Extracción Automática del Conocimiento R-PL5

Carlos Javier Tacón Fernández Alicia Tomás Martínez
Zamar Elahi Fazal Roura

April 24, 2018

0.1 Ejercicio completo de detección de datos anómalos

Desarrollo por parte de cada alumno del enunciado y la solución de un ejercicio en el que se realicen análisis con R de detección de datos anómalos utilizando todos los me?todos vistos en teoría e introduciendo modificaciones sobre el ejercicio hecho en clase.

Para la realización de la tercera parte de la práctica vamos a usar un dataset encontrado en UCI que refleja datos de ausentismo en el trabajo. [1].

Objetivos

Eliminar los 'outliers' o valores atípicos no es un procedimiento estándar y quizá tampoco sea lo más correcto, sin embargo eliminarlos puede cambiar en gran medida el análisis de los datos, permitiendo hallar nuevos datos entre datos semejantes. Para entender bien y visualizar cómo afecta el tratamiento de outliers vamos a:

- Estudiar los dator tratandolos si es preciso
- Hallar la regresión linear previa al tratamiento de outliers.
- Realizar varios métodos para el tratamiento de outliers.
- Hallar la regresión linear posterior al tratamiento de outliers.

Los métodos que vamos a usar y demostrar para el tratamiento de los valores aítpicos son:

- K-Vecinos
- B
- C
- D

Manipulación de los datos.

Nuestra muestra trata sobre las ausencias que realizan ciertos empleados en una empresa. Vamos a echar un rápido vistazo a nuestra muestra.

```
#Primero la cargamos
data<- read.csv(file="data/Absenteeism_at_work.csv", header=TRUE, sep=";")
colnames(data)
```

```
##  [1] "ID"                          "Reason.for.absence"
##  [3] "Month.of.absence"            "Day.of.the.week"
##  [5] "Seasons"                     "Transportation.expense"
##  [7] "Distance.from.Residence.to.Work" "Service.time"
```

```
##  [9] "Age"                                "Work.load.Average.day"
## [11] "Hit.target"                         "Disciplinary.failure"
## [13] "Education"                          "Son"
## [15] "Social.drinker"                     "Social.smoker"
## [17] "Pet"                                "Weight"
## [19] "Height"                             "Body.mass.index"
## [21] "Absenteeism.time.in.hours"
```

Vamos a mirar cómo están distribuidos los datos:

```
head(data$ID)

## [1] 11 36  3  7 11  3
```

Podemos observar que se crea una fila por cada ausencia y no por cada empleado, observando la descripción del dataset [1] vemos que existen valores constantes (cómo edad, peso, altura...) y otros valores como 'Reason of absense' qué son la razón de por qué se hace una fila por ausencia si no que queremos un conjunto similar de empleados de edades similares y número de ausencias parecidas.

```
#Filas por ausencia
nrow(data)

## [1] 740

#Filas por empleado
nrow(table(unique(data$ID)))

## [1] 36
```

Para nuestro estudio **no** querémos eliminar a los empleados que tengan *ciertos tipos de ausencia*. Queremos eliminar aquellos empleados outliers por edad y número de ausencias. Por ejemplo, para nuestra empresa un empleado muy mayor apunto de jubilarse puede ser un valor atípico (suelen hacer más visitas al centro médico) o alguien de 28 años que es un hipocondríaco que tiene más de 100 ausencias y también puede ser un valor atípico.

Por lo tanto tenemos que manipular la estructura de los datos para hacer una fila por empleado y obtener su numero de ausencias.

