Modelo de regresión lineal: análisis descriptivo de los datos

Contents

1 Introducción

Vamos a leer el archivo de datos *kidiq.csv*:

```
d = read.csv("datos/kidiq.csv")
str(d)
```

```
## 'data.frame':    434 obs. of  5 variables:
##  $ kid_score: int  65 98 85 83 115 98 69 106 102 95 ...
##  $ mom_hs   : int  1 1 1 1 1 0 1 1 1 1 ...
##  $ mom_iq   : num  121.1 89.4 115.4 99.4 92.7 ...
##  $ mom_work : int  4 4 4 3 4 1 4 3 1 1 ...
##  $ mom_age  : int  27 25 27 25 27 18 20 23 24 19 ...
```

donde se recogen datos de las siguientes variables:

- **kid_score** : puntuacion de un test cognitivo en niños de 3-4 años
- **mom_hs** :
  - mom_hs = 1 : las madres han terminado secundaria (high school)
  - mom_hs = 0 : las madres no terminaron secundaria
- **mom_iq** : puntuación de la madre en otro test cognitivo
- **mom_work** :
  - mom_work = 1 : la madre no trabajó en los primeros tres años del niño
  - mom_work = 2 : la madre trabajó en el segundo o tercer año
  - mom_work = 3 : la madre trabajó a tiempo parcial el primer año
  - mom_work = 4 : la madre trabajó a tiempo completo el primer año
- **mom_age** : edad de la madre

Estamos interesados en estudiar si la puntuación obtenida por los niños (variable *kid_score*) está relacionada con las otras variables. Luego:

- La variable respuesta es numérica, cuantitaviva.
- Los regresores son de dos tipos: cuantitativos (mom_iq, mom_age) y cualitativos (mom_hs, mom_work).
- Los factores cualitativos se tienen que representar como factores:

```
d$mom_hs = factor(d$mom_hs, labels = c("no", "si"))
d$mom_work = factor(d$mom_work, labels = c("notrabaja", "trabaja23", "trabaja1p", "trabaja1c"))
```

Se guardan en formato Rdata ya que se van a utilizar mucho:

```
save(d,file="datos/kidiq.Rdata")
# para leer estos datos: load(kidiq.Rdata)
```

2 Análisis descriptivo de las variables cualitativas

```
(t1 = table(d$mom_hs))
```

```
##
##  no  si
##  93 341
```

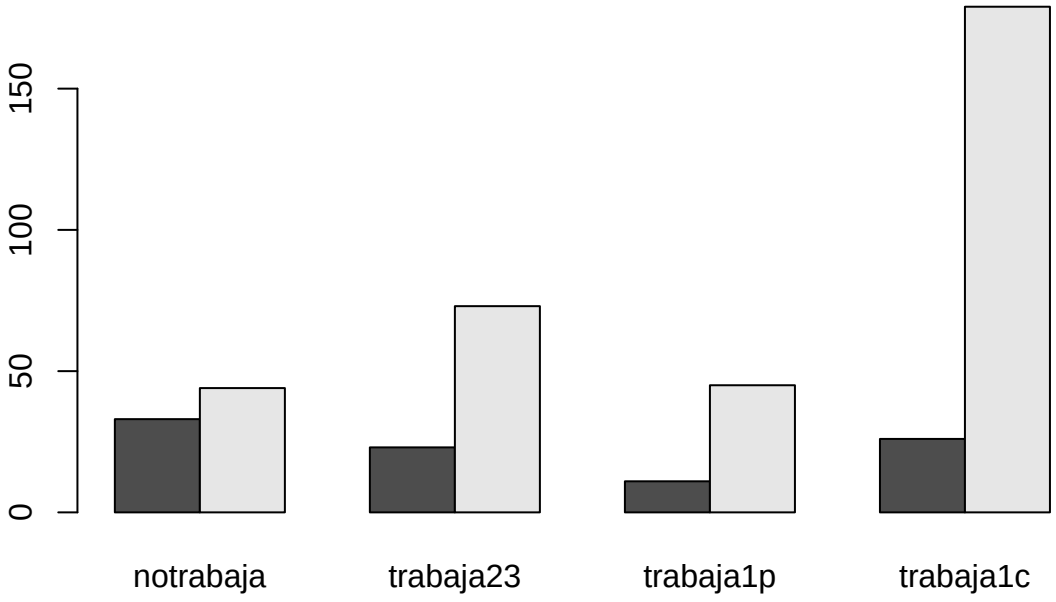
```
(t2 = table(d$mom_work))
```

```
##
## notrabaja trabaja23 trabaja1p trabaja1c
##        77        96        56       205
```

```
t3 = table(d$mom_hs,d$mom_work)
addmargins(t3)
```

```
##
##       notrabaja trabaja23 trabaja1p trabaja1c Sum
##   no         33        23        11        26  93
##   si         44        73        45       179 341
##   Sum        77        96        56       205 434
```

```
barplot(t3, beside = T)
```



