

Regressio

Miquel Miró Nicolau

13/11/2020

En aquest document farem un primer ús de la regressió lineal simple (SLR, per l'anglès *Simple linear regression*). Aquesta tècnica ens permet cercar dependències lineals entre dues variables

Dades

Primerament creem un *dataframe* per provar la SLR. En aquest cas crearem un dataset amb dues característiques, el pes i l'altura.

```
pes = c(75, 80, 42, 56, 104)
altura = c(180, 175, 150, 160, 200)

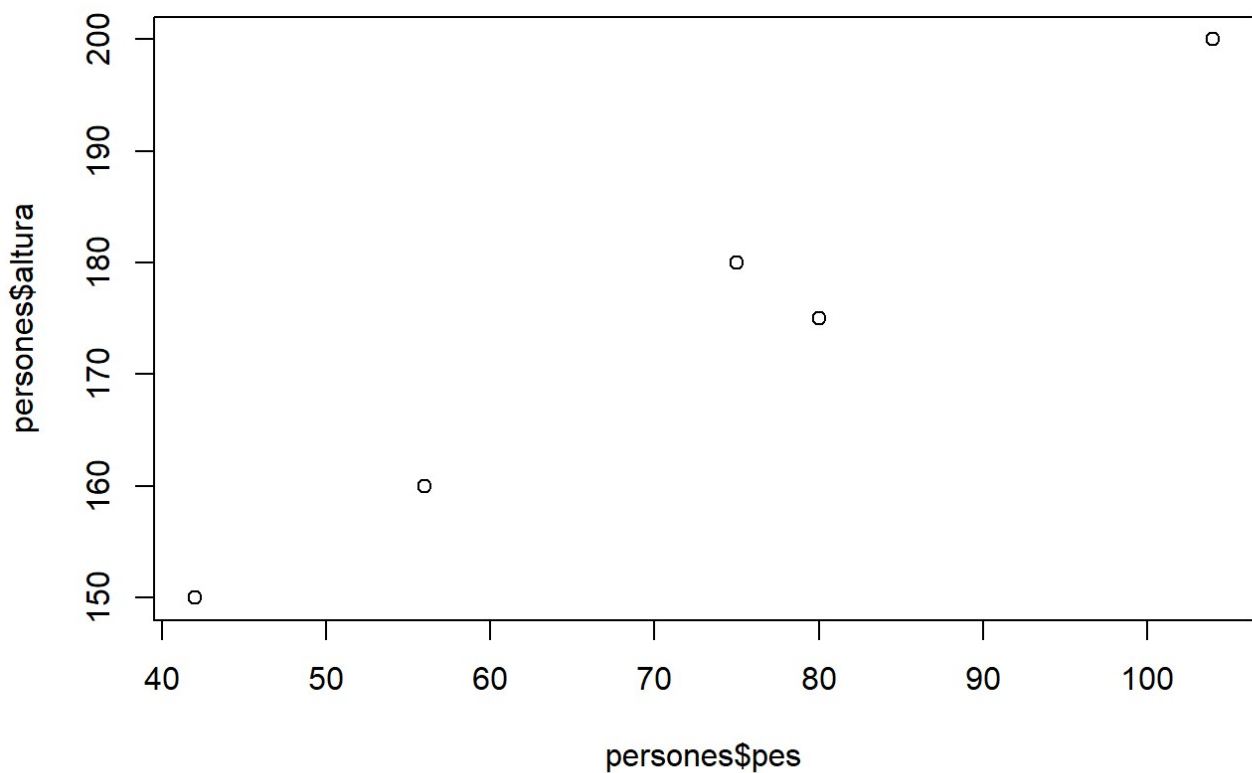
persones = data.frame(pes, altura)
persones
```

```
##   pes altura
## 1  75    180
## 2  80    175
## 3  42    150
## 4  56    160
## 5 104    200
```

Plot

Una vegada que tenim les dades podem fer un gràfic que ens permeti observar a simple vista si hi ha, o no, una dependència lineal entre les dues variables.

```
plot(persones$pes, persones$altura)
```



El gràfic anterior ens permet observar que una major altura suposa un major pes, per tant, en aquest cas, podem afirmar que una de les condicions per poder fer una regressió lineal simple es compleix.

