Regresión Lineal

Juan Sebastián González Pinilla

Importar, limpiar y visualizar los datos

Nos basaremos en unos datos sobre el salario de personas de distintas nacionalidades, con diferentes edades. En este ejercicio nos basarémos únicamente del salario y la edad, la finalidad es predecir el salario que puede ganar una persona de 33, 60 y 25 años.

- Se usará la librería `readr`
- Se importará los datos con el delimitador ;
- Se crean variables basado en las columnas salario y edad

```
library(readr)
```

```
data = read_csv("Dia9.csv")
data
```

```
## # A tibble: 10 x 4
##    Pais     Edad Salario Compra
##    <chr>   <dbl>   <dbl> <chr>
##  1 France     44   72000 No
##  2 Spain      27   48000 Yes
##  3 Germany    30   54000 No
##  4 Spain      38   61000 No
##  5 Germany    40     NaN Yes
##  6 France     35   58000 Yes
##  7 Spain     NaN   52000 No
##  8 France     48   79000 Yes
##  9 Germany    50   83000 No
## 10 France     37   67000 Yes
```

```
Edad=data$Edad;Edad
```

```
##  [1]  44  27  30  38  40  35 NaN  48  50  37
```

```
Salario=data$Salario;Salario
```

```
##  [1] 72000 48000 54000 61000   NaN 58000 52000 79000 83000 67000
```

Creamos dos dataset nuevos con los NaN eliminados

```
dataEdad=data[!is.na(data$Edad),]
dataSalario=data[!is.na(data$Salario),]
```

Reemplazamos los NaN del dataframe original con la media de los datos dataframe nuevos (porque no contienen el NaN)

```r
data$Edad[is.na(data$Edad)] = mean(dataEdad$Edad)
data$Salario[is.na(data$Salario)] = mean(dataSalario$Salario)
data
```

```
## # A tibble: 10 x 4
##    Pais     Edad Salario Compra
##    <chr>   <dbl>   <dbl> <chr>
##  1 France   44    72000  No
##  2 Spain    27    48000  Yes
##  3 Germany  30    54000  No
##  4 Spain    38    61000  No
##  5 Germany  40    63778. Yes
##  6 France   35    58000  Yes
##  7 Spain    38.8  52000  No
##  8 France   48    79000  Yes
##  9 Germany  50    83000  No
## 10 France   37    67000  Yes
```

Modelar la ecuación de Regresión

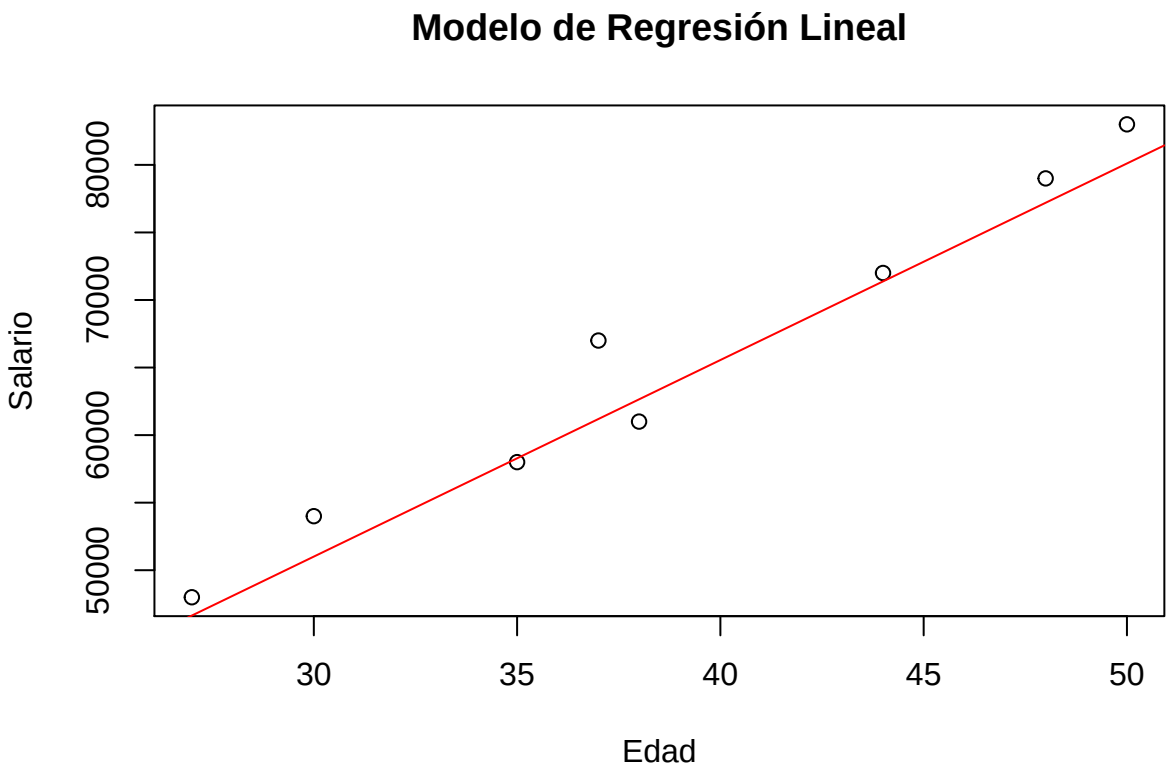
Modelar el salario por la variable edad

```r
lm_Sal_Ed = lm(Salario~Edad,data=data)
lm_Sal_Ed
```

```
##
## Call:
## lm(formula = Salario ~ Edad, data = data)
##
## Coefficients:
## (Intercept)         Edad
##        7362         1455
```

Visualizamos los valores y la curva del modelo de regresión lineal

```r
plot(Edad,Salario,main="Modelo de Regresión Lineal")
abline(lm_Sal_Ed, col = "red")
```



Predicción

Creamos un marco de datos invisible, que será el valor que vamos a predecir

```
valor_predecir = data.frame(Edad = c(33,60,25));valor_predecir
```

```
##   Edad
## 1   33
## 2   60
## 3   25
```

Predecirmos el valor creado con base en el modelo de regresión

```
predict(lm_Sal_Ed, newdata = valor_predecir)
```

```
##        1        2        3
## 55372.00 94652.85 43733.23
```

Conclusiones

Basado en el modelo lineal que se estimó a partir del conjunto de datos de `data`, se predijo que:

- El salario promedio para un trabajador de 33 años sería de alrededor de 55.372 dólares al año
- El salario promedio para un trabajador de 60 años sería de alrededor de 94.652,85 dólares al año
- El salario promedio para un trabajador de 25 años sería de alrededor de 43.733,23 dólares al año