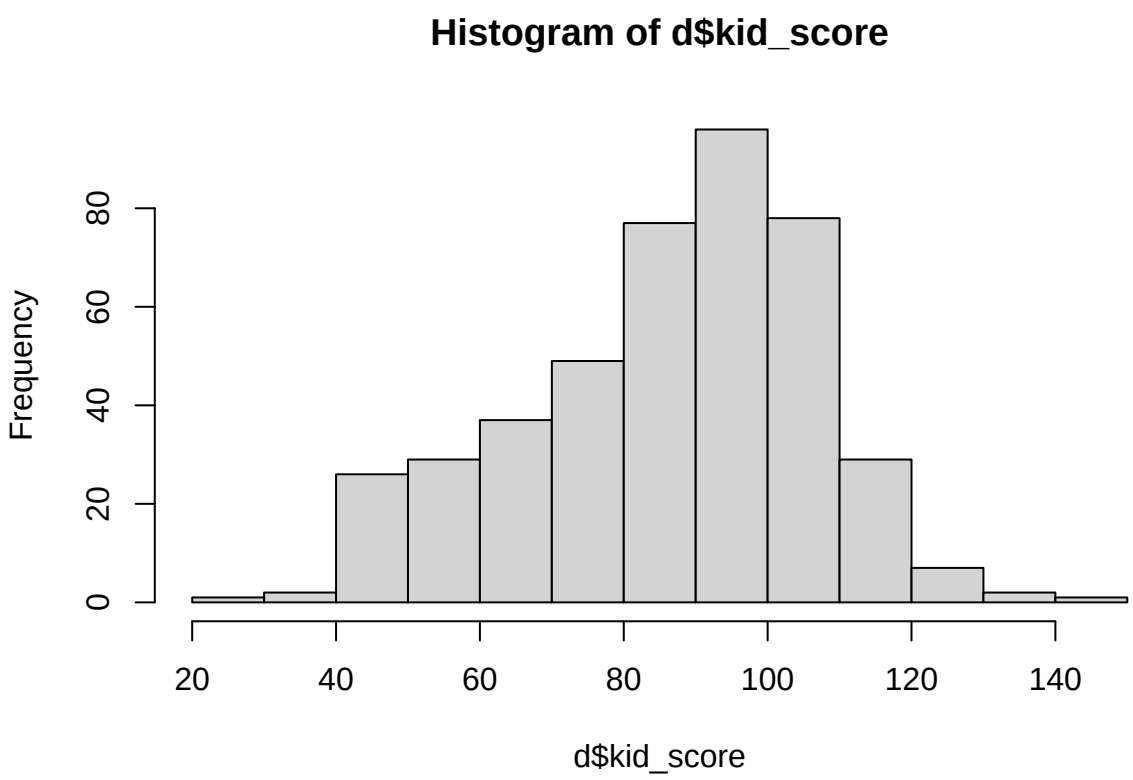
3 Análisis descriptivo de las variables cuantitativas

```
summary(d[,c("kid_score","mom_iq","mom_age")])
```

