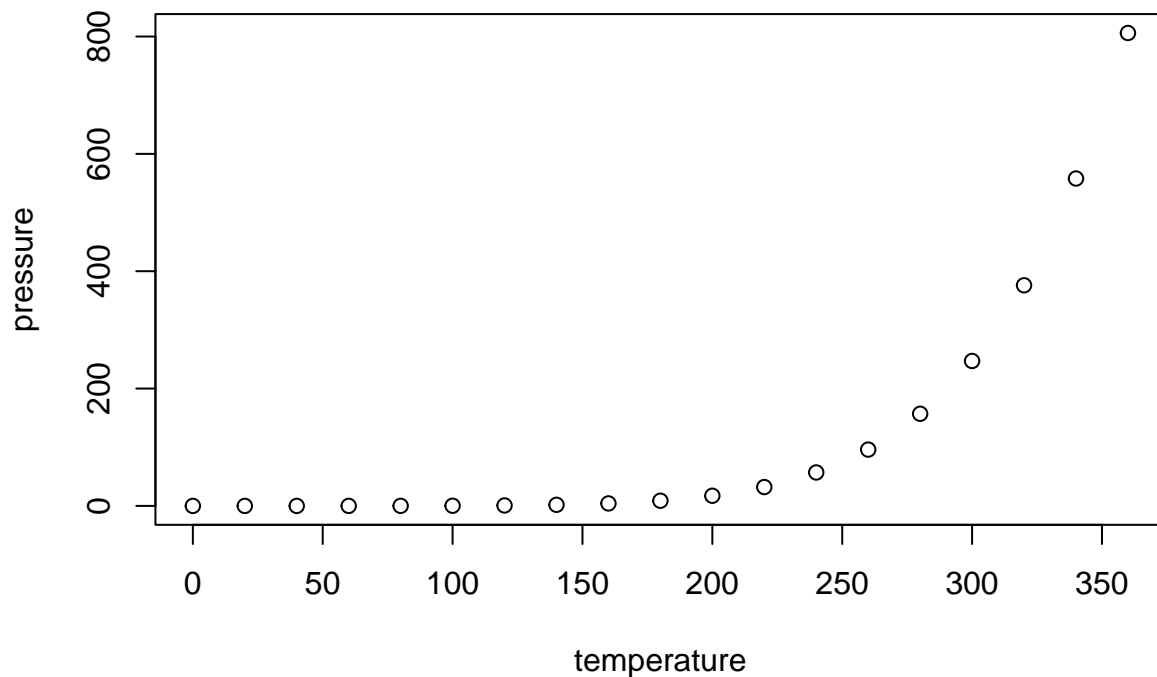
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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.