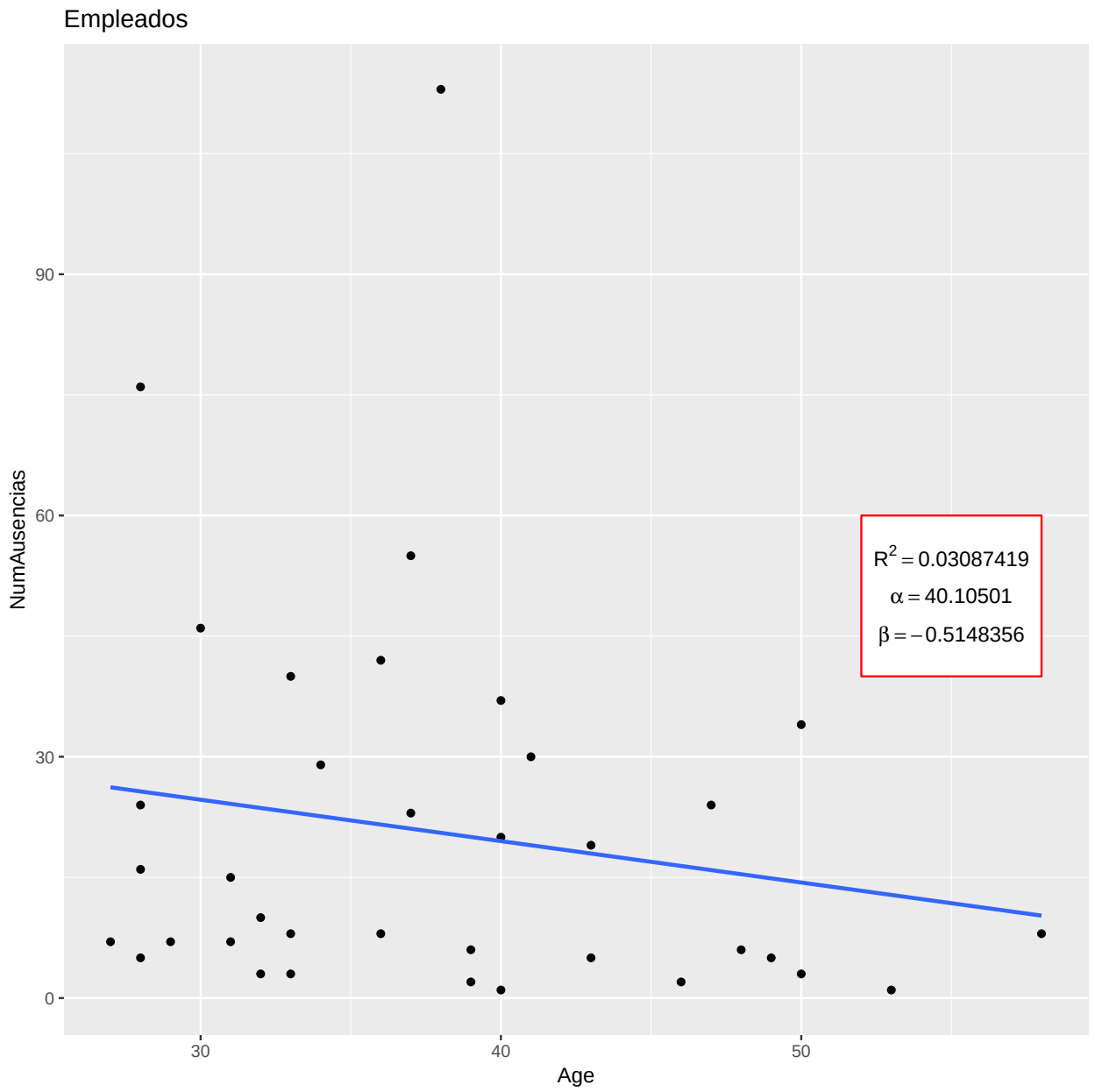
Esta parte es sencilla, creamos una nueva tabla de datos quitando los duplicados y obteniendo el número de ausencias cómo la frecuencia absoluta de cada empleado [ID] (ya que había una fila para cada ausencia de un empleado).

```
empleados = data[!duplicated(data$ID),]
#Ordenamos
empleados = empleados[order(empleados$ID),]
#Sacar número de ausencias con la frecuenci absoluta.
empleados$NumAusencias = as.numeric(table(data$ID))
#Vemos que hay tantas filas como empleados
nrow(empleados)

## [1] 36
```

Estudio de la muestra

```r
#Usamos ggplot2 - library(ggplot2)
# Referencia: http://t-redactyl.io/blog/2016/05/creating-plots-in-r-using-ggplot2-part-11-linear-regression-plots.html
lm <- lm(NumAusencias ~ Age, data = empleados)
alpha <- as.numeric(lm$coefficients[1])
beta <- as.numeric(lm$coefficients[2])
Rcuadrado <- summary(lm)$r.squared
dispersion <- ggplot(empleados, aes(x=Age,y=NumAusencias)) + geom_point() + geom_smooth(method='lm',se=FALSE) +
  ggtitle("Empleados")+
  annotate("rect", xmin = 52, xmax = 58, ymin = 40, ymax = 60, fill="white", colour="red") +
  annotate("text", x=55, y=55, label = paste("R^2 == ", Rcuadrado), parse=T) +
  annotate("text", x=55, y=50, label = paste("alpha == ",alpha), parse=T) +
  annotate("text", x=55, y=45, label = paste("beta == ",beta), parse=T)
dispersion
```



La calidad de la recta de regresión es muy baja **(0.0308742)**, vamos a ver si eliminando valores atípicos podemos mejorarla.

K-Vecinos

Para la realización de esta parte de la práctica vamos a hacer uso de la funcion knn [2]disponible en el paquete class, y el paquete gmodels para garantizar la precision de los valores

Para ello cargamos los paquetes y guardamos una copia de los datos:

References

[1] Data Set, *https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work*, UCI.

[2] Uso Knn, *https://www.analyticsvidhya.com/blog/2015/08/learning-concept-knn-algorithms-programming/*, analyticsvidhya.