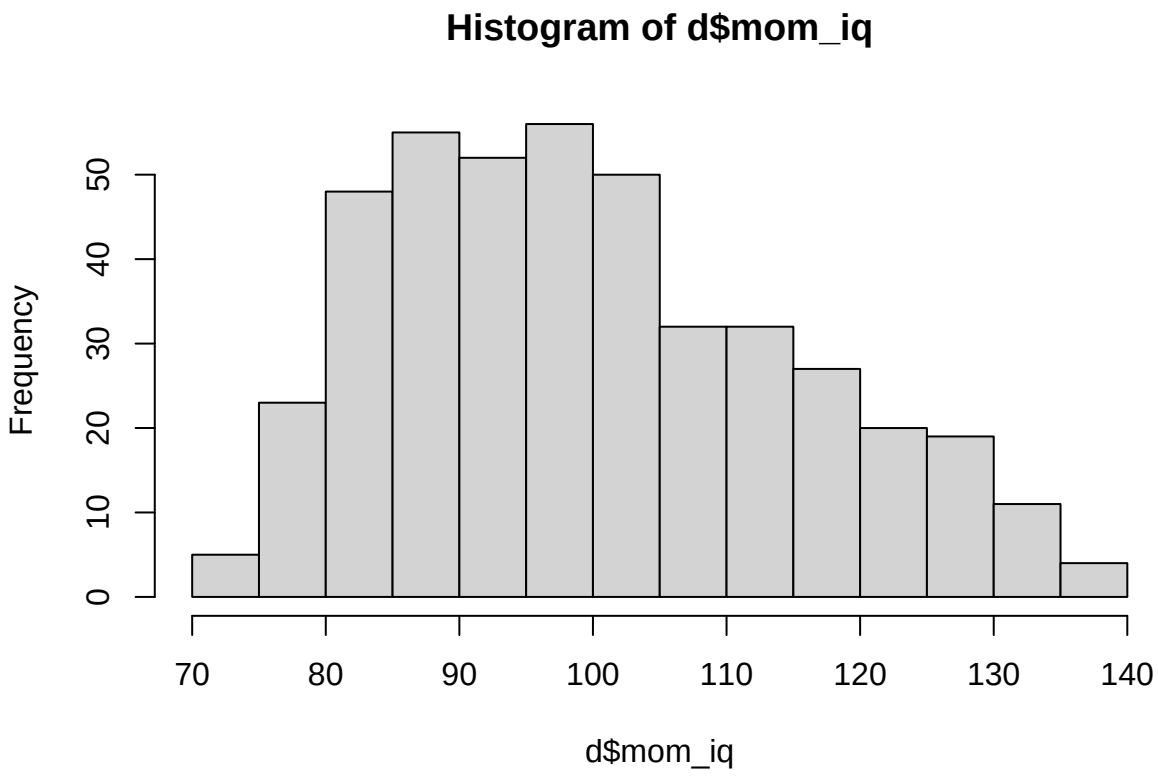
```
##    kid_score          mom_iq          mom_age
##  Min.   : 20.0   Min.   : 71.04   Min.   :17.00
```

```
##  1st Qu.: 74.0   1st Qu.: 88.66   1st Qu.:21.00
##  Median : 90.0   Median : 97.92   Median :23.00
##  Mean   : 86.8   Mean   :100.00   Mean   :22.79
##  3rd Qu.:102.0   3rd Qu.:110.27   3rd Qu.:25.00
##  Max.   :144.0   Max.   :138.89   Max.   :29.00
```

```
hist(d$kid_score)
```