SLR

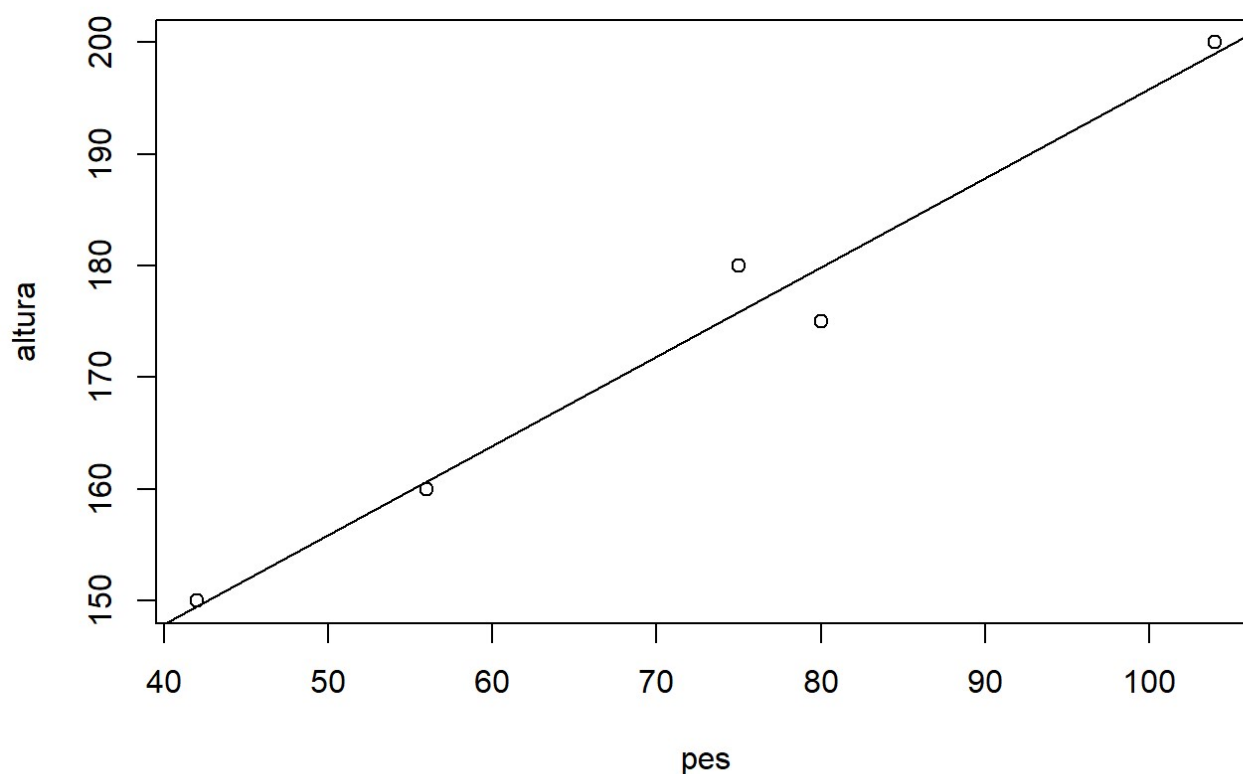
Realitzam la regressió lineal amb la comanda `lm`.

```
model = lm(persones$altura ~ persones$pes)
model
```

```
##
## Call:
## lm(formula = persones$altura ~ persones$pes)
##
## Coefficients:
## (Intercept)  persones$pes
##      115.9422      0.7991
```

Observam amb el resultat que s'ha creat un model per prediure l'altura de les persones amb la seqüent equació $\text{Altura} = 115 + 0.799 \cdot \text{pes}$. Podem fer un gràfic a on s'observi el model i les dades

```
plot(persones)
abline(model)
```



Finalment volem obtenir una mètrica numèrica de la qualitat del model proposat. Per obtenir-la emprarem la funció `summary` del model.

```
summary(model)
```

```
##
## Call:
## lm(formula = persones$altura ~ persones$pes)
##
## Residuals:
##      1      2      3      4      5
## 4.1231 -4.8725  0.4944 -0.6934  0.9484
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  115.9422     5.8995   19.65 0.000288 ***
## persones$pes    0.7991     0.0792   10.09 0.002073 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.758 on 3 degrees of freedom
## Multiple R-squared:  0.9714, Adjusted R-squared:  0.9618
## F-statistic: 101.8 on 1 and 3 DF, p-value: 0.002073
```

La mètrica que ens interessa és la *Multiple R-squared*, sabem que el valor màxim d'aquesta mètrica és 1 i indica el millor model. En el nostre cas, ja que el valor és de 0.9714, podem afirmar que és un bon model