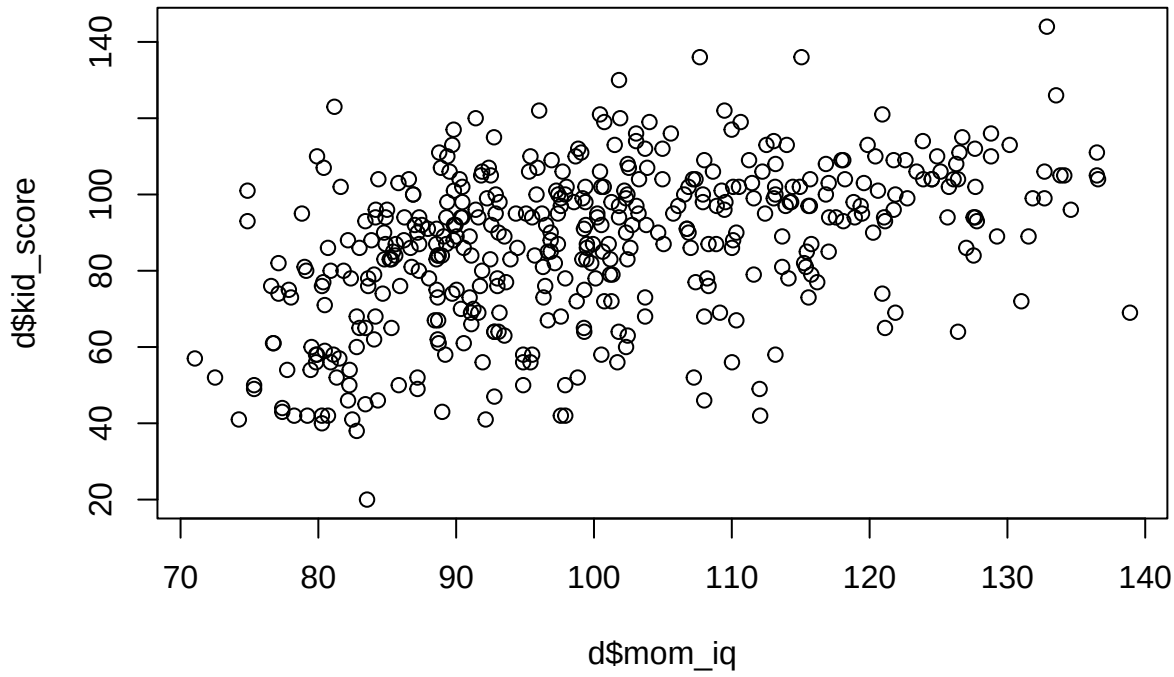



```
hist(d$mom_iq)
```

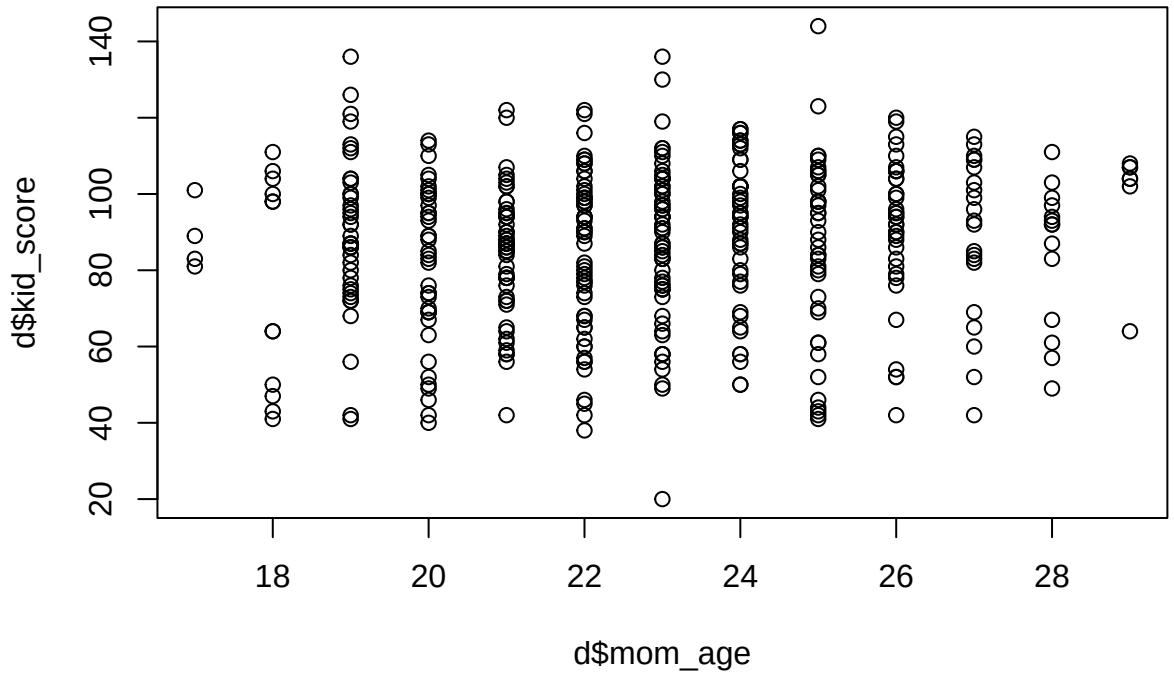



```
hist(d$mom_age)
```


```
plot(d$mom_iq,d$kid_score)
```




```
plot(d$mom_age,d$kid_score)
```



Parece que hay cierta relación lineal entre kid_score y mom_iq. Sin enmbargo, no está clara dicha relación entre kid_score y mom_age.